Universidade do Minho
Escola de Ciências

Ângela Cristina Franco Ferreira

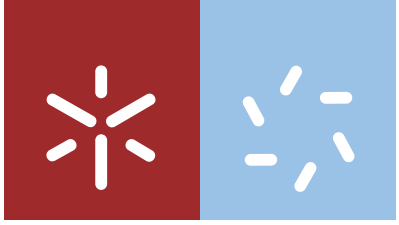
Joint Modelling of Longitudinal Data and Time Until Premature Termination in Psychotherapy

July 2020

Joint Modelling of Longitudinal Data and Time Until Premature Termination in Psychotherapy

Ângela Cristina Franco Ferreira

UMinho | 2020



Universidade do Minho
Escola de Ciências

Ângela Cristina Franco Ferreira

Joint Modelling of Longitudinal Data and Time Until Premature Termination in Psychotherapy

Master Dissertation
Master Degree in Statistics

Dissertation supervised by
Professor Inês Pereira Silva Cunha Sousa
Professor Eugénia Maria Ribeiro Pereira

July 2020

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

ACKNOWLEDGEMENTS

I'd really like to take the opportunity to thank my supervisor, Dr. Inês Sousa (Professor in Statistics at the School of Sciences, University of Minho) for providing me this interesting topic, as well as valuable guidance throughout this journey. I appreciate your enthusiasm, energy, and scientific proactivity. Thank you for all the advice, ideas, and patience in guiding me.

My sincere thanks also goes to Dr. Eugénia Ribeiro (Professor in Psychology, University of Minho), my co-supervisor, who provided me the opportunity to join their research team, and who gave permission to access the data used in the present work. Her advice and guidance have helped me, not only to develop my research abilities, but also to become a better clinician.

I would like to extend my thanks to the teaching academic crew who, throughout this years of study in Statistics, facilitated a stimulating, productive, and enjoyable learning experience and deserve many thanks, namely: Dr. Susana Faria, Dr. Arminda Gonçalves, Dr. Luís Machado, Dr. Raquel Menezes, Dr. Maria Conceição Serra, and Dr. Irene Brito.

I am also deeply thankful to all my friends and colleagues for support and encouragement that motivated me to persevere through the toughest times. Special thanks are owed to Dulce Pinto and Carina Magalhães for friendship and emotional support, as well as to Pedro Afonso for provided me with valuable and useful suggestions on this project.

Lastly, I would like to thank my family and especially my partner, Tiago, for all your love, personal support and great patience at all times.

STATMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

JOINT MODELLING OF LONGITUDINAL DATA AND TIME UNTIL PREMATURE TERMINATION IN PSYCHOTHERAPY

Abstract: Joint modelling enables the simultaneous study of longitudinal and survival processes, exploiting the association between them. A particular case, adopted in the present work, is the shared random effects model, using a linear mixed effects model to represent the longitudinal process linked to a Cox regression model to represent the survival process. The primary purpose of this work is to briefly review the shared random-effect model methodology, along with independent survival and longitudinal models, and detail its implementation and evaluation, through a real data set. The focus is on the hazard of premature termination in psychotherapy, investigating the effect of two known process variables: therapeutic alliance quality and treatment outcome. Additionally, we aim tho infer which risk factors affect both the hazard of premature termination and these process variables. A data set of 97 clients, along with 12 variables, was collected from a university clinic, over a period of three years. These clients were assigned to the Unified Protocol for transdiagnostic treatment of emotional disorders. The benefits of joint modelling were highlighted through the comparison of joint models and separate survival and longitudinal methods. Results showed that, the therapeutic alliance quality and the treatment outcome mean progression were significantly associated with the hazard of premature termination for these clients. We conclude that independent analysis bring up bias parameter estimates, and an assumption of association between the two processes in a joint model of premature termination data is necessary.

Keywords: Joint modelling, premature termination in psychotherapy, therapeutic alliance, treatment outcome

MODELAÇÃO CONJUNTA DE DADOS LONGITUDINAIS E TEMPO ATÉ AO ABANDONO PREMATURO DA PSICOTERAPIA

Resumo: A modelação conjunta permite o estudo simultâneo de processos longitudinais e de sobrevivência, dando conta da possível associação entre estes. Uma abordagem em particular, adotada no presente trabalho, é a modelação conjunta com efeitos aleatórios partilhados, que utiliza o modelo linear misto para representar o processo longintudinal vinculado ao modelo de regressão de Cox para representar o processo de sobrevivência. O principal objetivo do presente trabalho é apresentar uma breve revisão da literatura acerca desta metodologia, assim como dos modelos independentes longitudinal e de sobrevivência, e detalhar a sua implementação e avaliação, através de uma aplicação a um conjunto de dados reais. Focamo-nos no risco de abandono prematuro da psicoterapia, investigando o efeito de duas variáveis processuais bem conhecidas: a qualidade da aliança e os resultados terapêuticos. Adicionalmente, pretendemos aferir os preditores que afetam o risco de abandono prematuro e cada uma das variáveis processuais. Consideraram-se os dados de 97 clientes, para 12 variáveis, recolhidos numa clínica universitária, durante um período de três anos. Estes clientes participaram num ensaio clínico com a aplicação do Protocolo Unificado para o tratamento transdiagnóstico de perturbações emocionais. As vantagens da modelação conjunta foram comprovadas pela comparação dos modelos conjuntos aos modelos individuais. Os resultados mostraram existir um efeito significativo da evolução, quer da qualidade da aliança, quer dos resultados terapêuticos, no risco de abandono prematuro da psicoterapia. Assim se concluiu que, análises independentes produzem estimativas dos parâmetros enviesadas, e que por isso é necessário considerar a associação entre os dois processos através da modelação conjunta de dados relativos ao abandono prematuro da psicoterapia.

Palavras-Chave: Abandono prematuro da psicoterapia, aliança terapêutica, modelação conjunta, resultados terapêuticos

CONTENTS

ACRONYMS

A

AIC Akaike Information Criterion.

B

BAI Beck Anxiety Inventory.

BDI-II Beck Depression Inventory.

C

CPHM Cox Proportional Hazards Model.

L

LMEE Linear Mixed Effects, Exponential Serial Correlation Model.

LMEG Linear Mixed Effects, Guassian Serial Correlation Model.

M

MAR Missing at Random.

MCAR Missing Completely at Random.

MNAR Missing Not at Random.

O

OLS Ordinary Least Square Model.

OQ-10.2 Outcome Questionnaire, short version.

OQ-45.2 Outcome Questionnaire.

P

PH Proportional Hazard.

PT Premature Termination.

U

UP Unified Protocol.

W

WAI Working Alliance Inventory.

LIST OF FIGURES

LIST OF TABLES

A missing value hides a meaningful value.

Little & Rubin (2019, p.16)

1

INTRODUCTION

Premature Termination (PT; also called dropout), or client's unilateral termination of services prior to completion of a recommended course of treatment, is a widespread problem in outpatient psychotherapy that unables the delivery of effective interventions. The most recent meta-analytic review, developed by Swift & Greenberg (2012), found that, across 669 studies and almost 84 000 clients, the average weighted dropout rate was approximately 20%, with higher rates among clients who were younger, had personality or eating disorder diagnosis, and were seen by trainee clinicians.

Treatment dropout has been linked to various damaging consequences for clients, including poorer treatment outcomes (Cahill et al., 2003; Pekarik, 1992), lower likelihood of recovery (E. Anderson & Lambert, 2001), higher dissatisfaction with the treatment received (Björk et al., 2009), and more likely to start and stop treatment, namely psychiatric support, on multiple occasions (Carpenter et al., 1979). Additionally to negative impact on clients, premature termination also influences service providers and mental health agencies by way of under-usage of their time and loss of revenue, due, for instance, to the missed appointments and lengthier waiting lists (Reis & Brown, 1999). Society as a whole is burdened by premature termination since others in need see their access to treatment denied, and dropout's associates, like family members and friends, are exposed to the client's continued impairment, as well as to the demand to persist in providing physical, emotional and even financial support (Cahill et al., 2003).

Clinicians and researchers long have sought to estimate whether and when PT is likely to occur, and what factors predict its occurrence, in order to prevent this type of termination and promote treatment effectiveness. And for almost as long, inconclusive and mixed findings in this area have been emerged, in part due to the recurrent application of subotimal, or even inappropriate, analytic techniques. In fact, psychotherapy termination data poses statistical model challenges to researchers, given their particular features, that, when overlooked, can cause considerable bias of results (K. Anderson, 2015; Corning & Malofeeva, 2004).

PT is an event in time which arises during a longitudinal process occurring over a series of sessions, and, as such, these data must be analysed longitudinally. Another feature of these data is that premature terminators do not seem to be an homogeneous group, since clients who terminate prematurely early in the treatment process appear to be different from those who dropout later, not only in observed outcomes at previous time points, but also in covariates that contribute to their termination status (K. Anderson, 2015).

Mechanisms behind dropout or PT (i.e., reasons for treatment discontinuation), poses another important methodological challenge to researchers. Despite modern longitudinal methods, as linear mixed effects models, do not require complete data at each measure point, they are based on *Missing at Random (MAR)* assumption (Little, 1995; Little & Rubin, 2019), which assumes that the probability of a dropout depends only on those variables observed in the model (e.g., the last observed value or covariates). However, in some cases, dropout may be non-ignorable, because it is likely to depend on unobserved or latent variables (e.g., owing to deterioration or improvement, clients could believe the treatment is either not useful or not need, and could decide to terminate participation). In these circumstances, besides MAR does not hold, clients that provide data (i.e., survivors) may differ from those who stop doing so (i.e., premature terminators). When this is not considered, it may lead to an overestimation of the treatment effect, since longitudinal profiles will probably reflect more an artefact caused by selective dropout, than genuine change over time (Wolke et al., 2009). Therefore, it is important to investigate missing mechanisms using models that consider non-ignorable dropout, that is *Missing Not at Random (MNAR)* models, as those proposed under joint modelling approach, based on a joint distribution of observations and dropout indicators.

Predictors of PT in psychotherapy, by turn, in most previous studies, have been conceptualized as static or constant over the data collection period (Corning & Malofeeva, 2004). Reviews have focused primarily on clients characteristics, including both demographic (e.g., age and gender) and psychological variables (e.g., diagnosis and readiness to change), as also on therapists characteristics (e.g., experience level, age and gender), treatment (e.g., theoretical orientation) and setting variables (e.g., type of clinic), in order to identify the factors that influence premature termination (Swift & Greenberg, 2015). Moreover, measurements usually are taken once, for each participant, and the variables are cast as static predictors, even they can or do change (e.g., symptom level) (Corning & Malofeeva, 2004). Some researchers, however, argued that a number of potentially meaningful predictors, which vary over the course of therapy, should be analysed and incorporated into appropriate statistical models; namely variables that capture the relational nature of the therapeutic process (e.g., therapeutic alliance and agreement on the presenting problems) (Samstag

et al., 1998). In terms of clinical implications, by paying attention to both static and time-variant predictors, clinicians may be able to reduce rates of premature termination, not only by making adaptations to the client and setting characteristics where need (e.g., particularly efforts on retention with younger clients and those with personality or eating disorder diagnosis), but also by applying specific strategies in therapy course (e.g., addressing motivation and clients preferences, as also repairing alliance ruptures) (Samstag et al., 1998).

Therapeutic alliance, defined as the agreement on the goals and tasks of therapy, in a context of a positive effective bond between therapist and client, has been found as an important predictor of overall outcome of therapy, as well as of PT (Bordin, 1979). For example, Samstag, Batchelder, Muran, Safran and Winston (1998) verified that, when compared to a good outcome group and even a poor outcome completer group, the dropout group rated the relationship as more problematic (i.e., the therapeutic alliance scores were significantly worse). In the same line, Sharf, Primavera, and Diener (2010), in a meta-analysis, found a significant relationship between the strength of the therapeutic alliance and PT; that is, weaker alliances were associated with an increased likelihood of drooping out, across settings, theoretical orientations, and perspectives (of both clients and therapists). According to the authors, the session-by-session monitoring of specific alliance and interpersonal patterns is useful either to predict the overall evaluation of treatment progress as also to inform therapists regarding specific intervention choices.

Based on the aforementioned considerations, in the present work, we adopted a longitudinal framework, in order to properly address PT data. Recall that, in many longitudinal studies, usually different types of outcomes are collected, such as repeated measurements of one or more response variables also called longitudinal data (e.g., depression and anxiety severity at a set of time points), and event times, also called survival data (e.g., time to recovery or time to dropout). Furthermore, we have implicit outcomes, like missing data, which need to be proper handling, under penalty of lead to biased estimates. Repeated measurement and survival data require different statistical methods, and are traditionally analyzed separately, as we will explain in the section 3.1 and 3.2 (Diggle et al., 2008). However, when the longitudinal outcome and the time-to-event (in our case dropout) mechanism are associated, separate analysis are not suitable, and a joint modelling approach is required, as will be explained in section 3.3 (Asar et al., 2015).

In the present work, we compared results of different statistical models in the context of longitudinal data, including separate analysis and joint modelling approach, where predictors of premature termination in psychotherapy are investigated, using a real data set from a psychology university clinic in the north of Portugal. Specifically, this work aims:

1. to produce an exploratory study on PT in psychotherapy in a specific outpatient context, regarding a particular period of time;

2. to explore separate statistical models for survival and longitudinal data;

3. to develop a statistical model on joint modelling of longitudinal variables progression and time to dropout; and,

4. to compare results of different models in order to conclude on the importance of joint models.

This dissertation is organized as follows: in Chapter 2, we introduce an overview of the Unified Protocol treatment program data, namely a description of variables, methodology and procedures employed, in order to conduct the empirical study. A review of the literature on statistical methodology, namely survival analysis (section 3.1) and longitudinal analysis (section 3.2), accounting for separate techniques to model survival and repeated measures data, respectively, will be exposed in Chapter 3. Additionally, in section 3.3, we will focus on the joint modelling approach, exploiting the main methodologies applied in this field along with a detailed description of the one employed in this particular study. Chapter 4 concerns the presentation of the main results, from survival (section 4.2), longitudinal (section 4.3) and joint modelling analysis (section 4.4). Finally, Chapter 5 provides a discussion of the main results and methodological limitations along with suggestions for future works.

2

PREMATURE TERMINATION IN PSYCHOTHERAPY DATA

To implement the empirical study, we used a real data set on *Unified Protocol (UP)* for transdiagnostic treatment of emotional disorders program collected at a psychology university clinic in the north of Portugal, over a period of three years (from 2015 October 1 to 2018 September 30). This clinic is a public association aimed at providing low-cost, high quality clinical services to members of community. Registry data collection and analysis was submitted to ethical appreciation and approved by the university's Ethics Committee for Research in Social and Human Sciences.

2.1. PARTICIPANTS

2.1.1 *Clients*

During the study period, 255 referrals for individual therapy were received, but 51 (20%) of individuals failed to attend first appointment. Of those 204 individuals that attended the first appointment, 130 (64%) continued to treatment and 74 (36%) terminated for a range of different reasons (e.g., failed to complete assessment, just asked for evaluation, were referred to another treatment or service). At intake, participants were invited to engage in a clinical trial to analyses the efficacy of Unified Protocol (Barlow et al., 2011) for transdiagnostic treatment of emotional disorders. Of those 130 individuals that initiated therapy, 33 (25%) were found to be unsuitable for UP or refused to participate (these clients were referred or choose a more preferable treatment). This resulted in 97 subjects for analysis. Figure 1 summarizes client enrollment and flow from 2015 October 1 to 2018 September 30.

Thus, of the 97 subjects who were assigned to UP treatment, nearly half (n = 51) completed the full course of treatment, as schedule by the treatment manual, or terminated at an earlier stage that was mutually agreed upon between the therapist and client (in cases where the client had reached the therapeutic goals). Of remaining 46 clients, 31 dropped out (i.e., terminated unilaterally without the therapist's

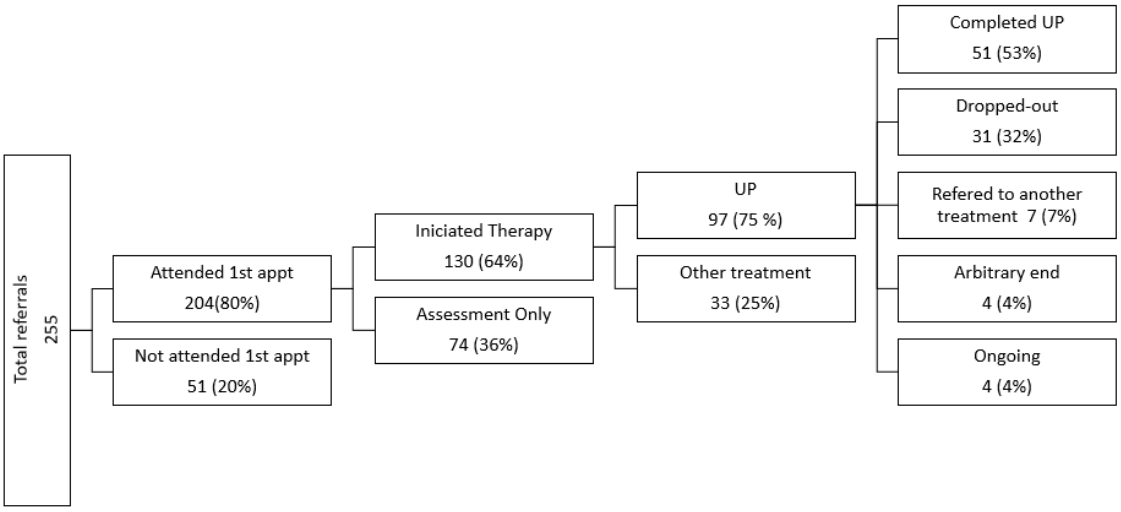


Figure 1: Flowchart showing attrition rates from October 2015 to September 2018

approval or knowledge), 7 were referred to more adequate treatment given Axis-II comorbility detected during treatment, 4 terminated prematurely in consequence of an arbitrary circumstance (e.g., the client graduated and moved to another city), and 4 cases were still ongoing at the data collection end point.

To be eligible for UP, clients had to receive a principal (most interfering and severe) diagnosis of an emotional disorder (Bullis et al., 2019), namely an anxiety disorder, depressive disorder, obsessive-compulsive or related disorders, somatic symptom or related disorders, and trauma or stress-related disorders as assessed using the Diagnostic Interview for DSM-5 Anxiety, Mood, and Obsessive-Compulsive and Related Disorders (DIAMOND; Tolin et al., 2018); be 18 years or older; be fluent in Portuguese; agreeing to the videotaping of sessions and the completion of research questionnaires, and provide informed consent. The exclusion consisted of those conditions that in clinical context would have required prioritization for immediate treatment, namely : (1) current significant suicidal risk; (2) presence of psychosis, mania, or organic mental disorder; and (3) current history of substance abuse or drug dependence, with the exception of nicotine and caffeine.

2.1.2 *Therapists*

The clinic staff comprised seven therapists (n = 5 women), with an average age of 31 years (range = 23 – 35 years). The doctoral-level therapists (n = 3) provided therapy services to just over half the clients (52%), the master's-level therapists (n = 3) saw 45% of the clients, and the intern (n = 1) saw 3% of the clients. Doctoral and master's level therapists had an average clinical practice experience of 4 years (range

= 2-8 years). Treatment was provided under the close supervision of four senior team members. All of the therapists met weekly for group supervisions of 90 minutes.

2.1.3 *Treatment*

The UP program consisted of a maximum of 16-20 individual sessions of approximately one hour each. Sessions took place weekly, although at the end of treatment, namely last two sessions, have two-week intervals. Overall, clients received on average 12.14 sessions (SD = 6.14, range 1-20 sessions). About 50% of the clients attended less than 13 sessions and 75% less than 18 sessions.

The UP consists of five core treatment modules that were designed to target key aspects of emotional processing and regulation of emotional experiences: a) increasing present-focused emotion awareness, b) increasing cognitive flexibility, c) identifying and preventing patterns of emotion avoidance and maladaptive emotion-drive behaviors (EDB's), d) increasing awareness and tolerance of emotional-related physical sensations, and e) interoceptive awareness and situation-based emotion-focused exposure. These modules are preceded by a first one, focused on enhancing motivation and readiness for change and treatment engagement, as well as an introductory module educating clients on the nature of emotions and providing a treatment rationale for understanding their emotional experiences. The last module consists of reviewing progress over treatment and figured out relapse prevention strategies.

2.2. MEASURES

Along the permanence in UP program, different types of information concerning the clients and their clinical condition were collected. Firstly, information about baseline characteristics of the client (e.g., sex and age) was considered. Clients' symptomatology level and therapeutic alliance quality were usually recorded in each session (one per week). For this purpose, the *Outcome Questionnaire, short version (OQ-10.2)* was administered at the beginning of each session in order to evaluate client's symptomatology through therapy, and the *Working Alliance Inventory (WAI)* was administered at the end of each session in order to evaluate therapeutic alliance quality (according client's perspective). Finally, the event that conduced client to abandon the treatment program was checked through Termination Report Form, therapist's notes and clients' records (e.g., e-mail). When no information was available for a given participant, the principal researcher of this work contacted them in order to check out the dropout main motivation. Following, we present the instruments used to gather information.

Diagnostic Interview for Anxiety, Mood, and Obsessive-Compulsive and Related Neuropsychiatric Disorders (**DIAMOND**; Tolin et al., 2018) is a structured interview, used, at intake, to determine eligibility and gather demographic and clinical information about the participants. The demographic variables assessed were age, sex, relationship status, professional status, and education. The clinical variables assessed were diagnosis, the presence of co-mobility, ans medication use at intake.

Beck Depression Inventory II- (**BDI-II**; Beck et al., 1996; Portuguese version by Coelho & Barros, 2002) is a 21 item self-reported scale that assesses the severity of three dimensions of depression: cognitive, affective, and somatic symptoms. Each sentence is rated on a 4-point Linkert scale, ranging from 0 to 3. The total score ranges from 0 to 63, with higher values indicating more severity of depressive symptoms. The Portuguese version of the *BDI-II* was used, which has shown good to excellent validity and reliability in a large number of studies (Campos, 2011). The inventory was administered at the intake as well as at the end of UP treatment.

Beck Anxiety Inventory - (**BAI**; Beck et al., 1988; Portuguese version by Quintão et al., 2013) consists of 21 items commonly used to evaluate clinical anxiety. Each item is rated on a 4-point Linkert scale. The possible range of total score goes from 0 to 63, with higher values indicating higher levels of anxiety. In this study, the Portuguese version of the *BAI* was used, which has shown good psychometric properties, namely adequate reliability values, with Cronbach's alpha of .92. The inventory was administered at the intake as well as at the end of UP treatment.

Outcome Questionnaire - (**OQ-45.2**; Lambert et al., 1996; Portuguese version by Machado & Fassnacht, 2015) consists of a self-report measure that comprises 45 items designed to evaluate therapeutic progress and outcome in three dimensions: subjective discomfort, interpersonal functioning and social role performance. Items are rated on a 5-point Linkert scale, ranging from 0 (never) to 4 (always). Total scores, reflect the client's symptomatology level, ranging from 0 to 180, with higher scores indicating higher impairment. The *OQ-45.2* presents substantial evidence for validity and reliability, as well good internal consistency, with Cronbach's alpha of .89 (Machado & Fassnacht, 2015). The Reliable Change Index (RCI), calculated for the Portuguese population, was 15 points, and the cutoff was 62 points (Machado & Fassnacht, 2015). In this study, OQ-45.2 was used at intake and at the end of UP treatment to assess the presence of clinically significant symptomatology, as also to evaluate treatment progress. Additionally, a shortened version of the outcome questionnaire (OQ-10.2; Lambert et al., 1998) was administrated at the commencement of each follow-up session. The OQ-10.2 consists of 10 items that assess changes in the clients' symptomatic distress over short periods. It presents analogous value of internal consistency (α = .87; Goates-Jones & Hill, 2008).

Working Alliance Inventory - (**WAI**; Horvath & Greenberg, 1989) consists of a self-report measure designed to assess therapeutic alliance's quality between the therapist and the client. The WAI yields a global score as well as a score for each of the three subscales: (1) agreement on tasks; (2) agreement on goals; and (3) development of a bond. In this study, the short version of the client's WAI form was used following each attended session (WAI-SR; translated into Portuguese and adapted by Ramos, 2008). The WAI-SR includes 12 items rated on a 5-point Likert scale, ranging from 1 (seldom) to 5 (always). Higher scores in this measure reflect a better working alliance. Psychometric studies of the Portuguese WAI-SR have shown adequate reliability values, with Cronbach's alpha of .85 for the total scale, .72 for the task subscale, .64 for the bond subscale and .80 for the goals subscale.

Termination Form At the close of therapy, the therapists completed a termination form indicating the length of treatment (i.e., number of therapy sessions attended) and the type of termination. To record the type of termination, therapists chose from among five responses options: "mutual termination "(i.e., both client and therapist agree on the treatment conclusion), "client decision" (i.e., client calls, e-mail or shows for session and announces his/her decision to leave treatment), "client no-show" ,"referral to another treatment or agency", and "other" (i.e., termination is caused by an arbitrary event, like end of graduation or move to another city). In order to reflect both, the logic of survival analysis and the termination classifications found in the literature, these five options were recoded into premature termination (PT) - the event of interest, and censored cases. In this study, we define PT based on the unilateral initiative of the client to abandon or terminate the therapy, after at least one therapeutic session, without the therapist' agreement or knowledge. Consequently, "client decision" and "client no-show" were recoded instances of PT. One the other hand, when information on time to event was not available due to loss to folow-up (given arbitrary or controlled reasons) or non-occurrence of outcome event before the close of the data collection period, client is said to be censored. In this sense, "mutual termination", "referral to another treatment or agency" (which, at this clinic, implied the client had agreed) and "arbitrary end" (i.e., cases whose clinically meaningful endpoint cannot be known because the client left therapy for non-therapy-related reasons) were recoded into censored cases. Similarly, "ongoing cases" - cases that continue through the close of the data collection period - were also recoded into censored cases.

To summarize, regarding survival endpoint we have two groups: PT (the event of interest) and censored cases. Two process variables, namely therapeutic alliance quality and symptomatology level or treatment outcome, in the present study, are the longitudinal responses. Finally, a total of ten baseline clients variables were considered as potential predictors of either hazard to premature termination (regarding survival process) and mean progression of therapeutic alliance, as well as treatment outcome evolution (regarding

longitudinal process). Note that, for separate survival analysis, therapeutic alliance and symptomatology level both measured at last session observed, were tested as predictor variables. Table 1 presents a brief description of the variables for 97 clients from the UP trial.

Table 1: Variables description

	Variable	Code/ Values
At intake	Numeric:	
	Age	Years
	Symptomatology level (OQ-45.2)	Score: 0-180; cutoff = 62; non-clinical $\leq$ 62
	Depressive severity (BDI-II)	Score: 0-63; cutoff = 13; non-clinical $\leq$ 13
	Anxiety level (BAI)	Score: 0-63; cutoff = 21; non-clinical $\leq$ 21
	Categorical:	
	Sex	1 - Male; 2 - Female
	Professional status	1 - Employed; 2 - Unemployed; 3 - Student
	Relationship status	1 - Single; 2 - Married/ in a relationship; 3 - Divorced
	Education level	1 - 9 to 12 years; 2 - University
	Diagnostic area	1 - Anxiety; 2 - Mood; 3 - Others
	Medication	0 - No; 1 - Yes
Follow up sessions	Numeric:	
	Therapeutic alliance quality (WAI)	Score: 0-60
	Symptomatology level (OQ-10.2)	Score: 0-40
	Time	Session: 1-20
	Categorical:	
	Status	0 - Censored; 1 - Premature Termination (PT)

3

STATISTICAL METHODOLOGY

3.1. SURVIVAL ANALYSIS

Survival analysis refers to a class of techniques designed to studying the occurrence of events in a longitudinal framework. A survival outcome is the time, from a defined origin (e.g., time from diagnosis or beginning of treatment) until an event of clinical interest (e.g., recovery, relapse or dropout). Note that survival times T, can be either observed or censored. The latter meaning that observation of the subject in question is terminated before the occurrence of a target event during data collection. The only information available for these subjects is a maximum time T_C, up to which it is known not to have observed the event.[1]

Censoring is non-informative if it is statistically independent of the event of interest (e.g., subject withdraws from the study for reasons not related to her prognosis, such as moved to a different city). Otherwise, we speak of informative censoring, which occurs when non-observation of the subject is due to occurrence, or imminent occurrence, of the event of interest (e.g., subject is about to dropout and, as a result, stops answering the questionnaires during the period of data gathering).

Next, we will present some concepts and notation to describe the distribution of time to the event of interest.

3.1.1 *Notation and Definition of Concepts*

Let T be the non-negative random variable representing the time until the event of interest and let $t_1, ..., t_n$ be the random sample from T_i on $i = 1, ..., n$ subjects. However, for certain individuals we

1 Note that, this is right-censoring. It must be distinguished from left-censoring, where for a subset of the subjects under study, the event of interest is only known to occur before a certain time point; and from interval-censoring, where for a subset of the subjects under study, the event of interest is only known to occur between two certain time points. For all censoring types, we do not know the exact value of T.

will not observe T_i, but rather a realization c_i of a random variable C_i, representing the censored time. Therefore, the observed survival data is the realization $S_i = min\{T_i, C_i\}$, $i = 1, ..., n$ of a random variable S, that is , the minimum between failure time and censored time.

Note that, a common assumption in survival analysis is non-informative censoring, meaning that random variables C and T are independent. Therefore, if $T \leq C$, $S = T$, and a failure time is observed; if $C < T$, $S = C$, and censoring time is observed. To distinguish failure from censored time we consider $\delta_i = I(T_i \leq C_i)$, a censoring indicator that takes value 1 if the observed is the failure time and 0 otherwise.

The random variable T of interest can be described in several ways, by different distributions. Clinical studies often focus on estimating the **survivor function** - the probability of an individual surviving beyond time t (i.e., the probability of being event-free at time t). Formally, for the continuous case,

$$S(t) = P(T > t).$$

Likewise, when considering a discrete distribution of the random variable T, the discrete survivor function for individual i at time j is the probability that individual i will survive beyond time j, that is,

$$S(t_{ij}) = P(T_i > t_{ij}).$$

Two key properties of this function are that $S(0) = 1$, that is, at the beginning of the study, when $t = 0$, the event has not yet occurred for any subjects; and $\lim_{t\to\infty} S(t) = 0$, which means that everyone will experience, at some moment, the event. However, in practice, the latter assumption may not be required, given the restricted follow-up period.

Another way to describe the distribution of the time event is the **hazard function** - the conditional probability that the event occurs in a given short period of time, given that it did not occur earlier. So, in the general continuous case the hazard function, $h(t)$, is defined as,

$$h(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t},$$

where the numerator is a conditional probability. By dividing this probability by Δt, we transform it to a rate. And, by taking the limit as the width of the interval (i.e, Δt) becomes infinitesimally small, we are obtaining the *instantaneous rate* at which the event occurs at a given time. The hazard function magnitude summarizes exactly the right answer to - whether and, if so, when events occur (i.e., the risk of event occurrence in each period) (Willett & Singes, 1993). A basic property is that, the hazard function in the continuous case is never negative, and can vary from zero to infinity.

For the discrete case, the hazard function can be defined as,

$$h(t_{ij}) = P(T_i = t_{ij} | T_i \geq t_{ij}),$$

where, once again, the hazard function $h(t_{ij})$ is the conditional probability of the event of interest occurring to the ith individual in the jth interval, given that it has not occurred previously. In this case, each interval is 1 time unit.

3.1.2 *Non-Parametric Survival Models*

The most well-known estimator of the survival function, called the **product-limit estimator**, has been proposed by Kaplan & Meier (1958). This is a non-parametric estimator, widely used in survival analysis, that does not make any assumptions for the underlying distribution of the event times. To introduce this estimator, let $t_1, ..., t_k$ denote the unique event times in a given sample. Using the law of total probability, the probability of surviving any time point t can be written as the product of the conditional probabilities:

$$P(T > t) = P(T > t | T > t-1) \times P(T > t-1 | T > t-2) \times ... \times P(...).$$

To estimate survival probabilities at each unique event time, we utilize the above expansion, and in the calculation of the conditional probabilities, we account for censoring by suitable adjusting the number of subjects at risk (i.e., the subjects who have not experienced the event and are not censored), which leads to the product-limit estimator:

$$\hat{S}_{KM}(t) = \prod_{i:t_i \leq t} \frac{r_i - d_i}{r_i}, \tag{1}$$

where r_i denotes the number of subjects still at risk at the unique event t_i, and d_i is the number of events at t_i. This estimate is a step function with jumps at observed event times t_i.

The variance of the product-limit estimator can be determined by Greenwood's formula, given by,

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{r_i(r_i - d_i)}.$$

On the other hand, the non-parametric estimator for the discrete-time hazard function is given by the proportion of subjects entering each time period (i.e., number of subjects at risk) who left the study during that period (i.e., number of events). Formally,

$$\hat{h}(t) = \frac{d_i}{r_i} \tag{2}$$

A hazard function defined this way, since it is a probability, varies between 0 and 1.

As aforementioned, any of the two functions $S(t)$ and $h(t)$ defines uniquely a specific probability distribution for random variable T, and each of them provides the investigator a different view of the data. On top of that, both functions are related, in a way we can deduce one from other. So, the sample hazard function can be used to estimate the sample survivor function indirectly in time periods that censoring precludes its direct computation. The sample survival probability for any time period is just 1 minus the hazard probability for that period multiplied by the sample survival probability from the previous period.

Finally, in order to compare time-to-event between two or more groups of subjects differing for a given characteristic or randomly allocated to different treatments, the most usually non-parametric approach adopted is the **Mantel-Haenzel test** (1959), currently called the **log-rank test**, where the null hypothesis being tested is: no difference between (true) survival curves. Note that, the log-rank test requires the assumption that within each level (i.e., group) the populations are homogeneous in survival experiences. Nevertheless, this assumption is rarely realistic, specially in clinical studies, as populations are made heterogeneous in survival experience by demographic variables and possible risk or prognostic factores (i.e., effect of covariates).

3.1.3 *Semi-Parametric Cox Proportional Hazards Model*

The regression method, introduced by Cox (1972), is nowadays widely used to adjust for, as well as to assess the effects of several covariables simultaneously on the hazard rate. This method is known as the *Cox Proportional Hazards Model (CPHM)*, given by:

$$h(t|\boldsymbol{x}_i) = h_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i) \tag{3}$$

where $h_o(t)$ is the baseline hazard rate, the exponential function $exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)$ is the relative risk function, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iq})$ represents a vector of q covariates mesured at baseline to predict event time, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of regression parameters to estimate. Basically, the Cox model formula says that the hazard of experiencing an event for each individual at time t is the product of two quantities. The first of these, $h_0(t)$, is called the baseline hazard function. The second quantity is the exponential expression to the linear sum of $\boldsymbol{W}^T\boldsymbol{\beta}$, where the sum is over the q explanatory W variables. Note that, when all covariates are equal to zero, the relative risk function is equal to one, such that the hazard rate corresponds to the baseline hazard.

The hazard rate ratio between two individuals, denoted 1 and 2, with covariates vector, $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively, is:

$$\frac{h(t|\boldsymbol{x}_1)}{h(t|\boldsymbol{x}_2)} = \frac{h_0(t)exp(\boldsymbol{\beta}^T\boldsymbol{x}_1)}{h_0(t)exp(\boldsymbol{\beta}^T\boldsymbol{x}_2)} = exp\{(\boldsymbol{x}_1 - \boldsymbol{x}_2)\boldsymbol{\beta}^T, \tag{4}$$

for all $t \geq 0$. Since the baseline hazard function $h_0(t)$ appears in both the numerator and denominator, it is canceled out of the formula, and the final expression does not involve time, t. Thus, one of the crucial assumptions in Cox regression is that the hazard ratio is constant and independent of time, justifying the name of *Proportional Hazard (PH)* model.

If we assume that $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are exactly the same, except for the k^{th} covariate, then equation 4 becomes:

$$\frac{h(t|\boldsymbol{x}_1)}{h(t|\boldsymbol{x}_2)} = exp\{(\boldsymbol{x}_1 - \boldsymbol{x}_2)\boldsymbol{\beta}^T\} = exp(\beta_k), \tag{5}$$

where $exp(\beta_k)$ is the hazard ratio or the relative risk of the k^{th} covariate. Hence, regarding the interpretation of regression coefficients vector β, for a given particular time point t, $exp(\beta_k)$ denotes the relative increase in the risk for an event that results from one unit change in the k^{th} covariate, while the other covariates are kept unchanged.

The CPHM is said to be a semi-parametric model since the baseline hazard function, $h_0(t)$, is a non-parametric (i.e., non specified) component, and the relative risk function, $exp(\boldsymbol{\beta}^T\boldsymbol{x})$ is parametric. To estimate the regression coefficients, Cox (1975) proposed the maximization of partial log-likelihood method, given by the following function:

$$PL(\boldsymbol{\beta}) = \prod_{t_j} \frac{exp(\boldsymbol{\beta}^T\boldsymbol{x}_{i_j})}{\sum_{l \in R_j} exp(\boldsymbol{\beta}^T\boldsymbol{x}_l)}$$

where i_j is the index of the individual who experiences an event at time t_j, $R_j = \{l|Y_l(t_j) = 1\}$ is the risk set of individuals at t_j, and $Y_l(t)$ is the indicator for individual l just before time t. Basically, this function is constructed based on conditional probabilities, that individual i experienced the event, given that someone did, at the set of observed event times.

The maximum partial likelihood estimate for $\hat{\boldsymbol{\beta}}$ is obtained by differentiating the log partial likelihood function, $logPL(\boldsymbol{\beta})$. By turn, a 95% confidence interval of the relative risk (equation 5) can be obtained by exponentiating the lower and upper limits of the standard 95% confidence interval for the regression coefficient, $\hat{\beta}_k \pm 1.96SE(\hat{\beta})$. For more details on parameter estimates in the partial likelihood see (Klein & Moeschberger, 2011).

To conclude, the main advantage of CPHM is that the estimation of the regression coefficients does not depend on the baseline hazard function $h_0(.)$, since it gets absorbed when the coefficients are estimate by the method of partial log likelihood (i.e., it is nowhere to be seen in the log likelihood calculation). Hence, the shape of the baseline hazard is irrelevant and there is no need to assume a known distribution (Klein & Moeschberger, 2011).

3.1.4 *Time-Dependent Cox Model*

In the relative risk model previously introduced, the covariates are assumed to be constant in time, as typically occurs when treatment, sex, age, and clinical features at study entry are considered. However, in many studies it may also be of interest to investigate whether time-dependent covariates are associated with risk of an event.

An additional advantage of the Cox regression model (Cox, 1972) is that, besides it allows for fixed covariates that do not change over time, it can also incorporate time-dependent covariates. So, the extended model is given by:

$$h_i(t|\boldsymbol{y_i}(t), \boldsymbol{x}_i) = h_0(t) \exp\{\boldsymbol{\beta}^T \boldsymbol{x}_i + \boldsymbol{\zeta} \boldsymbol{y}_i(t)\} \tag{6}$$

where, as in section 3, $\boldsymbol{x}_i$ denotes a vector of baseline covariates, and $\boldsymbol{y}_i(t)$ denotes a vector of time-dependent covariates. In this sense, the hazard h for person i at time t to experience the event of interest is a product of a baseline hazard function $h_0(t)$, which has no particular parametric form, and the exponential of a linear combination of explanatory variables, some of which can be functions of time. Estimation of β and ζ is also based on the maximization of partial log-likelihood method (for more details see Rizopoulos).

The interpertation of the regression coefficients vector $\boldsymbol{\zeta}$ is exactly the same as for $\boldsymbol{\beta}$. Namely, if we assume for simplicity that there is only a single time-varying covariate, then at any particular time point t, $exp(\zeta)$ denotes the relative increase in the risk for an event at time t that results from one unit increase in $y_i(t)$ at the same time point. Moreover, given that $y_i(t)$ is time-varying, model 6 no longer assumes that the hazard ratio is constant over time Rizopoulos.

To the correct use of this (extended) Cox model, it is crucial to distinguish between two different categories of time-dependent covariates, namely, *external* or *exogenous* covariates and *internal* or *endogenous* covariates (Diggle et al., 2002). The reason why it is important to distinguish between then is that and endogenous covariate requires special treatment compared to an exogenous one.

An exogenous covariate exists/develops independently of the survival of a subject in the study. A standard example for an external covariate is the time of the day or the season of the year, or yet environmental factors like air pollution. The value that an external covariate takes in time are not influenced by the life experience of the subject, since the value are generated by a mechanism which is external to the individual. Conversely, an endogenous covariate only exists/can be recorded as long as the participant is alive, as instance blood pressure measured over time. The value of an endogenous covariate at time t carries information about the life experience of the individual up to that time. Hence, the first important characteristic of endogenous covariates is that typically require the survival of the subject for their existence. Another feature of endogenous covariates is that they are typically measured with error. This measurement error primarily refers to the biological variation induced by the patient herself rather than to the error induced by the procedure that determines the value of covariate. In particular, measuring the same client twice, even on the same day, we do not expect to observe exactly the same value for an endogenous covariate. Thus, for such covariates, it would be more reasonable to assume that the observed marker levels are actually a contaminated with biological variation version of the true marker levels. The final important implication with endogenous covariates is that their complete path up to any time t is not fully observed. That is, the levels of a biomarker or any other clinical parameter for a client are only known for the specific occasions that this patient visited the study center to provide measurements, and not in between these visit times (Rizopoulos, 2012).

The Cox proportional hazards model can be extended to handle exogenous time-dependent covariates, but it is not appropriate when the time-dependent covariates are of endogenous nature. This is because the extenden Cox model assumes that time-dependent covariates are predictable processes, measured without error, and have their complete path fully specified (Rizopoulos, 2012).

3.1.5 *Model Diagnostic Procedures*

After a model has been fitted to an observed set of survival data, the adequacy of the model needs to be assessed. Many model-checking procedures are based on quantities known as residuals. These are values that can be calculated for each individual in the study, and have the feature that their behaviour is known, at least approximately, when the fitted model is satisfactory (Kleinbaum & Klein, 2012).

The residual that is most widely used in the analysis of survival data is the **Cox-Snell residual** (Cox & Snell, 1968), given by:

$$r_i^{CS} = \hat{H}_0(t_i) exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i),$$

where $\hat{H}_0(t_i)$ is an estimate of the baseline cumulative hazard function at time t_i, with $\hat{H}_0(t_i) = -log\hat{S}_i(t_i)$. If the model fitted to the observed data is satisfactory, the residuals should approximately have an exponential distribution with mean one (Collett, 2015). This can be checked using an exponential Quantile-Quantile (QQ) plot.

A key assumption of the Cox model is proportional hazards (PH). That is, with time-fixed or constant covariates, the relative hazard for any two subjects obeys the relationship presented previously in equation 4, where the proportionality constant is independent of time. In contrast, if there are one or more explanatory variables in the model whose coefficients vary with time (i.e., time-dependent covariates), the proportional hazards assumption will be violated. In order to check this assumption, for time-fixed variables that have a small number of levels, a simple graphical test can be made by looking at the survival curves. If the Kaplan-Meier curves cross for two or more levels of a predictor of interest, then the PH assumption is not met; in contrast, parallel curves indicate that the PH assumption is satisfied (Kleinbaum & Klein, 2012). Another graphical option could be to use the **Schoenfeld residuals**, which represent the difference between the observed covariate and expected values, given the risk set at that time. In this sense, whenever we represent them ranked by its event time, they should be flat and centered around zero. Otherwise, a plot that shows a non-random pattern against time is evidence of violation of the PH assumption. Schoenfeld (1982) proposed a chi-squared goodness-of-fit test statistic for the proportional hazards regression model, where a correlation of zero (the null hypothesis) indicates that the model met the proportional assumption (i.e., the residuals are independent of time). In the case of a violated proportional hazard assumption, results of the Cox's model cannot be trusted. Nevertheless, determining which factors have time-varying effects can be quite useful in itself by gaining insights into the data, and, by turn, the extended models of the Cox regression are required.

3.1.6 *Conclusion*

Survival analysis has several advantages when applied to the study of psychotherapy termination (Corning & Malofeeva, 2004). First, this approach, offers a time-based conceptualization and analysis of the data, appropriate to the time-oriented nature of psychotherapy, which recognizes the population heterogeneity in survival experience given demographic variables and possible risk or prognostic factors. Second,

survival analysis allows for adequate handling of cases for which the event of interest is not observed over the study's data collection period. These cases are called censored, and they usually arise as a result of the therapy ending, or the therapy continuing past the data collection period for some clients, as well as an end for arbitrary reasons (i.e., when a external circumstances, as instance move to a different city, cause a disruption in the observation of an entire risk period). Third, when modelling survival data, a breadth of predictors can be incorporated into statistical models, including not only time-invariant or static variables (i.e., that vary across individuals but not over time), as also time-variant predictors (i.e., that vary over time). However, as aforementioned, a drawback of the CPHM is that it's not account for measurement error in the endogenous covariables. In fact, the Cox model with internal time-dependent variables is sometimes misused, and at list, considerable care must be taken in interpreting the results of a model including such covariates (Rizopoulos, 2012). In section 3.3, we will introduce a modelling framework especially designed to account for the special features of endogenous time-dependent covariates.

3.2. LONGITUDINAL ANALYSIS

Longitudinal data results from the observation of subjects that are measured repeatedly over time on one or more response variables. The main feature of longitudinal studies is that they permit the direct assessment of change in the response variable over time, distinguishing differences among subjects in their baseline levels (cohort effects) from changes over time within subjects (aging effects) (Diggle et al., 2002).

For the analysis of repeated measurements it is common to assume independence between subjects. However, this assumption is not adequate for measurements within the same subject, once we expect a positive correlation (i.e., intraindividual correlation). Ignoring correlation in longitudinal data, by the inappropriate application of standard statistical tools (e.g., *t-test* and simple linear regression), could lead to incorrect inferences or inefficient estimates of the regression coefficients, as also to the sub-optimal protection against bias causes by missing data (Diggle et al., 2002). For this reason, correlation structure of data takes a prominent role to estimate regression parameters in longitudinal analysis (Fitzmaurice et al., 2004).

Statistically, there are three potential sources of variability that have an impact on the correlation among repeated measures: (1) between-subjects heterogeneity, (2) within-subjects biological variations, and (3) measurement error (i.e., the remaining random error) (Fitzmaurice et al., 2004). Between-subjects heterogeneity (1) reflects natural variation in individuals' propensity to respond, since some individuals con-

sistently respond higher than the average, while others consistently respond below the average. Basically, each individual has his or her own subject-specific propensity to respond, which drives from unobservable genetic, biological, environmental, social, or behavioral factors (or some combination of these factors), and is shared by all of the repeated measures obtained on that individual. As a result a pair of repeated measures on the same individual will be expected to be more similar than single observations obtained from two randomly selected individuals. A popular approach to handle between-subjects variability is to specify the individual-specific "random effects" (e.g., randomly varying intercepts and slopes) with the assumption of a known distribution.

The second component of variability in longitudinal data is the within-subject variation. Considering the inherent within-individual biological variability, each sequence of repeated measurements on any individual might vary randomly around their long-run average (or "true" underlying biological process). Consequently, random deviations or departures from an individual's underlying response trajectory are expected to be more similar than those obtained from several randomly selected individual (i.e., there is a subject-specific dependence). Besides that, measurements (on the same individual) taken very closely together will typically be more highly correlated than measurements that are further separated in time. So, in recognizing patterns of correlation (i.e., serial correlation), researchers can account for intraindividual correlation by specifying within-subject covariance structures on repeated measurements.

A final source of variability in longitudinal data is random measurement error. This is an ubiquitous component of almost all studies, longitudinal or not, and results from imprecision of the measurement procedure. As regularly specified in general linear and generalized linear regression models, this random term for uncertainty can be estimated as regression residuals (Liu, 2016).

Next, we will introduce some vector and matrix notation, as well as present a general linear regression model for longitudinal data.

3.2.1 *Notation and Definition of Concepts*

In assuming that N subjects are measured repeatedly over time, we let Y_{ij} represent a response variable observed at time t_{ij}, for observation $j = 1, ..., n_i$ on subject $i = 1, ..., m$. The set of repeated outcomes for the i^{th} subject can be grouped into a $n_i \times 1$ vector,

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}, \qquad i = 1, \dots, m;\ j = 1, \dots n_{i_m}.$$

Associated with each response, Y_{ij} , there is a $p \times 1$ vector of explanatory variables, given by:

$$\boldsymbol{x}_{ij} = \begin{pmatrix} x_{ij1} \\ x_{ij2} \\ \vdots \\ x_{ijp} \end{pmatrix}, \qquad i = 1, \dots, m;\ j = 1, \dots, n_i.$$

The vectors of explanatory variables can be grouped into a $n_i \times p$ matrix denoted by $\boldsymbol{X}_i$, where:

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}_{i1}^T \\ \boldsymbol{x}_{i2}^T \\ \vdots \\ \boldsymbol{x}_{in_i}^T \end{pmatrix} = \begin{pmatrix} x_{i11} & x_{i12} & \dots & x_{i1p} \\ x_{i21} & x_{i22} & \dots & x_{i2p} \\ \dots & \dots & \ddots & \dots \\ x_{in_i1} & x_{in_i2} & \dots & x_{in_ip} \end{pmatrix}, \qquad i = 1, \dots, m,$$

Thus, the matrix $\boldsymbol{X}_i$ is simply an ordered collection of the values of the p explanatory variables for the i^{th} subject at each of the n_i measurement occasions, where the rows correspond to the explanatory variables associated with the responses at the n_i different measurement occasions, and the columns correspond to the p distinct explanatory variables.

The explanatory variables can be measured at baseline or can be time dependent. In the former case, the same values of the explanatory variables are replicated in the corresponding rows of $\boldsymbol{X}_i$. In the latter case, the values taken by the explanatory variables can vary over time (for at least some individuals) and the values in the corresponding rows of $\boldsymbol{X}_i$ can be different at each measurement occasion.

Most longitudinal analysis are based on a regression model such as the linear model,

$$Y_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \dots + \beta_p x_{ijp} + e_{ij}, \qquad i = 1, \dots, m;\ j = 1, \dots, n_i,$$

where $\beta_1, \dots, \beta_p$ are the unknown regression parameters relating the mean of Y_{ij} to its corresponding explanatory variables, and e_{ij} is a zero-mean random variable which represents deviations of the responses

from their corresponding predicted means. In using vector and matrix notation, the regression model can be expressed in an even more compact form:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{e}_i \tag{7}$$

where $\boldsymbol{X}_i$ is a $n_i \times p$ matrix, with the vector x_{ij} in the j^{th} row, and $\boldsymbol{e}_i = (e_{i1}, ..., e_{in_i})$. The vector of continuous response, $\boldsymbol{Y}_i$, is assumed to have a conditional distribution [2] (i.e., the mean of the longitudinal response vector is related to the explanatory variables via the linear regression model given above in equation 7) , that is multivariate normal, with mean response vector:

$$E(\boldsymbol{Y}_i|\boldsymbol{X}_i) = \boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\beta}$$

and covariance matrix,

$$\boldsymbol{\Sigma}_i = Cov(\boldsymbol{Y}_i|\boldsymbol{X}_i).$$

The multivariate normal distribution is completely specified by the vector of means, $\boldsymbol{\mu}_i$, and the covariance matrix, $\boldsymbol{\Sigma}_i$.

Finally, in order to obtain a complete specification of the model for response variable $\boldsymbol{Y}_i$, we need model the covariance structure. As we notice before, repeated observations on the same individual are not independent. In fact, this is an advantage, given that correlated observations provide more precise estimates of the rate of change than would be obtained from an equal number of independent observations of different individuals. Thus, although the correlation, or more generally, the covariance among the repeated responses, is not usually of intrinsic interest, it must be properly accounted in order to yield valid inferences about the regression parameters of primary interest.

According to Fitzmaurice et al. (2004), three broad approaches to modeling the covariance can be distinguished: (1) unstructured covariance, (2) covariance pattern models, and (3) random effects covariance structures. The first, also called fully parametrized, allows all of the parameters of the variance-covariance matrix to be different (see unstructured covariance in Table 2). As no explicit structure is assumed for

2 For simplicity of notation, we often replace $E(Y_i|X_i)$ by $E(Y_i)$ and $Cov(Y_i|X_i)$ by $Cov(Y_i)$. However, it should be clear from the context that it denotes, respectively, the conditional mean of the responses and the conditional covariance of the responses, given the explanatory variables. In a similar vein, in discussing the distribution of Y_i, it should be understood that we are always referring to the conditional distribution of Y_i given the explanatory variables.

the covariance among repeated measures, there are $q = n(n+1)/2$ unique parameters. A favorable aspect of unstructured covariance is the lack of assumptions about them, since they can take any form. However, the number of covariance parameters can be quite large, until larger than sample size, which potentially produces unstable estimates. Moreover, incomplete data across time are unallowable under this approach (i.e., it is only applicable when all individuals are measured at the same set of occasions).

An alternative approach are covariance pattern models, which place a structure on covariance matrix. Taking into account any trend in the dispersion matrix, the correlation among repeated measures is expressed as an explicit function of the time lag. Then, the covariance matrix can be adequately described with only a few parameters (the number of parameters depends on matrix form or structure). There are several possible forms for Σ. A simple form is that of compound symetry (as shown in Table 2), which specifies equal variances and equal covariances. Notice that, in this case, σ^2 is the variance of dependent variable at every time point, and the covariance equals to ρ for the pairwise association of the dependent variable for any two time points. Consequently, the number of variance-covariance parameters are $q = 2$. Another form that only depends on two parameters is the first-order autoregressive (AR1) structure (see Table 2). In this case, the covariance for time points j and j' equals

$$\sigma_{jj'} = \sigma^2 \rho^{|j-j'|},$$

where ρ is the AR(1) parameter and σ^2 is the error variance. Note that, in this case, the correlation decreases exponentially as the lag between the time points increases. A potential drawback within covariance pattern models is that they depend upon a reduced number of parameters to convey accurately the information of the covariance matrix. On the other hand, if the model assumptions hold, the covariance is described parsimoniously by a small number of parameters.

An alternative strategy for imposing structure on the covariance is through the introduction of random effects. The random effects models assumes the correlation between repeated measurements arises because each subject has an underlying (or latent) level of response that persists over time and influences all repeated measurements on that subject. This individual-specific effect is regarded as a random variable. A more detailed account will be given in the next section (3.2.2).

Table 2: The three most common covariance structures

Unstructured	Compound Symmetry	First-order Autoregressive
$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \dots & \sigma_{1n} \\ & \sigma_2^2 & \sigma_{21} & \dots & \sigma_{2n} \\ & & \sigma_3^2 & \dots & \sigma_{3n} \\ & & & \ddots & \dots \\ & & & & \sigma_n^2 \end{pmatrix}$	$\sigma^2.\begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ & 1 & \rho & \dots & \rho \\ & & 1 & \dots & \rho \\ & & & \ddots & \dots \\ & & & & 1 \end{pmatrix}$	$\sigma^2.\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots & \rho^{n-1} \\ & 1 & \rho & \rho^2 & \dots & \rho^{n-2} \\ & & 1 & \rho & \dots & \rho^{n-3} \\ & & & 1 & \dots & \rho^{n-4} \\ & & & & \ddots & \dots \\ & & & & & 1 \end{pmatrix}$

Note. As all matrices are symmetric, only their upper triangles are shown. Greek letters represent unknown parameters. The parameter ρ satisfies $|\rho| \leq 1$.

3.2.2 *Linear Mixed Effects Models*

The most widely used class of models for repeated measurement data is the **linear mixed effects models**, that incorporate both fixed effects and random effects. According to this approach the mean response is modeled as a combination of the population characteristics, that are assumed to be shared by all individuals (i.e., fixed effects), and subject-specific effects that are unique to a particular individual and vary randomly from one individual to another (i.e., random effects). In this sense, individuals in the population are assumed to have their own subject-specific mean response trajectories over time and a subset of the regression parameters (e.g., the intercept and slope) are regarded as being random (Diggle et al., 2002; Fitzmaurice et al., 2004).

The general linear mixed effects model is defined as:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{8}$$

where $\boldsymbol{Y}_i$ is the n_i-dimensional response vector for subject i, $1 \leq i \leq m$, m is the number of subjects, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are $(n_i \times p)$ and $(n_i \times q)$ dimensional matrices of known covariates, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of fixed effects, $\boldsymbol{b}_i$ is a $(q \times 1)$ vector of random effects, and $\boldsymbol{\epsilon}_i$ is an n_i-dimensional vector of residual components. Note that, $\boldsymbol{Z}_i$ is a known design matrix linking the vector of random effects $\boldsymbol{b}_i$ to $\boldsymbol{Y}_i$, and the

columns of $\boldsymbol{Z}_i$ are a subset of the columns of $\boldsymbol{X}_i$. The random components are assumed to be normally distributed with zero expectation and the following properties:

$$\begin{cases} \boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D}), \\ \boldsymbol{\epsilon}_{ij} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_i), \\ b_1, \ldots, b_m, \epsilon_1, \ldots, \epsilon_m \quad \text{independent}, \end{cases}$$

where $\boldsymbol{D}$ and $\boldsymbol{\Sigma}_i$ are the variance-covariance matrices for the random effects across subjects and for the within-subject random errors, respectively (Fitzmaurice et al., 2004; Liu, 2016).

The mean response trajectory over time for any individual can be described as

$$E[\boldsymbol{Y}_i|\boldsymbol{b}_i] = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i,$$

whereas the mean response profile in the population is given by

$$\begin{aligned} E(\boldsymbol{Y}_i) &= \boldsymbol{\mu}_i \\ &= E[\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i] \\ &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\, E[\boldsymbol{b}_i] \\ &= \boldsymbol{X}_i\boldsymbol{\beta}, \end{aligned}$$

since $E[\boldsymbol{b}_i] = \boldsymbol{0}$. The former is referred as the conditional mean of Y_i, given the subject-specific effect, and the latter is referred as the marginal mean of Y_i, where the averaging is over all individuals in the population (i.e., averaged over the distribution of the subject-specific effects)(Fitzmaurice et al., 2004). In a similar way, we can distinguish between conditional and marginal covariances. The conditional covariance of $\boldsymbol{Y}_i$, given $\boldsymbol{b}_i$, is:

$$Cov(\boldsymbol{Y}_i|\boldsymbol{b}_i) = Cov(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_i,$$

which describes the covariance among the longitudinal observations when focusing on the conditional mean response profile of a specific individual (i.e., the covariance of the i^{th} individual's deviations from his or her mean response profile). Note that, these deviations are positive and negative, and vary randomly about zero. Besides that, it is usually assumed that Σ_i is a diagonal matrix , $\sigma^2_\epsilon \boldsymbol{I}_{ni}$, where $\boldsymbol{I}_{ni}$ denotes an $n_i \times n_i$ identity matrix [3], and measurements errors are independently distributed with a common variance σ^2_ϵ.

3 Recall that, in the identity matrix, the diagonal elements are all 1 and the off-diagonal elements are 0.

On the other hand, the marginal covariance of $\boldsymbol{Y}_i$, averaged over the distribution of $\boldsymbol{b}_i$, is:

$$\begin{aligned} Cov(\boldsymbol{Y}_i) &= Cov(\boldsymbol{Z}_i\boldsymbol{b}_i) + Cov(\boldsymbol{\epsilon}_i) \\ &= \boldsymbol{Z}_i Cov(\boldsymbol{b}_i)\boldsymbol{Z}' + Cov(\boldsymbol{\epsilon}_i) \\ &= \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{\Sigma}_i \\ &= \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \sigma_\epsilon^2\boldsymbol{I}_{ni}, \end{aligned}$$

where the covariance matrix, $Cov(\boldsymbol{Y}_i)$, will, in general, have non-zero off-diagonal elements, thereby accounting for the correlation among repeated observations on the same individual. Thus, the introduction of random effects, $\boldsymbol{b}_i$, can be seen to induce correlation among repeated measurements of $\boldsymbol{Y}_i$. A particular property of the linear mixed effects models is that $Cov(\boldsymbol{Y}_i)$ has been described in terms of a set of covariance parameters, some defining the matrix $\boldsymbol{D}$ and some defining the matrix $\boldsymbol{\Sigma}_i$, which allows the explicit analysis of between-subject ($\boldsymbol{D}$) and within-subject ($\boldsymbol{\Sigma}_i$) sources of variation in the responses. In addition, the marginal covariance of $\boldsymbol{Y}_i$ can be expressed as an explicit function of the times of measurement (when times of measurement, or functions of time, are included in $\boldsymbol{Z}_i$), and by turn each individual can have a unique sequence of measurement times.

Given the specification of the variance/covarince components, the total variance of $\boldsymbol{Y}_i$, denoted $\boldsymbol{V}_i$, is given by:

$$\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}'_i + \boldsymbol{\Sigma}_i,$$

where the off-diagonal elements in $\boldsymbol{V}_i$ reflect dependence of the repeated measurements of the responses $\boldsymbol{Y}_i$. Therefore, $\boldsymbol{Y}_i$ can be expressed as:

$$\boldsymbol{Y_i} \sim MVN(\boldsymbol{X_i\beta}, \boldsymbol{V_i} = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{\Sigma}_i). \tag{9}$$

Diggle et al. (2002) proposed a useful model which can be viewed as an extension of the general linear mixed model, and where the covariance assumptions are relaxed by allowing an appropriate, more general, residual covariance structure $\boldsymbol{\Sigma}_i$ for the vector $\boldsymbol{\epsilon}_i$ of subject-specific error components. They propose to decompose the random term ϵ_i into two components in an additive way,

$$\boldsymbol{\epsilon}_i = \boldsymbol{\epsilon}_{(1)i} + \boldsymbol{\epsilon}_{(2)i}$$

where $\boldsymbol{\epsilon}_{(1)i}$ is a component of measurement process itself and $\boldsymbol{\epsilon}_{(2)i}$ is a component of serial correlation, suggesting that at least part of an individual's observed profile is a response to time-varying stochastic processes operating within that individual. Thus, Diggle et al. (2002) suggest using a single subject effect

(b_i), setting $Var(\epsilon_{(1)i}) = \tau^2 I_{n_i}$, and introducing a third random error vector, $\epsilon_{(2)i}$, which has a serial correlation structure.

The resulting linear mixed model can be written as:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_{(1)i} + \boldsymbol{\epsilon}_{(2)i} \tag{10}$$

$$\begin{cases} \boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D}), \\ \boldsymbol{\epsilon}_{(1)i} \sim N(\boldsymbol{0}, \tau^2 \boldsymbol{I}_{ni}), \\ \boldsymbol{\epsilon}_{(2)i} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{H}_i), \\ b_1, \dots, b_m, \epsilon_{(1)m}, \varepsilon_{(2)1}, \dots, \epsilon_{(2)m} \quad \text{independent}, \end{cases}$$

and the model is completed by assuming a specific structure for the $(n_i \times n_i)$ correlation matrix $\boldsymbol{H}_i$. Specifically, the random serial effect $\epsilon_{(2)i}$ results in a correlation between serial measurements, which depends only on the distance or time between the observations. The serial correlation matrix $\boldsymbol{H}_i$ then only depends on i through the number n_i of observations and through the time points t_{ij} at which measurements are taken. Diggle et al. (2002) suggest several parametric forms for the correlations which model it as a decreasing function of increasing time. Two frequently used are the exponential and Gaussian serial correlation functions which are shown in figure 2.

Given the additive formulation proposed by Diggle et al. (2002), $\boldsymbol{Y}_i$ can now be expressed as:

$$\boldsymbol{Y}_i \sim MVN(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \sigma^2\boldsymbol{H_i} + \tau^2\boldsymbol{I_i}). \tag{11}$$

where $\boldsymbol{H}_i$ is a matrix with (j,k) elements $h_{ijk} = \rho(|t_{ik} - t_{ij}|)$.

One very appealing aspect of this general linear mixed effects model is its flexibility in accommodating any degree of imbalance in longitudinal data, coupled with its ability to account for the covariance in a relative parsimonious way. In fact, it allows one to model correlation and variance with only few parameters, event when there are many different observation times, using only one random effect. Additionally, the model assumes a continuous outcome variable which is linearly related to a set of explanatory variables, and it expands on the ordinary linear regression model by incorporating lack of independence between repeated observations and more than one error term (Cnaan & Laird, 1997).

Under the assumption that the $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ are independently distributed as multivariate normal, and $\boldsymbol{Y}_i \sim MVN(X_i\beta, \sigma^2 V(\alpha))$, estimation of the parameters by maximum likelihood (ML) is straightfor-

ward. Thus, in the present study, we will consider simultaneous estimation of the parameters of interest, $\boldsymbol{\beta}$, and of the covariance parameters, σ^2 and α, using likelihood function for observed data, as given by:

$$L(\boldsymbol{\beta}, \sigma^2, \alpha) = -0.5\{nm\, log(\sigma^2) + m\, log(|\boldsymbol{V}(\alpha)|) + \sigma^{-2}(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})\} \quad (12)$$

For given α, the maximum likelihood estimator for $\boldsymbol{\beta}$ is the weighted last-squares estimator (Diggle et al., 2002), given by:

$$\hat{\boldsymbol{\beta}}(\alpha) = (\boldsymbol{X}'\boldsymbol{V}(\alpha)^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}(\alpha)^{-1}\boldsymbol{y}. \quad (13)$$

Substituting 13 in 12 we will have:

$$L(\boldsymbol{\beta}(\alpha), \sigma^2, \alpha) = -0.5\{nm\, log(\sigma^2) + m\, log(|\boldsymbol{V}(\alpha)|) + \sigma^{-2}RSS(\alpha)\}, \quad (14)$$

where the residual sum of squares (RSS) is given by,

$$RSS(\alpha) = \{y - X\hat{\beta}(\alpha)\}'V(\alpha)^{-1}\{y - X\hat{\beta}(\alpha)\}.$$

The maximum likelihood estimator for σ^2 is obtained differentiating 14 with respect to σ^2, for given α, as:

$$\hat{\sigma}^2(\alpha) = RSS(\alpha)/(nm),$$

where $N = \sum_{i=1}^{m} n_i$ is the total number of measurements on all m units.

Finally, the maximum likelihood estimate of α maximizes

$$L(\alpha) = -0.5[N\ \ log\{RSS(\alpha)\} + \sum_{i=1}^{m} log|V(\alpha)|].$$

Regarding the interpretation of the parameters, the *fixed effects* throughout parameter vector $\boldsymbol{\beta}$ describe patterns of change in the mean response over time (and their relation to covariates) in the population of interest. The interpretation is exactly the same as in a simple linear regression model. So, assuming we have p covariates in the design matrix, the coefficient $\boldsymbol{\beta}_j$, where $j = 1, \ldots, p$, denotes the change in the average $\boldsymbol{Y}_i$ when the corresponding covariate x_j is increased by one unit, while all other predictors are held constant. On the other hand, the random effects throughout parameter vector $\boldsymbol{b}_i$ describe how the trend over time for the i^{th} individual deviates from the population average. So, when combined with the fixed effects, $\boldsymbol{b}_i$ describes the mean response trajectory over time for any individual, which varies randomly

from one individual to another. Finally, the random errors are denoted by $\boldsymbol{\epsilon}_i$ and represent the deviation of $\boldsymbol{Y}_i$ from the subject-specific mean response, $\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}S_i$. Note that there is a change in notation from measurement or sampling errors in model (7), where e_i represents the deviation of Y_i from the mean response in the population. So, the random error in previous longitudinal regression model (7) have now been decomposed into two random components, $e_i = b_i + \epsilon_i$, a between-subject component and a within-subject component, as specified in model (8) (Fitzmaurice et al., 2004). Recall that, in general linear mixed model proposed by Diggle et al. (2002), the within-subject component, by turn, have been decomposed into two random components, $\epsilon_{(1)i}$ and $\epsilon_{(2)i}$, representing the measurement error and serial correlation, respectively (model 10).

3.2.3 *Variogram*

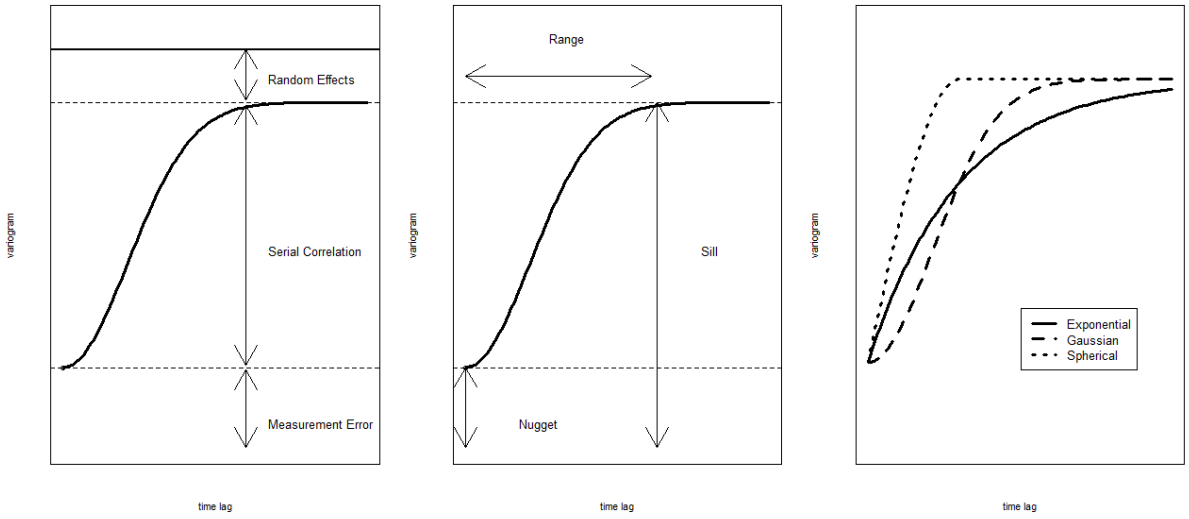
When analyzing longitudinal data, it is often essential to explore the covariance structure of the data before deciding which models might be most appropriate. The variogram (VG) is a helpful graphical tool, which displays the variability between data points as a function of distance or time lag (Diggle, 2005). So, for longitudinal data, the VG is defined as one-half the expected squared difference between residuals obtained on the same individual. The VG, denoted as $\gamma(h_{ijk})$, is given by:

$$\gamma(h_{ijk}) = \frac{1}{2}E(r_{ij} - r_{ik})^2, \tag{15}$$

where (h_{ijk}) is the time elapsed between the j^{th} and k^{th} repeated measurement on the i^{th} individual (Fitzmaurice et al., 2004).

The empirical or sample variogram, $\hat{\gamma}(h)$, is simply defined as one-half the average squared difference between pairs of residuals on the same individual whose corresponding observations are h units apart and average is taken over all pairs of observations for which $h_{ijk} = h$ (Fitzmaurice et al., 2004). One practical advantage of the variogram over, for instance, the covariance function, is that estimation from observed data is more straightforward, especially when the underlying stochastic process is observed at irregular time-points. So, with unbalanced longitudinal data, the empirical VG can be easily estimated by fitting a smooth curve to the scatterplot of the observed half squared differences between residuals from the same individual and the corresponding time lags (also called variogram cloud) (Diggle, 2005).

Three sources of variability - random effects (or between-client variance), serial correlation (or whitin-client variance), and measurement error (or white noise) - can be broken out with this graphical tool,



A. Decomposition of total variability.

B. Parameters of the variogram.

C. The three most common theorical variogram models.

Figure 2: Example of a theorical variogram

as displayed in Figure 2(A). Thus, the upper horizontal line represents the total variability present in the model. The curved line is actually the representation of the theorical variogram function; consequently, the amplitude of this curve represents the variability within the client (or serial correlation). The distance between the maximum of the curve and the upper horizontal line represents the variability between clients (or random effects). Finally, the distance between zero and the minimum value of the variogram line is the variability not explained in the model (white noise or measurement error) (Diggle et al., 2002).

Figure 2(B), by turn, shows the parameters of the variogram. Range indicates the lag time beyond observations appear independent (i.e., variance no longer increases); sill is the value that the variogram attains at the range (the value on the y-axis). From this point forward, one assumes that there is no more time dependence, as variance of the differences between the pairs of residuals becomes constant with time lag. The partial sill is the sill minus the nugget effect. Nugget represents the measurement error and it is estimated from the variogram as the value of $\gamma(h)$ for $h = 0$ (Diggle et al., 2002).

In order to select the appropriate statistical model, it is necessary to adjust the mathematical model (i.e., theorical variogram) which better represents the trend of the empirical variogram with relation to time lag. So, based on the shape of the variogram, we can select the parameterisation for the correlation function in the model which better fit the data. Several theoretical models are used to adjust the variogram including exponential, gaussian and spherical (see Figure 2 (c)) (Diggle et al., 2002).

3.2.4 *Model Diagnostic Procedures*

Once a linear mixed model is fitted, regression diagnostics are necessary in order to verify whether the statistical model fits data appropriately or any systematic discrepancies comes to light. According Diggle et al. (2002), simple plots can be a very effective tool in revealing inconsistencies between data and model, regarding both the mean structure (the fixed effects) and the covariance structure (the random effects and residual error). In this sense, one can: i) **superimpose the fitted mean response profiles on a time-plot of the average observed response within each combination of treatment and time**; and, ii) **superimpose the fitted variogram on a plot of the empirical variogram**. This procedures will allow the validation of mean structure and correlation structure, respectively, by graphical comparison.

Residuals, defined as the differences between the observed and fitted values of the response, have long been used for graphical and numerical examinations of the adequacy of standard regression models with independent observations. In the longitudinal setting, regression diagnostics are more complex because an individual's response is measured repeatedly over time (i.e., there are serial correlation). However, with relatively minor modifications, many of the residuals diagnostic developed for standard linear regression can be extended to the longitudinal setting, particularly to the mixed effect models (Fitzmaurice et al., 2004). So, given that the residuals from an analysis of longitudinal data are correlated, and potentially correlated with the covariates, and do not necessarily have constant variance over time, the first step is transform the residuals to ensure they have constant variance and zero correlation, thereby mimicking residuals from a standard linear regression. This can be achieved using a well-known technique called the Cholesky decomposition (or Cholesky factorization) (for more details, see Fitzmaurice et al.2004). Using the set of transformed residuals, all of the usual residuals diagnostics for standard regression can be applied.

In order to check for any systematic departures form the model for the mean response, we can construct a **scatterplot of the transformed residuals versus the transformed predicted values**. The fitting of a smooth curve to the scatterplot can often help in judging whether curvature is present. In a correctly specified model, the scatterplot should display no systematic pattern, with a random scatter around a constant mean of zero. Similarly, the transformed residuals also make it somewhat easier to assess the normal distribution assumption and to identify potential outliers that require further investigation, call on a **normal quantile plot** (or so-called quantile-quantile or Q-Q plot). So, on the basis of the ranks, we can plot the sample quantiles of the residuals against the quantiles expected if they have a normal distribution.

If the residuals depart discernibly from a straight line, then the assumption of normality may not be tenable. By turn, outliers, will appear far from the ends of the line (Fitzmaurice et al., 2004).

In order to assess the adequacy of the model for the covariance, the **variogram**, previously used to suggest appropriate models, can now be used as a diagnostic tool. The variogram, given by (15), can be expressed as:

$$\begin{aligned}\gamma(h_{ijk}) &= \frac{1}{2}E(r_{ij} - r_{ik})^2 \\ &= \frac{1}{2}E(r_{ij}^2 + r_{ik}^2 - 2r_{ij}r_{ik}) \\ &= \frac{1}{2}Var(r_{ij}) + \frac{1}{2}Var(r_{ik}) - Cov(r_{ij}, r_{ik}).\end{aligned}$$

.

When the variogram is applied to the transformed residuals, it simplifies to:

$$\gamma(h_{ijk}) = \frac{1}{2}Var(r_{ij}*) + \frac{1}{2}Var(r_{ik}*) - Cov(r_{ij}*, r_{ik}*) = \frac{1}{2}(1) + \frac{1}{2}(1) - (0) = 1.$$

Thus, in a correctly specified model for the covariance, a **smooth plot of the empirical variogram for the transformed residuals** should fluctuate randomly around a horizontal line centered at 1, and display no systematic curvature or trend over time (Fitzmaurice et al., 2004).

3.2.5 *Missing Data*

In longitudinal studies scope, it is very often the case that some subjects miss some of their planned measurements for a variety of reasons, originating missing data (Diggle et al., 2002). Depending on the features of the *missing data patterns*, we can distinguish two types of missingness, namely *monotone* and *non-monotone*. The earliest covers the case of loss to follow-up or dropout, when a subject is withdrawn from the study before its intended completion, so measurement sequence terminates prematurely.[4] The non-monotone missingness, also known as intermittent, is a more general type that covers cases in which the response of a subject is missing at one follow up time, but other measurements are observed following missing values (Little & Rubin, 2019).

4 Another case of monotone missingness is late entry, when a subject does note provide some of her initial response measurements but appears later and stays in the study until completion.

Missing data poses several challenges for analysis. The first and most obvious is a loss of information and a potential reduction in precision of inference on the population of interest (i.e., loss of efficiency). In fact, the reduction in precision is directly related to the amount of missing data, and, to prevent this, standard pieces of advice are reiterated in literature, namely: i) enroll more individuals to achieve the same levels of power in detecting important effects; ii) avoid missing values wherever possible, by taking energetic steps to retain subjects in the study; and, iii) collect covariates that are useful for predicting missing values (Little, 1995). Another issue is that missingness results in unbalanced data sets because not all subjects have same number of measurements at a common set of occasions. This creates complications for methods of analysis that require balanced data, but it does not pose any concern for the linear mixed-effect model introduced earlier. Finally, of much greater concern is the potential for biased and misleading inferences that can result if the reasons for missingness are related to outcomes of interest. In fact, the main concern of longitudinal analysis with missing data arises when there is an association between the longitudinal profile and the missing process. For example, if a patient drops-out the study because he/she believes that the treatment is not being effective, the missing values should not be dissociated from the measurement process. Therefore, it is necessary to distinguish between different reasons or mechanisms for missing values, in order to account a possible association (Little, 1995).

The appropriateness of different methods to the analysis of incomplete longitudinal data is determined by the **missing data mechanism** (i.e., reason for missing values). The missing data mechanism can be thought of as the probability model describing the relation between the missing data **R** and response data **Y** processes. So, in general, let $\mathbf{Y}^*$ denote the complete set of measurements which would have been obtained were there no missing values. To distinguish between the response measurements we actually collected from the ones we have planned to, let **R** denote a set of indicator random variable defined as:

$$r_{ij} = \begin{cases} 1 & \text{if } y_{ij} \textit{ is observed} \\ 0 & \textit{otherwise} \end{cases}$$

Therefore, we obtain a partition of the complete response vector $\mathbf{Y}^*$ into two subvectores, the observed data subvector $\mathbf{Y}^o$ containing those y_{ij} for which $r_{ij} = 1$, and the missing data subvector $\mathbf{Y}^m$ containing the measurements which would have been available had they not been missing, for whatever cause.

A taxonomy of missing data mechanisms, first proposed by Rubin (1976), and further developed by Little & Rubin (2019), is based on the conditional density of the missingness process **R** given the complete response vector $\mathbf{Y}^* = (\mathbf{Y}^o, \mathbf{Y}^m)$.

MCAR - Missing Completely at Random: refers to the probability of missing does not depend on either the observed or unobserved measurements. So, the conditional density is given by:

$$[\mathbf{R}|\mathbf{Y}^o, \mathbf{Y}^m] = [\mathbf{R}]$$

An example is when subjects drop out prior to the end of the study because they move to locations that are inaccessible to the research or, in a clinical trial, when subjects forgot the appointment originating missing data. In both scenarios, we have arbitrary reasons for missingness.

Under *MCAR* assumption, the observed data $\mathbf{Y}^o$ can be considered as a random sample of the complete data $\mathbf{Y}^*$. This, in turn, means that the distribution of the observed data does not differ from the distribution of the complete data, and we can obtain valid inferences using any valid statistical procedures for the data, while ignoring the process(es) generating the missing values (Rizopoulos, 2012). A standard procedure for testing for MCAR is to compare the equality of the empirical distributions of observed variables across the patterns (i.e., for respondents and non respondents subjects), using *t*-tests for location (Little, 1995).

MAR - Missing at Random: refers to the probability of missing depends on the observed data, but not on the unobserved measurements. In this scenarios, the conditional density is given by:

$$[\mathbf{R}|\mathbf{Y}^o, \mathbf{Y}^m] = [\mathbf{R}|\mathbf{Y}^o]$$

A standard example of MAR longitudinal data arises when a protocol stated that subjects, whose response value exceeds a specific threshold (e.g., diastolic blood pressure exceeding 110 mmHg), should "jump" or to be removed from the study - which can be seen as a form of planned drop-out. In this case, subjects leaves the study on doctors' advice based on previous observed longitudinal measurements, missingness is under the control of the investigator and is related to the observed components of $\mathbf{Y}^o$ only (Little, 1995).

Due the fact that the missing data mechanism depends on $\mathbf{Y}^o$, the distribution of $\mathbf{Y}^o$ does not coincide with the distribution of $\mathbf{Y}^*$, and therefore the observed data cannot be considered a random sample from target population (as in MCAR case). An important implication of this feature of MAR is that sample moments (e.g., mean and variance) are not unbiased estimates of the same moments in the target population. Thus, statistical based on these moments, without accounting for MAR (e.g., Generalized Estimation Equation - GEE), may prove misleading. In contrast, likelihood-based analysis based on the observed data, even without accounting for MAR, can provide valid inferences if the model for the measurement process $\mathbf{Y}^*$ is correctly specified (Rizopoulos, 2012).

MNAR - Missing Not at Random: refers to the probability of missing depends on observed and unobserved data. In this scenario, the conditional density is given by:

$$[\mathbf{R}|\mathbf{Y}^o, \mathbf{Y}^m] or [\mathbf{R}|\mathbf{Y}^o, \mathbf{Y}^m] = [\mathbf{R}|\mathbf{Y}^m]$$

An example of MNAR longitudinal data arises when subject decides to leave the study because he/ she feels ill on the day of their appointment, and the illness is related with the observed longitudinal profile, as also with those measurements that would have been observed if he/she have kept on the study (Sousa, 2011).

MNAR also is often called nonrandom missingness, and in the case of dropout, nonrandom dropout, given the observed data do not constitute a random sample from the target population. Under MNAR, as the missingness mechanism is nonignorable, we can only obtain valid inferences from an analysis that is based on the joint distribution of the measurement and missing process. (Rizopoulos, 2012).

Concluding, MCAR and MAR missing data mechanisms are known in the literature as ignorable, since longitudinal likelihood-based analysis can be performed while the missing data model can be left unspecified, or ignored (Little & Rubin, 2019). However, missing values originated by MNAR are said to be informative or nonignorable, and in such cases, models for the missing-data mechanism become useful, as we will explain on section 3.3. To perform this models and inform researchers about the mechanism that is creating missing values, one should collect as much information as possible about the reasons for missing value (Little, 1995).

3.2.6 *Conclusion*

Longitudinal approach enable direct study of change, giving its capacity to distinguish changes over time within individuals (ageing effects) from differences among people in their baseline levels (cohort effects) (Diggle et al., 2002). In this sense, longitudinal analysis is a powerful tool to the scope of psychotherapy research, which enables expanding our knowledge about client's process change. Particularly, answering questions as whether, how, and for whom psychotherapy (doesn't) work(s).

In a longitudinal study, individuals are measured repeatedly over time, and, inevitably, missing data is a ubiquitous issue. By missing data, we mean data that we planned to collect but did not get, considering different reasons and patterns. A particular case is dropout (i.e., loss during follow-up of participants who initially were in the study), which, as the remains, poses challenges for both reliability and validity of the estimates.

According Little & Rubin (2019), there are three drop-out mechanisms, which need be appropriately handled in order to achieve unbiased estimates. They are: MCAR - missing completely at random, MAR - missing at random, and MNAR - missing not at random. The first two are known in literature as ignorable, since longitudinal likelihood-based analysis can be performed ignoring them. However, the former is said to

be informative or nonignorable, since the reasons for dropout are related to observed (longitudinal profile) and unobserved (survival mechanism) data, requiring, by turn, the joint modelling of longitudinal and survival data (Rizopoulos, 2012). Following section introduces a modelling framework especially designed to account for nonignorable dropout mechanisms.

3.3. JOINT MODELLING OF LONGITUDINAL AND SURVIVAL DATA

Joint models for longitudinal and survival data are particularly relevant to many longitudinal studies, specially when repeated measurements on a response variable, an observation on a possibly censored time-to-event ("failure" or "survival"), and additional covariate information are collected on each participant, and the main interest relies on interrelationships between these variables (Tsiatis & Davidian, 2004). In fact, joint modelling approach is required whenever at least one of the following scenarios is suspicious or true (Teixeira et al., 2019):

1. repeated measurements are correlated with time-to-event (i.e., in the presence of informative censoring - when the reason for censure is related to the study outcomes)

2. repeated measures are measured with error, and/or

3. non-ignorable missing data are present

In these scenarios, the researcher needs to establish the statistical structure for accounting for potential lack of independence between survival (or missing data mechanism) and longitudinal processes. Thus, joint modelling methods are preferable to separate analyses, both to make optimal use of the available information and to obtain unbiased estimates of the model parameters (Asar et al., 2015).

Depending on specific applications, there are at least three different kinds of scientific objectives in joint modelling methodology. Firstly, the objective may be to analyse the longitudinal measurements Y while allowing for association with a time-to-event outcome T that is not of direct scientific interest. A widely occurring example of this is when T is a potentially informative (non-random) dropout time. A second possible objective is to analyse the distribution of time-to-event outcome T, while taking account of the association between T and Y. In this case, the joint model can be interpreted as a strategy for conducting a survival analysis in the presence of an endogenous time-varying variable measured with error. Finally, joint modelling can be undertake to analyse the effects of covariates of interest on both longitudinal and time-to-event processes, considering their joint evolution. In this case, the relationship between Y and T is of direct scientific interest.

3.3.1 *Joint Models*

The essence of likelihood approaches in joint modelling is the specification of the joint distribution $[Y, T]$ for Y and T, the random variables of repeated measurements and failure time, respectively. There are three types of likelihood approaches that can be classified according to how the joint distribution is factored: selection models, mixture models and random effects models. Their differences can be perceived in the following equations, where $U = (U_1, U_2)$ represents the latent random effects that links the longitudinal and survival processes:

$$\text{Selection models:} \qquad [Y, T] = [Y] \times [T|Y] \tag{16}$$

$$\text{Pattern-mixture models:} \qquad [Y, T] = [T] \times [Y|T]. \tag{17}$$

$$\text{Random effects models:} \qquad [Y, T, U] = [U][Y|U_1][T|U_2] \tag{18}$$

Selection models factorise the joint density into to product of the marginal density of the repeated measurements and the conditional density of the survival process given the repeated measurement, as in (16). So, besides to specify a model for the complete longitudinal data, a survival or selection model is required in order to characterizes the dropout probability as a function of covariates and complete data (Hogan et al., 2004).

Pattern-mixture models factorise the joint density in the alternative way, that is as the product of the marginal density of the survival process, and the conditional density of repeated measurement given the event time (i.e., dropout), as in (17). In this sense, mixture models treat the longitudinal response distribution as a mixture over dropout patterns (or times to dropout), regarding them as a source of heterogeneity. Applied to longitudinal studies, patter-mixture models are well suited to analysis where the number of dropouts is small, and can be perfectly used as an exploratory tool to check on longitudinal profile stratify by missing pattern groups (Hogan et al., 2004; Sousa, 2011).

Although mathematically the selection and the pattern-mixture models describe exactly the same joint distribution, they have different statistical interpretations. Therefore, the model strategy to adopt depends mostly on the nature of the statistical problem and the scientific questions to be answered. Selection models are mainly used when inference focus relies on the influence of a longitudinal variable, measured with error, in the estimation of the survival sub-model parameters. In this case, it is possible to quantify the effect of the longitudinal outcome on the survival hazard. On the other hand, when primary interest is

on the longitudinal trajectory, which might be associated with an event pattern, the pattern-mixture models are more commonly used. To a review of different approaches and applications that have been suggested in the literature, please see for example Sousa (2011).

Random effects models, the third general class of joint models, use latent frailties or random effects to induce dependence between the responses Y and the event time T. Some versions of these models also have been referred to as frailty models, shared parameter models, or models with random-coefficient-dependent dropout (Hogan et al., 2004). A key feature of random effects joint models is that they are specified conditionally on a latent variable or a unobserved random effect (as in 18), by assuming conditional independence between Y and T given a bivariate random effect $U = (U_1, U_2)$. They formalize the intuitive idea that a subject's pattern of response in a study as well as his/her propensity to dropout are likely to depend on many characteristics of that subject, including some which are unoservable as an underlying disease or illness progression. These unobervable characteristics are then included in the model by a random effect, rather than to a the actual outcome (Diggle et al., 2002; Sousa, 2011). Moreover, in terms of Little & Rubin (2019) taxonomy of missing data mechanisms, the dropouts equation 18 are completely random if U_1 and U_2 are independent, whereas if U_1 and U_2 are dependent then in general the dropouts are informative (Diggle et al., 2002). Consequently, in practice, in absence of dependence between longitudinal response and dropout mechanism, the analysis becomes as two independent longitudinal and survival analyses (Sharf & Diener, 2010).

Depending on the primary interest of the analysis, random effects joint models can be formulated to handle either a survival process with longitudinal covariates measured with error or a longitudinal process with informative censoring, while accounting for the strength of the link between longitudinal and survival processes. Along with the different formulations there are various types of random effects joint models, which make use of different submodels. A typical formulation is given and detailed following.

3.3.2 *Random Effects Joint Model*

The most commonly used formulation of the random effects model is the use of a linear mixed effects model linked to a semi-parametric Cox model. In this sense, to represent the longitudinal process, an oft used submodel is a linear mixed effects model, previously introduced in section 3.2.2, that incorporates random effects into a linear model in order to account for the within-subject correlation, as given by:

$$\text{Longitudinal process:} \qquad y_i(t_{ij}) = \boldsymbol{X}_{1i}(t_{ij})\boldsymbol{\beta}_1 + U_{0i} + U_{1i}t(_{ij}) + \epsilon_i(t_{ij}) \tag{19}$$

where $y_i(t_{ij})$ corresponds to the i^{th} individual at time t_{ij}, X_{1i} is the $(n_i \times p)$ design matrix for the fixed covariates, with corresponding regression parameters $\boldsymbol{\beta}$. The random measurement error term $\epsilon_i \sim N(0, \sigma^2 I_{n_i})$ is assumed to be independent of $W_i(t) = U_{0i} + U_{1i}t_{(ij)}$. Finally, W_i is the subject-specific random effects, where U_{0i} and $U_{1i}t_{(ij)}$ represent, for the individual i, the random intercept and slope effects, respectively. Although the random intercept and slope model is used here to represent the longitudinal process, others options can be considered like a random intercept only or a random quadratic model (Philipson et al., 2012). Additionally, as postulated in section 3.2.2, this model can be extended by incorporating a stochastic component which accounts for the within-individual fluctuations in their repeated measurements over time (i.e., serial correlation). In doing so, this allows the separation of the within-subject variations from measurement error, which by turn gives us a better representation of the true overall longitudinal process (Diggle et al., 2002).

Regarding the survival process, an often used submodel is a time-dependent Cox model, as an extension of the Cox proportional Hazards model, to allow the incorporation of time-varying covariates. This model was previously indroduced in section 3.1. Thus, in order to incorporate the longitudinal random effects into the survival model, a joint model commonly used is shown in equation 20:

$$\text{Survival Process:} \qquad h_i(t) = h_0(t)exp\{\boldsymbol{x}^{\boldsymbol{T}}{}_{2i}\boldsymbol{\beta}_2 + \gamma_0 U_{0i} + \gamma_1 U_{1i}t\}, \tag{20}$$

where $\boldsymbol{x}_{2i}$ represents the baseline covariates with corresponding regression parameters $\boldsymbol{\beta}_2$, $h_0(t)$ represents the baseline hazard, and γ_0 and γ_1 represent the effect of the longitudinal random intercept and slope on the survival process, that is, the association parameters. Equation 20 assumes a separate association for the influence of the random effects on the survival process. However, a common association can also be assumed by allowing $\gamma = \gamma_0 = \gamma_1$ (Henderson et al., 2000).

Maximum likelihood (ML) estimation, one widely used method to estimate the parameters of the joint model, adopted in the present work, involves maximizing the log likelihood of the joint distribution of the longitudinal and survival processes given the random effects that are assumed to underlie both processes. The joint likelihood is given by:

$$\prod_{i=1}^{N} \int f(\boldsymbol{Y}_i|\boldsymbol{U}_i, \boldsymbol{\theta}) f(\boldsymbol{T}_i, \boldsymbol{\delta}_i|\boldsymbol{U}_i, \boldsymbol{\theta}) f(\boldsymbol{U}_i|\boldsymbol{\theta}) d\boldsymbol{U}_i,$$

where $f(\boldsymbol{Y}_i|\boldsymbol{U}_i, \boldsymbol{\theta}), f(\boldsymbol{T}_i, \boldsymbol{\delta}_i|\boldsymbol{U}_i, \boldsymbol{\theta})$ and $f(\boldsymbol{U}_i|\boldsymbol{\theta})$ are the densities functions for the longitudinal process, survival process, and random effects, respectively. The event indicator is given by $\boldsymbol{\delta}_i$, equaling one if the event occurred and zero otherwise.

In order to maximize the likelihood of the observed data and estimate the parameters of interest, maximization algorithms are implemented, with the most popular algorithm being the Expectation Maximization (EM) proposed by Dempster et al. (1977). More details of this approach can be founded in Diggle et al. (2008) and Henderson et al. (2000).

To synthesize, in the present study we adopt the random effect methodology developed by Wulfsohn & Tsiatis (1997) with the extension of Henderson et al. (2000), making use of joineR package in R, developed by Philipson et al.(2012). The idea behind this methodology is to analyse simultaneously data arising from survival and longitudinal processes exploiting dependencies between the components. Next, we will introduce some joint model diagnostic tools, namely conditional residual analysis for longitudinal data with nonrandom informative dropout.

3.3.3 *Model Diagnostic Procedures*

After obtain a final joint model using the random effects approach described in the previous subsection, regression diagnostic are necessary in order to validate the model's assumptions. The standard tools to assess these assumptions are residual plots. Properties and features of residuals, when survival and longitudinal outcomes are separately modeled, were introduced in section 3.1.5 and 3.2.4, respectively. Even though the same type of diagnostic plots can be easily constructed to inspect the fit of joint models, they can be misleading because residuals can be markedly affected by knowledge of the dropout time and type, which therefore should properly be taken into account in an assessment of model adequacy.

The dropout mechanism implied by joint models is of nonrandom nature, and consequently the observed data, upon which the residual are calculated, do not constitute a random sample of the target population (Fitzmaurice et al., 2004). This in turn implies that residual plots based on the observed data alone can be misleading because these residuals should not be expected to exhibit standard properties, such as zero mean and independence. In practice, due to the impact of nonrandom dropouts, we cannot discern if a given systematic trend in residual plots or lack-of-fit is truly attributed to a model misspecification.

To overcome the issue caused by the nonrandom dropout and produce residuals for the longitudinal process that can be readily used in diagnostic plots, Rizopoulos et al., (2010) proposed a multiple imputation (MI) procedure for generating multiple sets of completed data. Basically, the authors suggested to augment the observed data with randomly imputed longitudinal responses under the complete data model, corresponding to the longitudinal outcomes that would have been observed had the subjects not dropped

out. The simulation scheme to conduct this kind of analysis is available in the *residuals()* method, and can be invoked using the logical argument MI, in package *JM*() Rizopoulos, 2010, 2012b).

3.3.4 *Conclusion*

Often, psychotherapy studies collect both longitudinal measurements and event times, recorded on the participants during follow-up times. Joint models are appropriate when the longitudinal outcome and survival endpoints are associated in some way, and by turn, they can serve distinct purposes. Joint modelling methodology can be applied i) to analyse the time-to-event outcome (survival analysis), accounting the measurement errors or missing data in time-dependent covariate; ii) to analyse the longitudinal outcome in the presence of informative (non-random) dropout; and iii) to analyse effects of unobservable (latent) variables on both longitudinal and time-to-event outcomes, simultaneously. In these cases, separate inferences based on the longitudinal model and the survival model may lead to biased or inefficient results. Joint models, on the other hand, incorporate all information simultaneously and provide valid and efficient inferences (Wu et al., 2012).

Over the last two decades, joint models for longitudinal and survival data have received much attention in the literature, and have been adopted in several clinical research fields. Examples of the application can be found on the scopes of immune deficiency disease (HIV/AIDS) (e.g., Tsiatis et al., 1995), leukemia and cancer (e.g., Brown & Ibrahim, 2003; Ibrahim et al., 2010), cardiovascular and kidney diseases e.g.,(e.g., Andrinopoulou et al., 2012; Garre et al., 2008), and dementia (e.g., Proust-Lima et al., 2019) to name a few. However, as far as we know, no study has yet applied the joint modelling approach to the psychotherapy data, not even to the dropout or premature termination research.

Using a real data set as an example for the comparison between the separate (longitudinal and survival) analysis and joint model approaches, the present work has a threefold purpose: i) discussing the parametrerisation and implementation of these models in R, using the joineR package, ii) draw attention of users of this package to the interpretations of model parameters, iii) reinforce the relevance of these models in psychotherapy research. In the next two chapters we present the method and results, respectively, from the empirical study.

4

JOINT ANALYSIS OF LONGITUDINAL DATA AND TIME UNTIL PREMATURE TERMINATION IN PSYCHOTHERAPY

The following chapter presents the results from UP study, regarding separate and joint analysis of survival and longitudinal processes. Firstly, in Section 4.1, we give an overview of UP sample regarding some descriptive statistics and group comparison (i.e., censored vs. PT cases) for baseline sociodemographic and clinical variables.

In Section 4.2, we present the survival analysis performed in order to understand what the possible risk factors were for premature termination in psychotherapy, for these clients. Namely, the Kaplan-Meier estimates and the Cox regression models main results are described.

Section 4.3 focuses on longitudinal analysis of therapeutic alliance quality (4.3.1) and treatment outcome (4.3.2) in order to identify risk factors related to the mean progression of its values. Beginning with an exploratory analysis for group (i.e., censored vs. PT cases) means over time as for variation among individuals, following with the presentation of the main results of each longitudinal model fitted to the data.

Finally, in Section 4.4 we present the main results for joint analysis, incorporating both factors risk for the premature termination in psychotherapy (survival process) and for the progression of therapeutic alliance and treatment outcome variables (longitudinal process).

4.1. SAMPLE CHARACTERISTICS

Ninety-seven clients were available for analysis. Approximately 74% of overall sample consisted of woman (n = 72), and the modal age was 24 years, with a range of 18 to 56 years (mean age = 28.97, SD = 10.42)

Almost half (51.54%) of the subjects were married or had a significant relationship, 40.21% were single, and 8.25% were divorced or separated. The majority have an university degree (56.7%), and 43.3% com-

pleted between 9 to 12 years of school. Slightly more than half (54.64%) were students at the time of the first appointment, while 39.18% worked full-time or part-time, and 6.19% were unemployed.

Recall that, according survival analysis framework, Premature Termination (PT), defined as the unilateral initiative of the client to discontinue treatment, after at last one therapeutic session, without the therapist' agreement or knowledge, is, in the present study, the event of interest. Thus, considering the time-to event-outcome, 31 (32%) clients experienced the event of interest - Premature Termination; and 66 (68%) were censored. Table 3 shows the descriptives for demographic characteristics of the sample at intake, according to time-to-event outcome. No meaningful group differences were found in any of the demographic variables (age, sex, relationship status, education level, and professional status).

Table 3: Demographic characteristics of the sample at intake, according to termination status

	PT (n = 31)		Censored (n = 66)			
	M(SD)	n(%)	*M(SD)*	n(%)	Statistics	p-value
Age at intake						
(years)	28.61 (10.12)		29.14 (10.63)		$W = 1028.5$	0.969
Sex						
Male		7 (7.22%)		18 (18.56%)	$\chi^2(1) = 0.06$	0.807
Female		24 (24.74%)		48 (49.48%)		
Relationship status						
Single		15 (15.46%)		24 (24.74%)	$\chi^2(2) = 1.30$	0.523
Married/ in a relationship		14 (14.43%)		36 (37.11%)		
Divorced		2 (2.06%)		6 (6.19%)		
Education level						
9-12 years		12 (12.37%)		30 (30.93%)	$\chi^2(1) = 0.164$	0.685
University		19 (19.59%)		36 (37.11%)		
Professional status						
Employed		14 (14.43%)		24 (24.74%)	$\chi^2(2) = 0.748$	0.688
Unemployed		2 (2.06%)		4 (4.12%)		
Student		15 (15.46%)		38(39.18%)		

Note. PT = Premature termination; M = mean; SD = standard deviation; W = non parametric Mann-Whitney test ; χ^2 = chi-square test.

Regarding clinical characteristics of the sample, approximately 29 % of overall sample was taking psychiatric medication at the beginning of treatment. The clients were assigned a diagnosis according to DSM-V (American Psychiatric Association, 2013) with the following distribution of principal diagnoses:

50.52% **anxiety disorders** (22.68% social phobia, 17.53% panic disorders, 6.19% unspecified anxiety disorder, 3.09% generalized anxiety disorder, 1.03% specific phobia), 41.24% **mood disorders** (35.05% major depression, 4.12% dysthymia, 2.06% unspecified depressive disorder), and 8.25% **others disorders** (3.09% obsessive-compulsive disorder, 1.03% trichotillomania, 3.09% illness anxiety disorder, and 1.03% post-traumatic stress disorder).

At intake, a considerable number of clients presented clinical symptomatology on the OQ-45.2 (79.38%), as also clinical severity of depressive symptoms on the BDI-II (78.35%), while a few presented clinical anxiety on BAI (30.93%). Figure 3 shows the distribution from QO-45.2, BDI-II and BAI data, at intake, for all sample.

Considering PT and censored subgroups, there was also no major difference (Table 4) in any of the clinical variables at intake (psychiatric medication, diagnoses areas, symptomatology level, depressive severity and anxiety level), as also in symptomatology level as measured by OQ-10.2 questionnaire at last session observed. However, regarding the therapeutic alliance quality, as measured by WAI inventory at last session observed, premature termination group differ significantly from censored group. As we can see in Table 4, clients who discontinue treatment (PT) show, on average, a worse therapeutic alliance quality at last session observed (i.e., immediately before treatment discontinuation), when compared to censored cases (49.03 vs 55.38 mean scores).

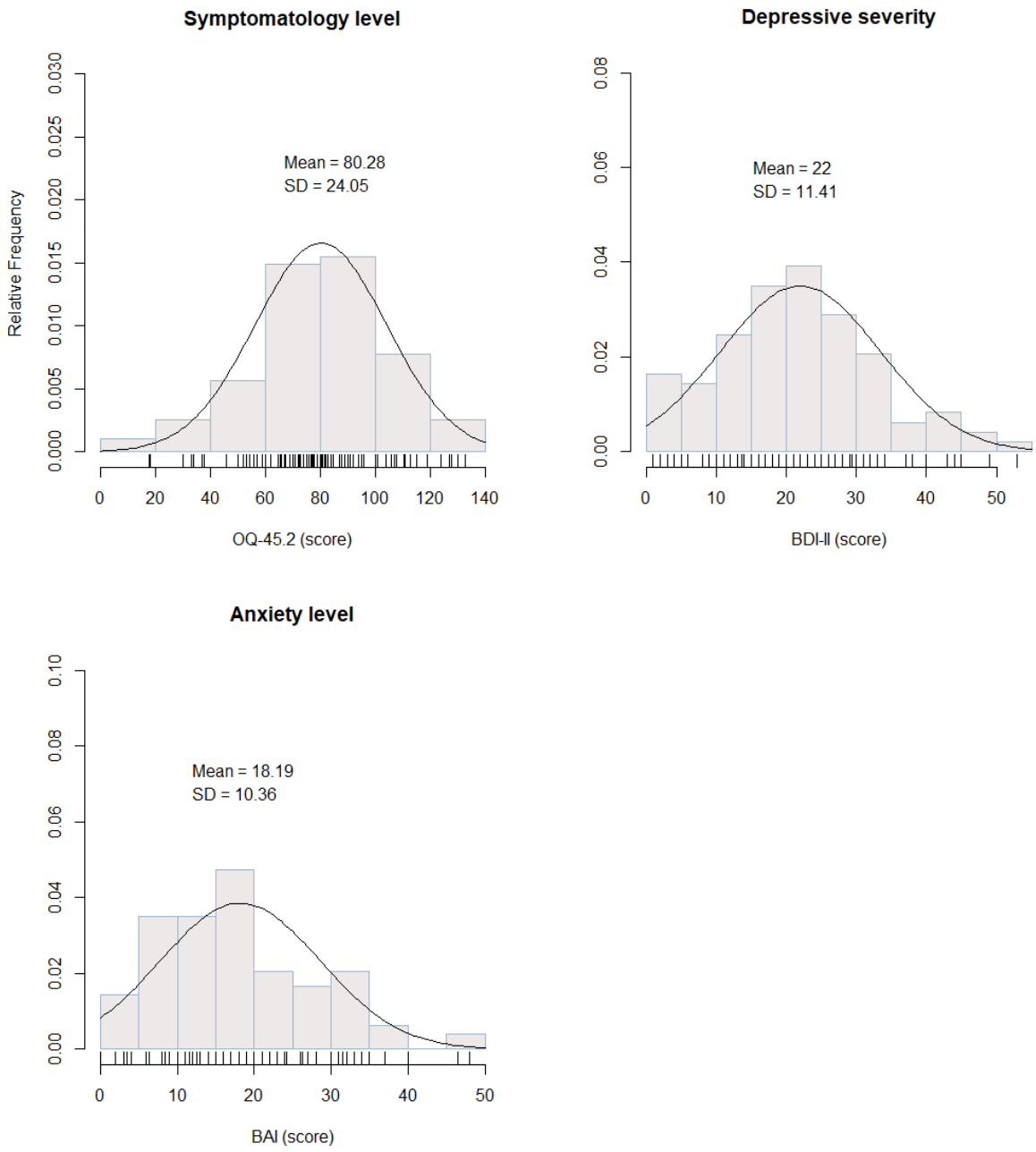


Figure 3: Frequency distributions graphs of clients' (a) symptomatology level, (b) depressive severity, and (c) anxiety level at intake

Table 4: Clinical characteristics of the sample at intake and last session observed, according to termination status

	PT (n = 31)		Censored (n = 66)			
	M(SD)	n(%)	*M(SD)*	n(%)	Statistics	p-value
Psychiatric medication at intake						
No		22 (23.40%)		45 (47.87%)	$\chi^2(1)$=0.003	0.954
Yes		8 (8.51%)		19 (20.21%)		
Diagnoses at intake						
Anxiety		14 (14.43%)		35 (36.08%)	$\chi^2(2)$=0.542	0.763
Mood		14 (14.43%)		26 (26.80%)		
Other		3 (3.09%)		5 (5.15%)		
Symptomatology level at intake						
clinical		22 (22.68%)		55 (56.70%)	$\chi^2(1)$=1.288	0.257
non-clinical		9 (9.28%)		11 (11.34%)		
Depressive severity at intake						
clinical		21 (21.65%)		55 (56.70%)	$\chi^2(1) = 2.174$	0.140
non-clinical		10 (10.31%)		11(11.34%)		
Anxiety level at intake						
clinical		10 (10.31%)		20 (20.62%)	$\chi^2(1)$= 0.003	1
non-clinical		21 (21.65%)		46 (47.42%)		
Therapeutic alliance at last session						
(WAI score)	49.03 (8.74)		55.38 (6.19)		$W = 1543$	4.78×10^{-5} ***
Symptomatology level at last session						
(OQ-10.2 score)	16.04 (7.16)		14.09 (7.50)		t(95)= -1.21	0.228

Note. PT = Premature termination; M = mean; SD = standard deviation; W = non parametric Mann-Whitney test ; χ^2 = chi-square test; t = Student's t-test. *** Indicates statistical significance at $p < .001$.

4.2. SURVIVAL ANALYSIS

Considering the time-to-event outcome, as aforementioned, 31 (32%) clients experienced the event of interest - premature termination; and 66 (68%) were censored. The sample distribution of event occurrences, namely the number and proportion of clients remaining in treatment (i.e., survival function), dropping out (i.e., hazard function), and completing treatment across the treatment sessions, is summarized in Table 5.

Table 5: Life table describing the number and proportion of dropouts by session for a sample of 97 clients

Session interval	Remaining in treatment	Dropping out	Censored	Hazard*	Survival**
$[1, 2)$	97	1		0.01	0.99
$[2, 3)$	96	3	2	0.03	0.96
$[3, 4)$	91	1	2	0.01	0.95
$[4, 5)$	88	6	1	0.07	0.88
$[5, 6)$	81	3	3	0.04	0.85
$[6, 7)$	75	3	4	0.04	0.82
$[8, 9)$	68	1	1	0.02	0.80
$[9, 10)$	66	4	2	0.06	0.76
$[10, 11)$	60	3	1	0.05	0.72
$[11, 12)$	56	2	1	0.04	0.69
$[12, 13)$	53	1	1	0.02	0.68
$[13, 14)$	51	2	1	0.04	0.65
$[15, 16)$	48		1	0.00	0.65
$[16, 17)$	47		8	0.00	0.65
$[17, 18)$	39	1	8	0.03	0.64
$[18, 19)$	30		18	0.00	0.64
$[19, 20)$	12		8	0.00	0.64
$[20, 21)$	4		4	0.00	0.64

Note. * The estimated hazard function is given by equation 2; ** The estimated survival function is given by equation 1.

Table 5 shows the greatest proportion of dropout occurred after completing the fourth treatment session (7%; n = 6). Besides that, nearly half of the clients (n = 17) who would ultimately dropout of treatment had already done so after completing the sixth treatment session. On the other hand, after session thirteen the hazard function is nearly zero, so the survivor function is flat, as we can see in Figure 4.

Figure 4 presents the Kaplan-Meier survival curve for all clients, considering right censoring and entire observed follow up time. Note that, the estimated survival probability (S_t), given by equation 1 and

represented by black line, is a step function that changes value only at the time of each event, that is when some participant drooped out. The dashed lines represent the 95% confidence intervals for the survival probability.

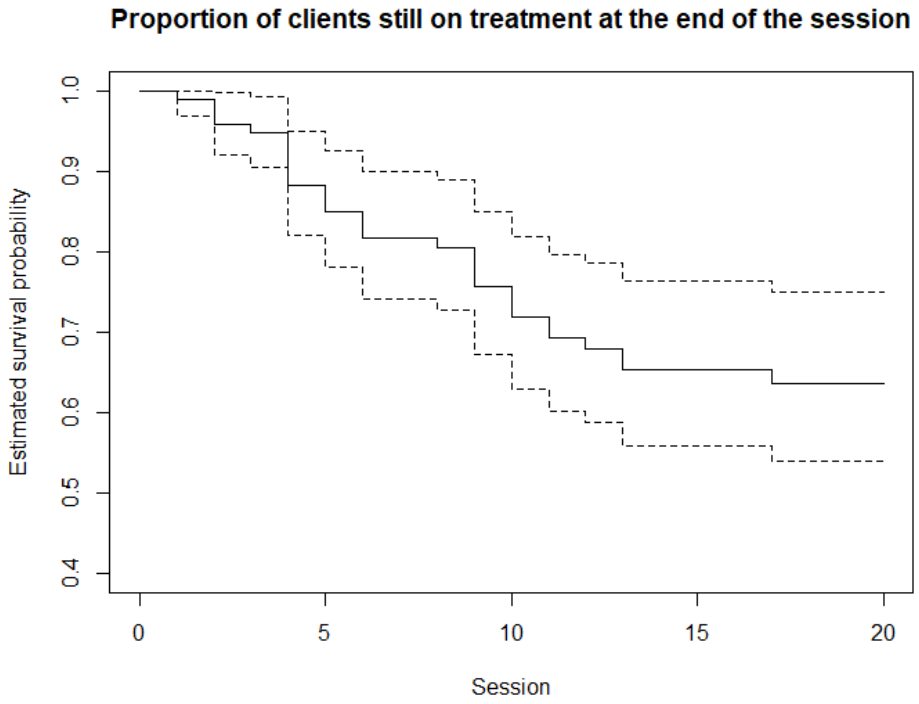


Figure 4: Kaplan-Meier curve estimate of UP clients

In order to estimate the relative risk of treatment PT we proceeded with the calculation of the Kaplan-Meier estimates stratified by category, considering the baseline covariables (i.e., measured at intake). To evaluate whether or not Kaplan-Meier survival curves were statistically different, the log rank test (Mantel-Cox test) was used. So, in Figure 5, the Kaplan-Meier survival curves comparing the probability of event free (i.e., still on treatment at the end of each session) for ten covariables are displayed, namely client's sex, age, relationship status, educational level, professional status, diagnoses areas, psychiatric medication, symptomatology level, depression severity and anxiety severity. The Log-rank test results are displayed in table 6.

Only one covariable turned out to have significant effect on client's risk of PT. The survival curves for clinical and non-clinical groups, regarding the depression severity (as measured by BDI-II at intake), differ marginally from each other ($\chi^2(1) = 3.4$, p=0.07). The respective Kaplan-Meier survival plot (Figure 5) shows the same probability of PT in the first sessions, but already on the fifth session, the clinical curve start to show a higher probability of survive per session when compared with the non-clinical curve. In other

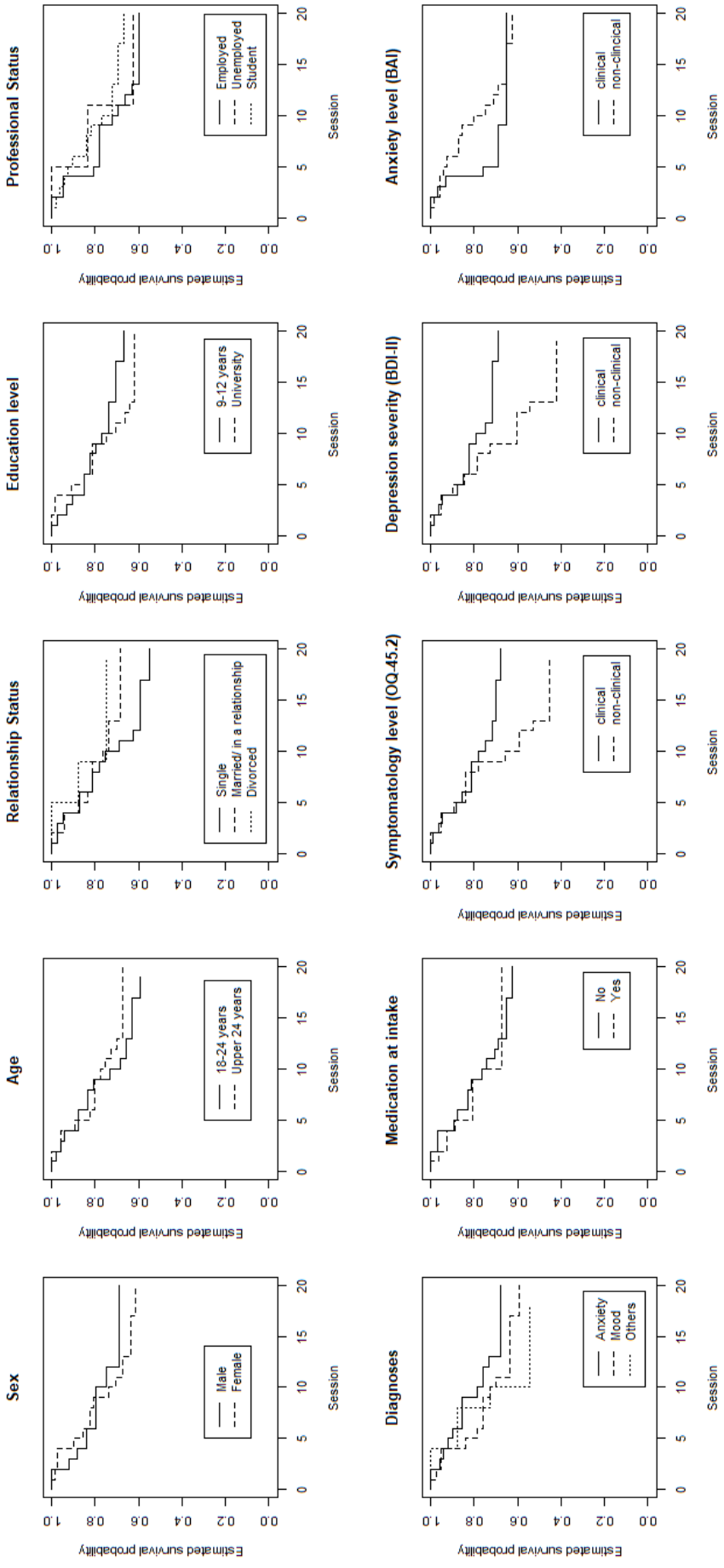


Figure 5: Kaplan-Meier curves estimates for the categorical variables measured at intake

words, after session five, the probability of PT is higher for non-clinical subgroup, regarding depression severity.

Table 6: Log-rank test results

	p-value
Sex	0.7
Age	0.6
Relationship status	0.5
Educational level	0.7
Professional status	0.8
Diagnoses	0.6
Medication at intake	0.9
Symptomathology level (OQ-45.2)	0.1
Depression severity (BDI-II)	0.07 †
Anxiety level (BAI)	0.7

Note. †indicates marginally significant difference at $p < .10$.

4.2.1 *Analysis with Cox Semi-Parametric Model*

The non-parametric methods described in previous section (i.e., Kaplan-Meier curves and log-rank tests), even useful in comparing the survival distributions of two or more groups, are unsuitable to describe the effect of a quantitative variable on survival times. Besides that, they compare groups for univariate analysis, taking into account only one factor at time and ignoring the impact of any others.

In order to assess simultaneously the effect of several risk factors, by considering both quantitative and categorical variables, we proceed to modeling techniques. So, for the analysis of the premature termination data in study, the Cox Proportional Hazards semi-parametric model (CPHM), given by the equation 3, described in section 3.1.3, was considered.

Univariate analysis for effect of explanatory variables

First, we started out by performing a univariate Cox regression analysis with single explanatory variables measured at intake (i.e., sex, age, relationship status, education level, professional status, diagnoses, psychiatric medication, symptomatology level, depression severity and anxiety level), as well as with two single explanatory variables measured at last session observed (i.e., therapeutic alliance quality and symptomatology level), in order to estimate the effect of those independent variables on survival of clients. Note that, some quantitative variables measured at intake, namely symptomatology level, depressive severity and anxiety level, were re-codified into a categorical type (i.e., clinical vs. non-clinical), so both scales were considered in the univariate analysis. Table 7 shows the estimates obtained.

Three covariables turned out to have marginally to highly significant effect on the client's risk of PT. Namely, there is an association between depressive severity (as measured by BDI-II at intake) and the risk of PT ($\hat{\beta}$ = 0.694, p = 0.071), meaning that the hazard to PT is two times higher for non-clinical group (HR = 2.003). Notably, this result is comparable to the log rank test (see Table 6), and may be interpreted in the same way. Likewise, there is an association between symptomatology level (as measured by OQ-45.2 at intake) and the risk of PT ($\hat{\beta}$ = - 0.014, p = 0.08), meaning that a unit decrease in the outcome questionnaire total score corresponds to a 1.4% fold increase in the risk for PT. Finally, there is an association between therapeutic alliance quality (as measured by WAI at last session) and the risk of PT ($\hat{\beta}$ = - 0.082, p = 3.53 $\times 10^{-5}$), meaning that a unit decrease in the working alliance inventory total score corresponds to a 8.6% fold increase in the risk for PT. To sum up, a decrease in symptomatology level score, depressive severity or therapeutic alliance quality is linked to a higher risk of PT.

Multivariate analysis for effect of explanatory variables

In a second phase, a multivariate model was adjusted and estimated by "step-wise backwards", starting with the sutured model with all baseline covariables (i.e., measured at intake) as also last observations for therapeutic alliance quality and symptomatology level, and then eliminating one-by-one the variables with lower significance (with a limit of 0.1 for the p-value, for inclusion of the variable). After adjusting several survival models with multiple covariates, we end up with a Cox proportional Hazard Model that incorporate four covariables with a significant effect on client's risk of PT, namely: diagnoses areas (anxiety, mood and others), symptomatology level at intake (OQ-45.2 score), symptomatology level at last session observed (OQ-10 score) and therapeutic alliance quality at last session observed (WAI score).

Table 8 shows the estimates obtained from multivariate Cox model adjusted. As one can see, the hazard to PT is 3.6 times higher for clients with a mood disorders and 4.1 times higher for clients with

Table 7: Estimated parameters (Est), hazard ratios (HR) and related 95% confidence intervals and p-value in analysis of UP data set with univariate CPHM for PT as the event

	Est	HR (95% CI)	p-value
Age at intake			
(years)	- 0.004	0.996 (0.961;1.03)	0.814
Sex (Ref = Male)			
(Female)	0.175	1.191 (0.513;2.764)	0.685
Relationship status (Ref = Single)			
(Married/ in a relationship)	- 0.325	0.723 (0.349;1.498)	0.383
(Divorced)	- 0.637	0.529 (0.121; 2.315)	0.398
Education level (Ref = 9-12 years)			
(University)	0.124	1.132 (0.549;2.333)	0.737
Professional status (Ref = Employed)			
(Unemployed)	- 0.144	0.866 (0.197;3.811)	0.849
(Student)	- 0.276	0.759 (0.366;1.573)	0.458
Diagnoses (Ref = Anxiety)			
(Mood)	0.316	1.372 (0.664;2.879)	0.403
(Others)	0.446	1.563 (0.448; 5.448)	0.484
Medication at intake (Ref = No)			
(Yes)	- 0.045	0.956 (0.426;2.148)	0.913
Depressive severity at intake (Ref = clinical)			
(non-clinical)	0.694	2.003 (0.942;4.257)	0.071 †
(BDI-II score)	- 0.019	0.981 (0.947;1.015)	0.273
Anxiety level at intake (Ref = clinical)			
(non-clinical)	- 0.152	0.859 (0.404;1.826)	0.692
(BAI score)	- 0.006	0.994 (0.958; 1.031)	0.736
Symptomatology level at intake (Ref = clinical)			
(non-clinical)	0.599	1.82 (0.836;3.959)	0.131
(OQ-45.2 score)	- 0.014	0.986 (0.971;1.002)	0.081 †
Symptomatology level at last session			
(OQ-10 score)	0.034	1.035 (0.988;1.084)	0.152
Therapeutic alliance at last session			
(WAI score)	- 0.082	0.921 (0.886;0.958)	3.5×10^{-5} ***

Note. Ref = Reference category; †Indicates marginally statistical significance; *** Indicates statistical significance at $p<.001$.

Table 8: Estimated parameters (Est), hazard ratios (HR) and related 95% confidence intervals and p-value in analysis of UP data set with multivariate CPHM for PT as the event

	Est	HR (95% CI)	p-value
Diagnoses (Ref = Anxiety)			
(Mood)	1.272	3.568 (1.399; 9.100)	0.008 **
(Others)	1.412	4.103 (1.012; 16.639)	0.048 *
Symptomatology level at intake			
(OQ-45.2 score)	- 0.041	0.960 (0.938; 0.983)	7.39×10^{-5}***
Symptomatology level at last session			
(OQ-10 score)	0.077	1.080 (1.010; 1.155)	0.025 *
Therapeutic alliance at last session			
(WAI score)	- 0.097	0.908 (0.865; 0.953)	8.75×10^{-5}***

Note. Ref = Reference category; †Indicates statistical significance at $p < .1$; * Indicates statistical significance at $p < .05$; ** Indicates statistical significance at $p < .01$; *** Indicates statistical significance at $p < .001$.

other disorders, when compared with clients with anxiety disorders. At baseline, an unit decrease in the outcome questionnaire (OQ-45.2 score) corresponds to a 4% fold increase in the risk for PT, but, at last session observed, an unit increase in the outcome questionnaire (OQ-10 score) corresponds to a 8% fold increase in the risk for PT. Besides that, there is an association between therapeutic alliance quality (as measured by WAI inventory at last session observed) and the risk to PT ($\widehat{\beta} =$ - 0.097, p $= 8.75 \times 10^{-5}$), meaning that a unit decrease in the working alliance inventory total score corresponds to a 9.2% fold increase in the risk for PT.

Figure 6 presents the plot of the Kaplan-Meier estimate and the multivariate Cox Proportional Hazard model for a subject with the following characteristics: mood disorder (the most frequent diagnoses observed in the data), an OQ-45.2 score at intake of 81, an OQ-10.2 score at last session of 14, and a WAI score at last session of 56 (i.e., the median values observed). This graph was built in order to graphically assess the model fit. The Cox Proportional hazards model curve remains within the 95% confidence interval of the Kaplan-Meier curve, which indicates a good fit to the data.

Along with this, Figure 7 shows the Kaplan-Meier estimates of the Cox-Snell residuals. Note that, the black solid line represents the Kaplan-Meier estimate of the survival function of the residuals (with the dashed lines corresponding the 95% confidence intervals), and the grey solid line denotes the survival function of the unit exponential distribution. Comparing the fit of the Kaplan-Meier estimate to the expected asymptotic distribution, even though some discrepancies occur (especially for residuals greater than 1.2),

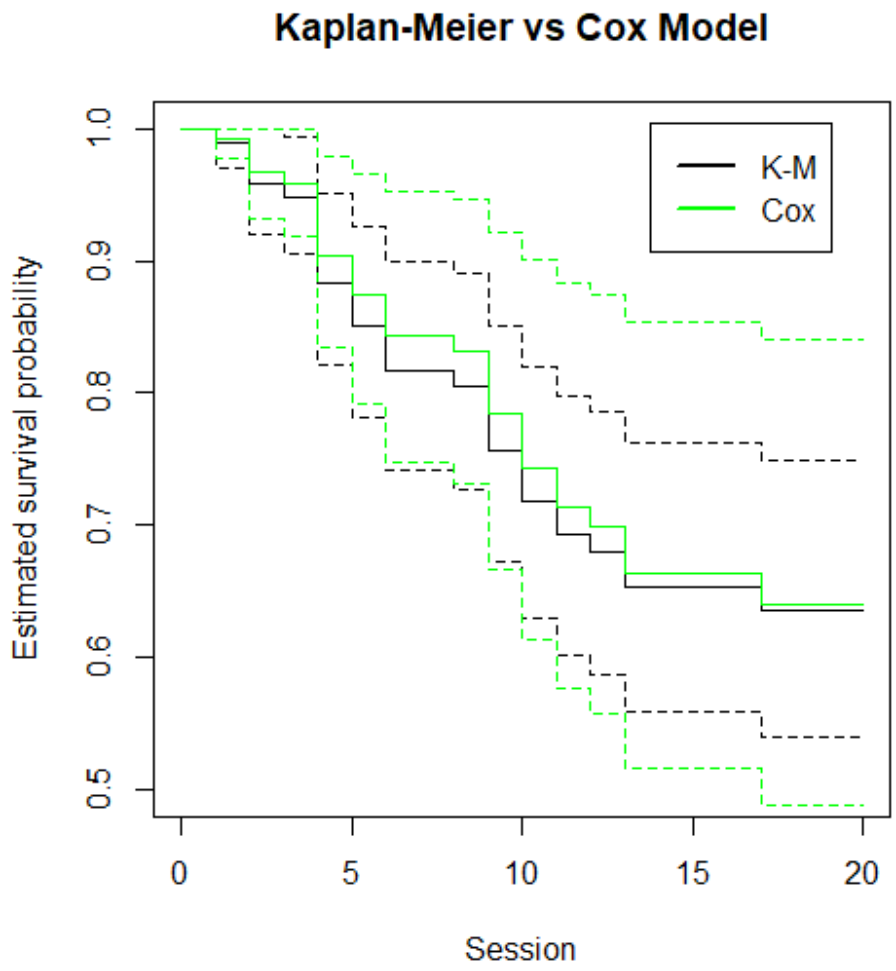


Figure 6: Kaplan Meier curve versus Cox Proportional Hazards curve for a combination of covariates (mood disorder, OQ-45.2 at intake=81, OQ-10.2 at last session=14, and WAI at last session=56)

the survival function of the unit exponential distribution lies within 95% confidence intervals of Kaplan-Meier estimate, which indicates, once again, a good fit of the multivariate Cox model to the data.

The goodness-of-fit (GOF) tests presented in Table 9 were conducted to investigate proportional hazards assumption. This approach provides chi-square statistics (i.e., correlation coefficients between transformed survival time and the scaled Schoenfeld residuals) for each variable in the model, as well as a global test. The idea behind the statistical test is that if the proportional hazards assumption holds for a particular covariate then the Schoenfeld residuals for that covariate will not be related to survival time. Consequently, a non significant p-value suggests that the proportional hazard assumption is reasonable, whereas a small p-value (i.e., less than 0.05) suggests that the variable being tested does not satisfy this assumption.

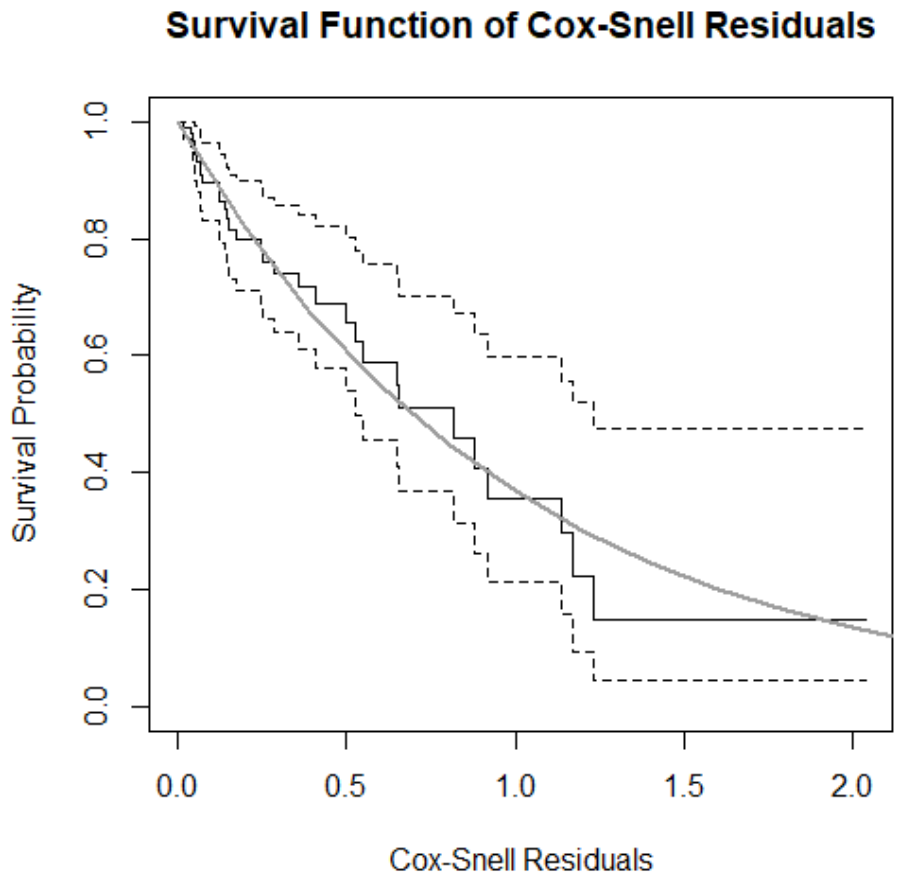


Figure 7: Superposition of Kaplan-Meier estimates of the survival function of the Cox-Snell residuals and survival function of the unit exponential distribution

Table 9: Chi-square statistics (χ^2) and p-values obtaining to investigate the PH assumption

	χ^2	p-value
Diagnoses areas	0.861	0.65
Symptomatology level at intake (OQ-45.2 score)	1.499	0.22
Symptomatology level at last session (OQ-10 score)	0.027	0.87
Therapeutic alliance at last session (WAI score)	0.001	0.99
Global	2.316	0.80

As shown in Table 9, there is no evidence to reject the null hypothesis of constant regression coefficients, both globally as well as for each covariate, which means that PH assumption was not violated.

Extended Cox Model

Following, the extended Cox Model, given by the equation 6, described in the section 3.1.4, was used in order to incorporate two time-dependent covariables, namely: symptomatology level (as measured by OQ-10.2 questionnaire at the beginning of each session) and therapeutic alliance quality (as measured by WAI inventory at the end of each session). First, we adjusted two univariate survival models with both covariates. Only the symptomatology level turned out to have a significant effect on the client's risk of PT. Specifically, there is an association between symptomatology level and the risk of PT ($\hat{\beta}$ = - 0.076, p = 0.012), meaning that a unit decrease in the outcome questionnaire total score corresponds to a 7.3% fold increase in the risk for PT (see Table 10).

Table 10: Estimated parameters (Est), hazard ratios (HR) and related 95% confidence intervals and p-value in analysis of UP data set with univariate time-dependent Cox model for PT as the event

	Est	HR (95% CI)	p-value
Symptomatology level			
(OQ-10 score)	- 0.076 (0.03)	0.927 (0.874; 0.984)	0.012 *
Therapeutic alliance			
(WAI score)	- 0.028 (0.026)	0.972 (0.923; 1.024)	0.29

Note. * Indicates statistical significance at $p < .05$.

In a second phase, a multivariate model was adjusted and estimated by "step-wise backwards", starting with the satured model with both time-fixed and time-dependent variables. Note that, when these new covariables were included in the satured model, we dropped out therapeutic alliance quality at last session observed and symptomatology level at last session observed, once they are a sub sample of time-dependent variables. After adjusting several survival models we end up once again with only one covariable with a significant effect on client's risk of PT, the symptomatology level (as measured by OQ-10.2 questionnaire).

4.3. LONGITUDINAL ANALYSIS

4.3.1 *Results of the Therapeutic Alliance Longitudinal Analysis*

The therapeutic alliance quality, as previously mentioned, was measured session by session, using the WAI inventory. Ninety-seven clients were available for analysis, which translates in a total number of 1179 measurements for this response variable. The median number of measurements per client is 14, varying between 1 and 20. Recall that, the total number of PT, the event of interest, is 31.

Group means over time

A *spaghetti plot* showing the therapeutic alliance quality individual progression (grey lines) of the longitudinal response for the both censored cases and premature terminators groups is presented in figure 8. The dashed black lines represent a non-parametric smooth spline of all observation points in the same plot, indicating the average trend of progression.

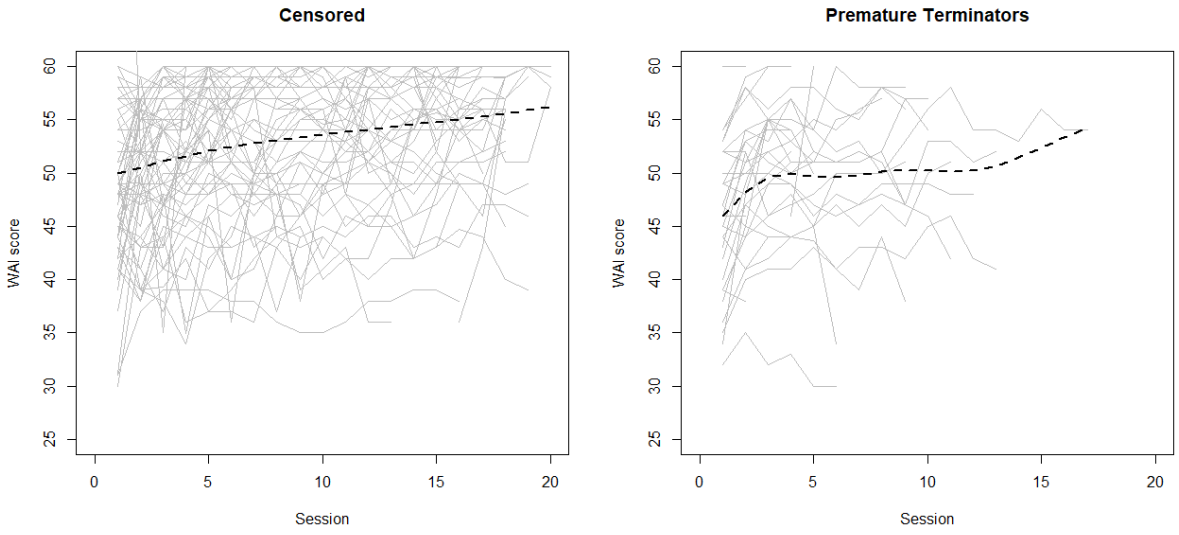


Figure 8: Spaghetti plot and smooth spline empirical mean of therapeutic alliance progression against the time from the beginning of treatment, for censored vs. PT cases

Considering Figure 8, we verify that, on average, the Premature Terminators start with lower values for therapeutic alliance quality, when compared to censored cases. Besides that, even the variable response progression for PT increases at a higher rate at the beginning of treatment, after session three slow down and the general trend of progression is lower when compared to censored cases.

In order to proceed with the inspection of a potential relation between the progression of the therapeutic alliance values and the PT event, a new spaghetti plot was built, considering the change in the response during the period before the event occurrence. This was achieved through recoding time so that one represents the event time (i.e., the last session observed for everyone).

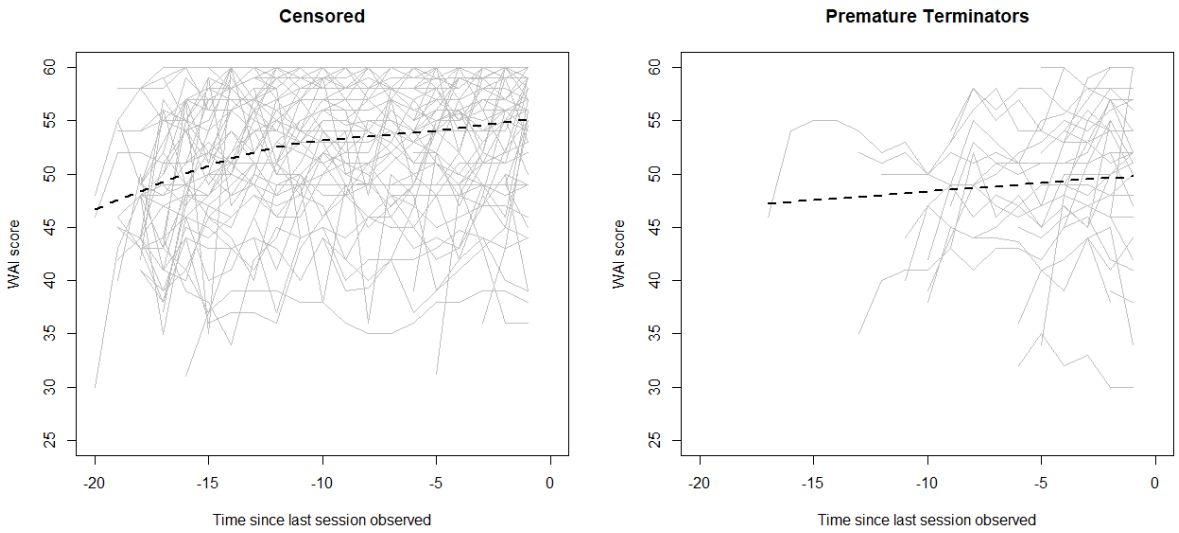


Figure 9: Spaghetti plot and smooth spline empirical mean of therapeutic alliance progression against the time until the event occurs, for censored vs. PT cases

Similarly to spaghetti plots with a positive time scale, in Figure 9 the smooth splines (dashed black lines) suggest that, on average, the therapeutic alliance quality progression increases at a higher rate for censored cases. In addition, premature terminators present a lower end point of the mean progression when compared to censored cases. And, it is evident, that the closer to the event time, the more therapeutic alliance quality fluctuate.

A linear-mixed effect analysis with an exponential correlation structure [1] was conducted in order to check if groups differences on therapeutic alliance mean progression are statistically significant.

In table 11 we present the results of two longitudinal models, using different time scales (i.e., time since first session vs time before the event occurs). As expected, premature terminators and censored cases differ significantly on both the starting and end points of the mean progression regarding therapeutic alliance quality. Intercept component of both models means that censored cases (i.e., the reference category for status variable) will start the progression of the therapeutic alliance quality with an expected value of 49.145 points, considering time since first session, and will terminate the progression of the therapeutic alliance quality with an expected value of 55.944 points, considering time since last session.

1 We are using the same correlation structure as the one obtained in the selected longitudinal model with multiple explanatory variables. The selection was advised on the empirical variogram as we opportunely will show.

Table 11: Estimated parameters (Est), standard errors (SE) and p-values for longitudinal models regarding group means differences over time. The response variable is therapeutic alliance quality

	Est (SE)	p-value
Intercept	49.145 (0.821)	<0.001 ***
Time (since 1^{st} session)	0.421 (0.053)	<0.001 ***
Status		
PT	-3.237 (1.476)	0.031 *
Time * Status (=PT)	0.097 (0.147)	0.509
Intercept	55.944 (0.827)	<0.001 ***
Time (since last session)	0.430 (0.054)	<0.001 ***
Status		
PT	- 5.990 (1.491)	<0.001 ***
Time* Status (=PT)	0.025 (0.149)	0.865

Note. PT = Premature termination. * Indicates statistical significance at $p < 0.05$; *** Indicates statistical significance at $p < 0.001$.

Premature terminators, by turn, have an expected decrease of -3.237 points on the intercept component at beginning of treatment, as well as an expected decrease of -5.990 points on the intercept component at last session observed. On the other hand, no meaningful differences between groups were found regarding the mean progression or rate of the therapeutic alliance quality throughout therapy (i.e., Time * Status).

Because therapeutic alliance mean progression differs according to the final event observed (i.e., PT vs censored), it becomes even more important investigate the association between longitudinal and survival mechanisms. This analysis, as we previously announced, requires a joint modelling approach, which will be presented further.

Variation among individuals

To explore the correlation structure present in the data, namely the association among repeated observations for an individual, and to determine the suitability of a linear mixed model to represent therapeutic alliance quality, an estimate of the empirical variogram $\gamma(u)$ was conducted. Figure 10 shows both basic quantities (u_{ijk}, h_{ijk}), where $h_{ijk} = 1/2(r_{ij} - r_{ik})^2$ is calculated from observed half-squared differences between pairs of residuals of an *Ordinary Least Square Model (OLS)*, considering therapeutic alliance quality as dependent variable and all attributes measured at intake, namely client's age, sex, relationship status, education level, professional status, diagnoses, medication, depressive severity, symptomatology and anxiety level, as independent variables), and $u_{ijk} = tij - tjk$ the corresponding time-

differences. The upper horizontal line represents the variogram based estimate of the process variance,

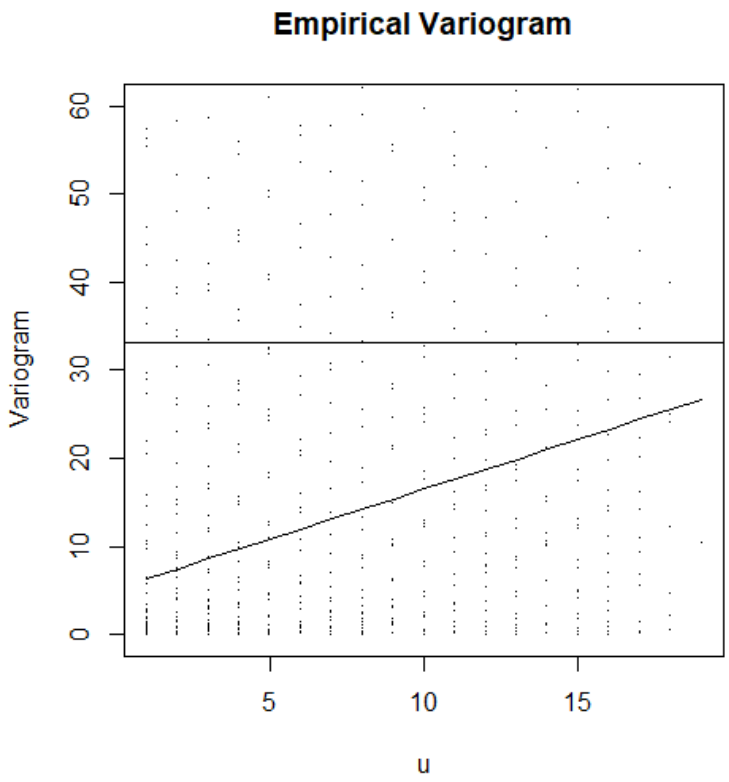


Figure 10: Empirical variogram from a OLS model considering therapeutic alliance (i.e., WAI score) as dependent variable and baseline attributes as independent variables

which is substantially larger than the value of the sample variogram (given by the amplitude of trend line), indicating that the positive correlation remains at large time separations. Further, we can see that the total variance in the data can be decomposed into three variance components: within-subject variance (given by the amplitude of trend line), between-subject variance (given by the distance between the maximum of the trend line and the upper horizontal line), and measurement error (given by the distance between zero and the minimum value of variogram line). Therefore, the empirical variogram from Figure 10 corroborates the adequacy of a longitudinal approach for these data, and indicates the need to include a random effect at subject level, as well as a random noise. Besides that, regarding the trend of the variogram line, a time correlation structure with an exponential parametrisation seems to be the one that which better fit the data.

Univariate analysis for effect of explanatory variables

With the motivation to understand which of the explanatory variables would affect the progression of the therapeutic alliance quality, and also how it would affect, firstly, we fitted single longitudinal models with single explanatory variables, using the same correlation structure as the one obtained in the selected longitudinal model with multiple explanatory variables, that will be presented next.

In table 12 the results from single models are shown, and, as we can see, time, professional status and diagnoses areas turned out to have significant effect on the mean progression of the therapeutic alliance quality. Specifically, there is an increase of 0.433 points per session, on average, of the therapeutic alliance quality; students present a lower starting point (- 2.097) of the mean progression, compared to employed clients; and, clients with mood disorder present also a lower starting point (- 2.431) of the mean progression of therapeutic alliance quality, compared to clients with anxiety diagnoses (the reference category). Overall, regarding diagnoses areas, there is an increasing effect of time on average therapeutic alliance quality, which is estimated to be 0.228 points per session higher than the rate of improvement among clients with anxiety disorders.

Multivariate analysis for effect of explanatory variables

Subsequently, a longitudinal analysis starting with all explanatory variables was performed. Specifically, a longitudinal linear mixed effects analysis with a random intercept at individual level (model given by equation 10) was conducted. Based on the analysis of the empirical variogram presented in Figure 10, two different correlation structures were tested: exponential and Gaussian. Therefore, two multivariate models were adjusted and estimated by "step-wise backwards", starting with the sutured model (with all pre-treatment covariables) and then eliminating one-by-one the variables with lower significance (with a limit of 0.1 for the p-value, for inclusion of the variable). After adjusting several longitudinal models with multiple covariates, we end up with a Longitudinal Mixed Effects Model that incorporate three covariables with a significant effect on mean progression of the therapeutic alliance quality, namely: time (i.e., session), symptomatology level (OQ-45.2 score at intake) and depression severity (BDI-II score at intake).

Table 14 presents the estimated parameters (Estimate) of the fitted longitudinal models, both with a random intercept an individual level, and with a serial correlation structure: one with Exponential structure - that is, *Linear Mixed Effects, Exponential Serial Correlation Model (LMEE)*; and the second with a Gaussian structure - that is, *Linear Mixed Effects, Guassian Serial Correlation Model (LMEG)*. Besides that, we compare the estimates to those obtained by fitting the simple OLS model, as well as the respective log Likelihood and *Akaike Information Criterion (AIC)* values.

Table 12: Estimated parameters (Est), standard errors (SE), and p-values for single Longitudinal Models. The response variable is therapeutic alliance quality

	Est (SE)	p-value
Time	0.433 (0.058)	<0.001 ***
Age at intake		
(years)	0.057 (0.058)	0.325
Sex (Ref = Male)		
(Female)	- 0.841 (1.369)	0.541
Relationship status (Ref = Single)		
(Married/ in a relationship)	0.321 (1.259)	0.8
(Divorced)	2.455 (2.239)	0.276
Education level (Ref = 9 to 12 years)		
(University)	- 0.025 (1.218)	0.984
Professional status (Ref = Employed)		
(Unemployed)	- 1.906 (2.555)	0.457
(Student)	- 2.097 (1.238)	0.094 †
Diagnoses (Ref = Anxiety)		
(Mood)	- 2.431 (1.427)	0.092 †
(Others)	1.917 (2.573)	0.458
(Time)	0.343 (0.066)	<0.001 ***
(Time * Mood)	0.228 (0.101)	0.024 *
(Time * Others)	0.208 (0.202)	0.303
Medication at intake (Ref = No)		
(Yes)	- 0.540 (1.352)	0.691
Depressive severity at intake (Ref = clinical)		
(non-clinical)	0.156 (1.472)	0.916
(BDI-II score)	0.009 (0.054)	0.874
Anxiety level at intake (Ref = clinical)		
(non-clinical)	- 0.168 (1.301)	0.897
(BAI score)	0.001 (0.059)	0.986
Symptomatology level at intake (Ref = clinical)		
(non-clinical)	1.547 (1.492)	0.303
(OQ-45.2 score)	-0.042(0.026)	0.103

Note. Ref = Reference category; †Indicates statistical significance at $p < .1$; * Indicates statistical significance at $p < .05$; *** Indicates statistical significance at $p < .001$.

Table 13: Estimated parameters (Est), standard errors (SE) and p-values for General Linear Model and Longitudinal Models. The response variable is therapeutic alliance quality

	LME		LMEG		OLS	
	Est (SE)	p-value	Est (SE)	p-value	Est (SE)	p-value
Intercept	52.147 (2.256)	<0.001 ***	53.209 (2.269)	<0.001 ***	52.061 (0.796)	<.001 ***
Time (Session)	0.451 (0.051)	<0.001 ***	0.454 (0.045)	<.001 ***	0.401 (0.036)	<0.001 ***
Depressive severity (BDI-II score)	0.224 (0.089)	0.013 *	0.223 (0.090)	0.014 *	0.174 (0.029)	<0.001 ***
Symptomatology level (OQ-45.2 score)	- 0.124 (0.042)	0.004 **	- 0.125 (0.043)	0.005 **	- 0.085 (0.014)	<.001 ***
Log Likelihood	- 3.066.545		- 3068.868		- 3851.002	
AIC	6149.091		6153.737		7712.003	

Note.†Indicates statistical significance at $p < .10$; * Indicates statistical significance at $p < .05$; *** Indicates statistical significance at $p < .001$.

Since LMEE model presented the higher Log Likelihood and the lower AIC values, we selected this as the model to describe the progression of the therapeutic alliance quality over time.

The intercept component of the LMEE model, in this particular case, means that a client with high OQ-45.2 score and lower depressive symptoms (BDI-II score) at intake will start the progression of therapeutic alliance quality with an expected value of 52.147 points. Overall, there is an increase rate of 0.451 points per session of the therapeutic alliance quality. Besides that, we can infer that depressive severity at intake affects the therapeutic alliance quality at a rate of 0.224 per unit increase in BDI-II score. Conversely, symptomatolgy level at intake affects the therapeutic alliance quality at a rate of -0.124 per unit increase in OQ-45.2 score. Concluding, higher scores for depressive severity and lower scores for symptomatology level, at intake, are linked to a higher starting point of the therapeutic alliance mean progression.

The correlation structure chosen to represent the variability of the data is the one that incorporates random effects at individual level with $\hat{v} \approx 9.495$, an exponential correlation structure to describe the variability within clients with $\hat{\rho}(u) \approx exp(-\frac{1}{15.640}|u|)$ and $\hat{\sigma}^2 \approx 29.534$, and a measurement error with variance $\hat{\epsilon} \approx 4.542$.

Figure 11 shows the superposition of the theoretical fitted variogram of both exponential and Gaussian correlation structure with the empirical variogram (of an OLS model considering time, depressive severity at intake and symptomatology level at intake as independent variables). This plot supports the choice of

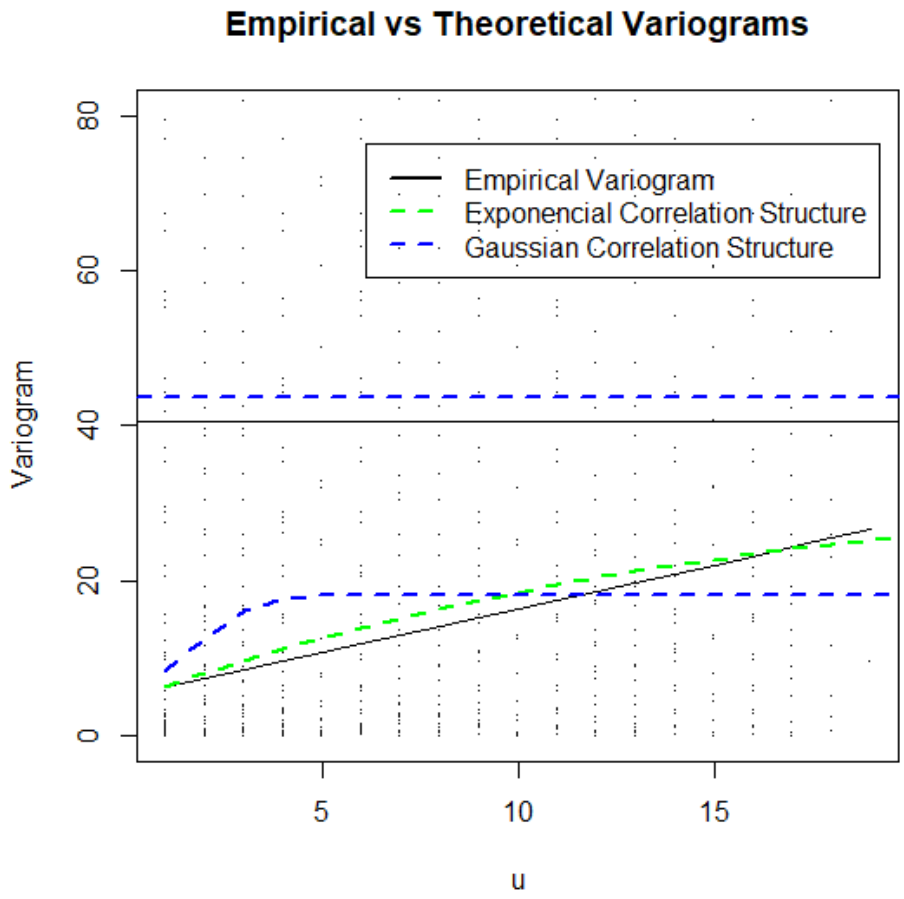


Figure 11: Superposition of empirical and theoretical variograms considering therapeutic alliance as dependent variable, and time, depression severity, and symptomatology level as independent variables

the LMEE model, since LMEE curve (green dashed line) is the one that best approaches the empirical curve.

In order to check assumptions of the final model fitted to the data, namely the normality of the random effect term and the residuals, normal probability plots (also called quantile-quantile) are presented in Figure 12. Despite some deviations from the expected normal lines towards the tails, the lines look straight and therefore pretty normal suggesting that the assumption is not violated.

Figure 13, by turn, presents the plot of the subject specific residuals versus the fitted responses. Observing this plot, the assumptions of homogeneous residual variance and linearity do not seem likely to be rejected.

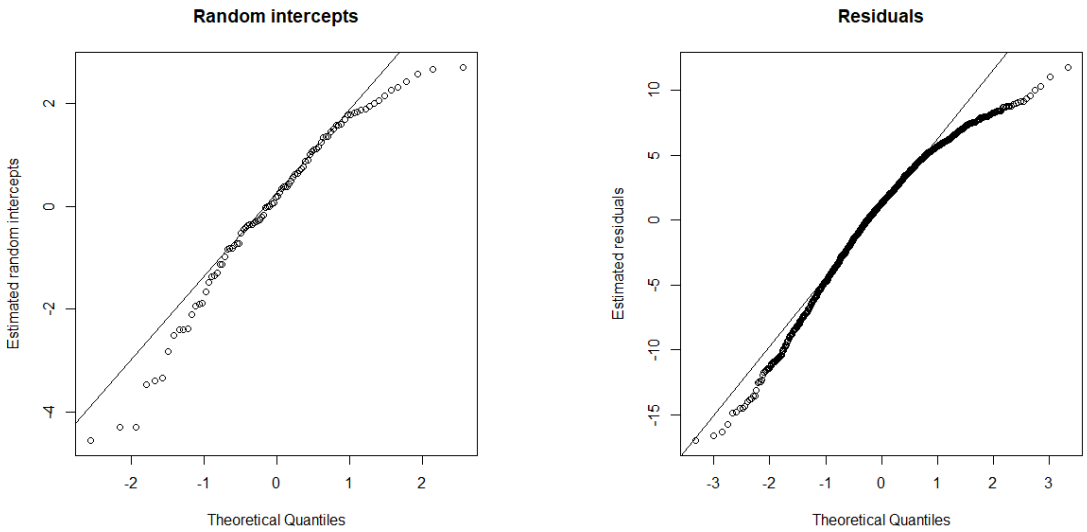


Figure 12: Quantile-quantile plots of predicted random intercepts and residuals for the LMEE model fitted to the data

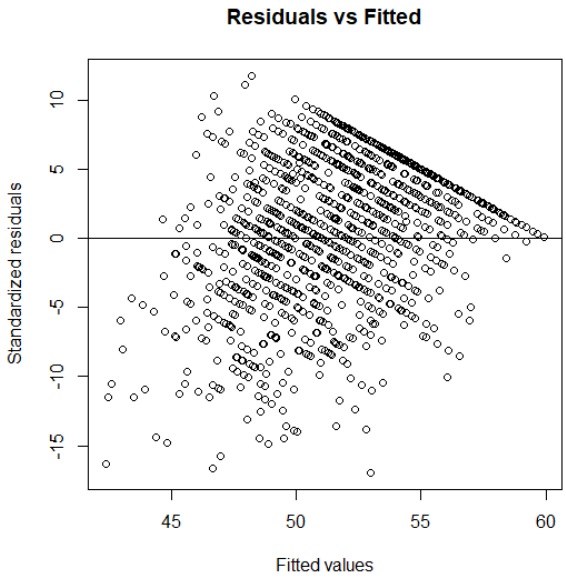


Figure 13: Residuals versus the fitted responses of WAI values

Group differences in explanatory variables effect

As PT in psychotherapy is our phenomenon of interest, we proceed to test group differences regarding the explanatory variables with significant effect on the mean progression of the therapeutic alliance quality. So, in order to investigate if explanatory effect on the therapeutic alliance mean progression differ for censored and PT cases, we added the variable status to the previous LMEE model and tested several interactions. The final model include one significant interaction between symptomatology level (as measured by OQ-45.2 at intake) and status (i.e., censored vs. PT).

Table 14: Estimated parameters (Est), standard errors (SE) and p-values for Longitudinal Model comparing groups for explanatory variables effect. The response variable is therapeutic alliance quality

	Est (SE)	p-value
Intercept	52.723 (2.753)	<0.001 ***
Time (Session)	0.431 (0.051)	< 0.001 ***
Status (Ref = Censored)		
PT	3.404 (4.147)	0.414
Depressive severity at intake		
(BDI-II score)	0.236 (0.085)	0.007 **
Symptomatology level at intake		
(OQ-45.2 score)	- 0.111 (0.044)	0.015 *
Symptomatology level at intake * Status (= PT)	- 0.086 (0.051)	0.098†
Log Likelihood	-3.062.209	
AIC	6144.418	

Note. PT = Premature termination.†Indicates statistical significance at $p < .10$; * Indicates statistical significance at $p < .05$; *** Indicates statistical significance at $p < .001$.

Table 14 presents the estimated parameters (Estimate) of the fitted longitudinal model, with a random intercept an individual level and an exponential correlation structure (i.e., LMEE model). Thus, regarding the symptomatology level effect, there is a marginal significant difference between censored and PT cases. While symptomatology level at intake affects the therapeutic alliance quality at a rate of - 0.111 per unit increase in OQ-45.2 score for censored cases, the rate almost double (- 0.192) for PT cases. In other words, higher scores in OQ-45.2 at intake (i.e., higher impairment) are associated to a lower starting point regarding the mean progression of therapeutic alliance quality (i.e., worse therapeutic alliance quality), and for PT group this relation/ rate is significantly more accentuate (i.e., 1.775 times higher compared to censored cases).

4.3.2 *Results of the Treatment Outcome Longitudinal Analysis*

The clients' symptomatology level or treatment outcome throughout therapy was measured by OQ-10.2 questionnaire at commencement of each session attended. Ninety-seven clients were available for analysis, which translates in a total number of 1184 measurements for this variable response. The median number of measurements per client is 14, varying between 1 and 20.

Group means over time

A *spaghetti plot* showing the symptomatology individual progression (grey lines) for the both censored (N=62) and PT (N=31) groups is presented in Figure 14. The dashed black lines represent a non-parametric smooth spline of all observation points in the same plot, indicating the average trend progression.

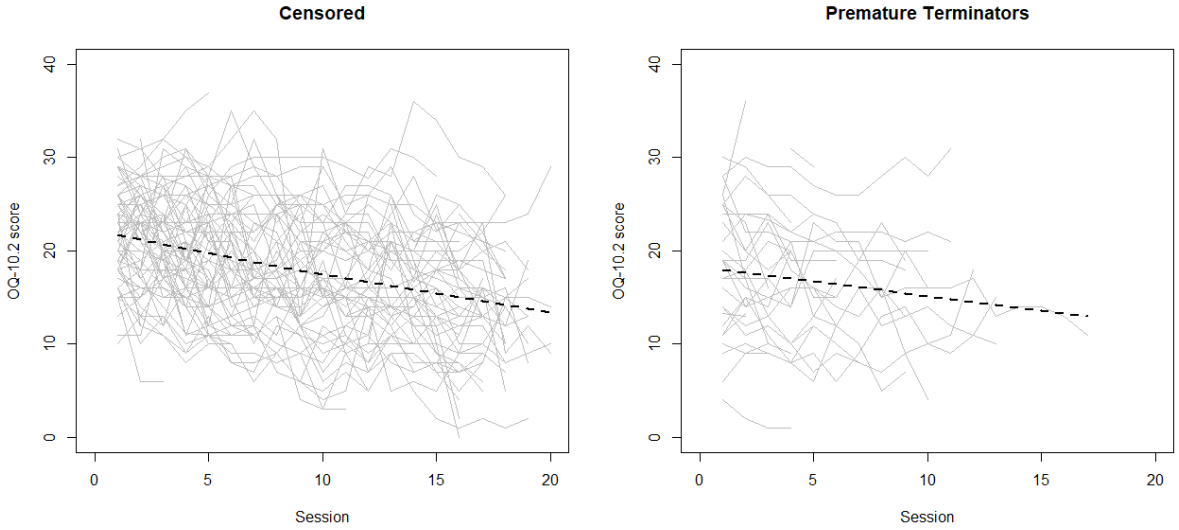


Figure 14: *Spaghetti* plot and smooth spline empirical mean of treatment outcome progression against the time from the beginning of treatment, for censored vs. PT cases

Considering Figure 14, we verify that, on average, premature terminators start with lower values for symptomatlogy level, when compared to censored cases, and for both groups the average trend of progression is decreasing over time.

A similar graphical presentation of treatment outcome progression is showed in Figure 15, but now considering the change in the response during the period before the event occurrence. In this case, the treatment outcome trend is decreasing for censored cases, but remains almost constant for premature terminators. Notably, this rises the suspicion that symptomatology level throughout treatment for premature

terminators remained unchanged. Additionally, it is evident, that the closer to the event time, the more a client's symptomatology level scores fluctuate.

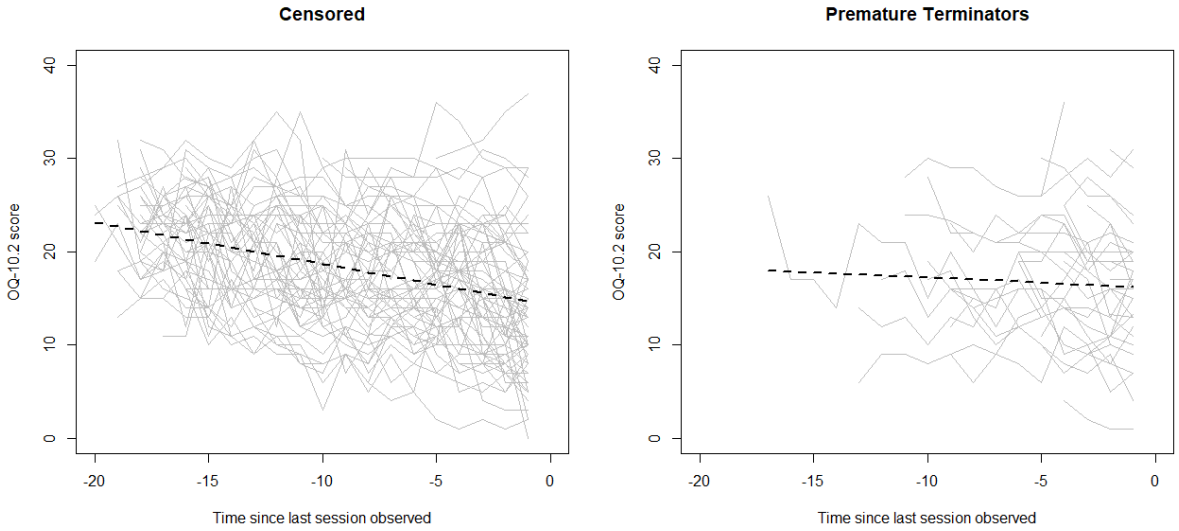


Figure 15: *Spaghetti* plot and smooth spline empirical mean of treatment outcome progression against the time until the event occurs, for censored vs. PT cases

In fact, after fitting a linear-mixed effect model with an exponential correlation structure [2], considering either time since first session or time since last session observed, we verify according Table 15 that premature terminators and censored cases differ significantly on starting points of the mean progression regarding symptomatology level. Intercept component of first model means that censored cases (i.e., reference category) will start the progression of symptomatology level with an expected value of 22.468 points. Premature terminators, by turn, have an expected decrease of -3.710 points on the intercept component at begining of treatment. No meaningful differences between groups were found regarding time since last session observed, which means that, on average, censored and premature terminators will terminate the progression of treatment outcome at identically expected values (13.951 vs 15.790 points, respectively). Likewise, no meaningful differences between groups were found regarding the mean progression rate of treatment outcome throughout therapy (i.e., Time * Status).

Once again, as symptomatology level mean progression differs according to the event outcome (i.e., PT vs. censored), a joint modelling approach is required in order to investigate the possible association between longitudinal and survival mechanisms. This analysis will be presented forward.

2 We are using the same correlation structure as the one obtained in the selected longitudinal model with multiple explanatory variables. The selection was advised on the empirical variogram as we opportunely will show.

Table 15: Estimated parameters (Est), standard errors (SE) and p-values for longitudinal models regarding differences between groups (Censored vs PT). The response variable is symptomatology level

	Est (SE)	p-value
Intercept	22.468 (0.795)	<0.001 ***
Time (since 1^{st} session)	- 0.532 (0.050)	<0.001 ***
Status		
PT	- 3.710 (1.429)	0.011 *
Time * Status (= PT)	0.154 (0.136)	0.259
Intercept	13.951 (0.801)	<0.001 ***
Time (since last session)	- 0.535 (0.051)	<0.001 ***
Status		
PT	1.839 (1.436)	0.204
Time * Status (= PT)	0.198 (0.138)	0.151

Note. PT = Premature termination. * Indicates statistical significance at $p < 0.05$; *** Indicates statistical significance at $p < 0.001$.

Variation among individuals

In order to explore the correlation structure present in the data, an estimate of the empirical variogram $\gamma(u)$ was conducted. Figure 16 shows both basic quantities (u_{ijk}, h_{ijk}), where $h_{ijk} = 1/2(r_{ij} - r_{ik})^2$ is calculated from observed half-squared differences between pairs of residuals of an ordinary least squares (OLS) model (considering symptomatology level as dependent variable and attributes measured at intake, namely client's age, sex, relationship status, education level, professional status, diagnoses, medication, depressive severity, symptomatology and anxiety level, as independent variables), and $u_{ijk} = t_{ij} - t_{ik}$ the corresponding time differences (or lag). The upper horizontal line represents the variogram based estimate of the process variance, which is substantially larger than the value of the sample variogram, given by the amplitude of trend line, indicating, by turn, that the positive correlation remains at large time separations. The total variance can be decomposed into three components: within-subject variance, given by the amplitude of the trend line; between-subjects variance, given by the distance between maximum of the trend line and the upper horizontal line; and, measurement error, given by the distance between zero and the minimum value of variogram line. Consequently, the empirical variogram form Figure 16 corroborates the adequacy of a longitudinal approach for these data, and indicates the need to include a random effect at subject level, as well as a random noise. Besides that, considering the trend of the

variogram line, a time correlation structure with an exponential parametrisation seems to be the one that which better fit the data.

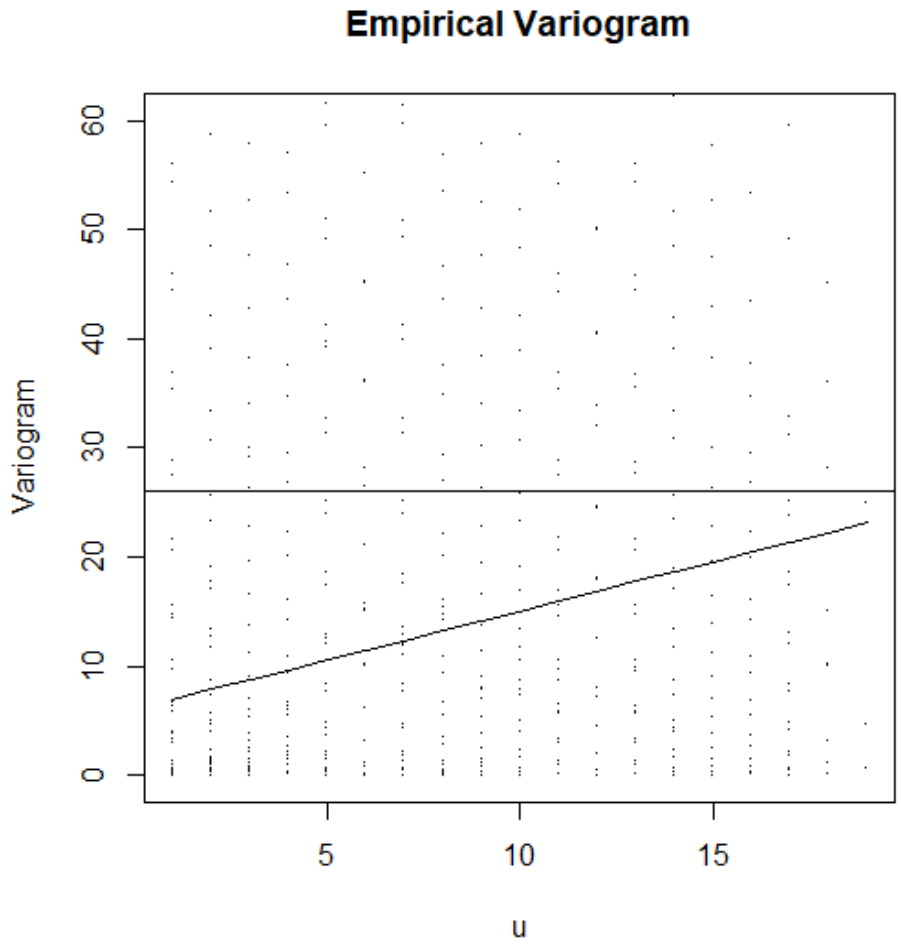


Figure 16: Empirical variogram from a OLS model considering treatment outcome as dependent variable, and time, and baseline attributes as independent variables

Univariate analysis for effect of explanatory variables

With the motivation to understand which of the explanatory variables would affect the progresison of the treatment outcome, firstly, we fitted single longitudinal models with single explanatory variables, using the same correlation structure as the one obtained in the selected longitudinal model with multiple explanatory variables, that will be presented further. Note that, some quantitative variables measured at intake, namely symptomatology level, depressive severity and anxiety level, were re-codified into a categorical type (i.e., clinical vs. non-clinical), so both scales were considered in the univariate analysis. Table 16 presents the results of the single longitudinal models.

Table 16: Estimated parameters (Est), standard errors (SE), and p-values for single Longitudinal Models. The response variable is symptomatology level over treatment

	Est (SE)	p-value
Time	- 0.498 (0.046)	<0.001 ***
Age at intake		
(years)	- 0.086 (0.056)	0.131
Sex (Ref = Male)		
(Female)	1.408 (1.347)	0.299
Relationship status (Ref = Single)		
(Married/ in a relationship)	- 3.697 (1.185)	0.002 **
(Divorced)	0.281 (2.113)	0.895
Education level (Ref = 9 to 12 years)		
(University)	0.448 (1.196)	0.709
Professional status (Ref = Employed)		
(Unemployed)	7.009 (2.429)	0.005 **
(Student)	2.983 (1.175)	0.013 *
Diagnoses (Ref = Anxiety)		
(Time)	- 3.361 (0.058)	<0.001 ***
(Mood)	3.108 (1.352)	0.024 *
(Others)	- 3.181 (2.438)	0.195
(Time * Mood)	- 0.309 (0.088)	5×10^{-4}***
(Time * Others)	- 0.226 (0.178)	0.203
Medication at intake (Ref = No)		
(Yes)	0.184 (1.326)	0.890
Depressive severity at intake (Ref = clinical)		
(Time)	- 0.556 (0.050)	<0.001 ***
(non-clinical)	- 7.982 (1.509)	<0.001 ***
(Time * non-clinical)	0.321 (0.119)	0.007 **
(BDI-II score)	0.354 (0.051)	<0.001 ***
(Time)	- 0.220 (0.103)	0.033 *
(Time * BDI-II score)	- 0.012(0.004)	0.004 **
Anxiety level at intake (Ref = clinical)		
(non-clinical)	- 3.638 (1.224)	0.004 **
(BAI score)	0.231 (0.053)	<0.001 ***
Symptomatology level at intake (clinical)		
(non-clinical)	-8.434 (1.208)	<0.001 ***
(OQ-45.2 score)	0.184 (0.017)	<0.001 ***

Note. Ref = Reference category; †Indicates marginally statistical significance; ** Indicates statistical significance at $p<.01$; *** Indicates statistical significance at $p<.001$.

Results show that time, relationship status, professional status, diagnose, depressive severity at intake, as also anxiety and symtomatology level at intake have a significant effect on the mean progression of the treatment outcome. Overall, there is a decline of - 0.498 points per session of the symptomatology level mean progression. Regarding relationship status, clients who are married or in a significant relationship present a lower starting point -3.697 of mean progression, compared to those who are single (the reference category). Conversely, unemployed and students present a higher starting point (7.009 and 2.983, respectively) of the symptomatology level mean progression, when compared to employed clients.

Moving to clinical attributes, mood disorders are related to an increase of the starting point of the symptomatology mean progression (3.108), and the average rate of decline is estimated to be -0.309 points per session higher than the rate of decline among clients with anxiety disorders (the reference category). Besides that, depressive severity at intake affects the symptomatology mean progression at a rate of 0.354 per unit increase in BDI-II score. In a similar vein, taking depressive severity as categorical variable, we can infer that non-clinical group presents a lower starting point (-7.982) of the mean progression of symptomatology level, when compared to clinical group (i.e., the group with higher BDI-II scores), and, additionally, the average rate of decline is estimated to be 0.321 points per session lower than the rate of decline among clinical group. Also, anxiety level at intake, treated whether as categorical or quantitative variable, present a significant effect on the mean progression of the symptomatology level, with non-clinical group showing a lower starting point (-3.683) of the mean progression, compared to clinical group. Finally, as expected, symptomatology level at intake, affects the mean progression of response variable at a rate of 0.184 per unit increase in OQ-45.2 score. Considering categorical scale, we can also say that, non-clinical group presents a lower starting point (-8.434) of the mean progression of symptomatology level, compared to clinical group.

Multivariate analysis for effect of explanatory variables

Subsequently, a longitudinal analysis starting with all explanatory variables was performed. Specifically, a longitudinal linear mixed effects analysis with a random intercept at individual level (model given by equation 10) was considered. Based on the analysis of the empirical variogram presented in Figure 16, two different correlation structures were tested: exponential and Gaussian. Therefore, two multivariate models were adjusted and estimated by "step-wise-backwards", starting with the satured model (i.e., with all pre-treatment covariables) and then eliminating one-by-one the variables with lower significance (with a limit of 0.1 for the p-value, for inclusion of the variable). After adjusting several longitudinal models with multiple covariates, beginning, as already mentioned, with the satured model, we end up with a Longitu-

dinal Mixed Effects Model that incorporate four covariables with a significant effect on mean progression of the treatment outcome, namely: time (i.e., session attended), clients' age (i.e., years), symptomatology level (i.e., OQ-45.2 score) and depression severity (i.e., BDI-II score) at intake.

Table 18 presents the estimated parameters (Estimate) of the fitted longitudinal models, both with a random intercept an individual level, and with a serial correlation structure: one with Exponential structure (LMEE), and the second with a Gaussian structure (LMEG). Besides that, we compare the estimates to those obtained by fitting the simple OLS model, as well as the respective log Likelihood and AIC values.

Table 17: Estimated parameters (Est), standard errors (SE) and p-values for General Linear Model and Longitudinal Models. The response variable is symptomatology level

	LMEE		LMEG		OLS	
	Est (SE)	p-value	Est (SE)	p-value	Est (SE)	p-value
Intercept	22.491 (1.002)	<0.001 ***	22.403 (1.004)	<0.001 ***	22.414 (0.414)	<0.001 ***
Time (Session)	- 0.489 (0.045)	<0.001 ***	- 0.500 (0.038)	<0.001 ***	- 0.404 (0.031)	<0.001 ***
Professional Status (Ref = Employed)						
(Unemployed)	5.074 (2.003)	0.013 *	5.296 (2.035)	0.011 *	4.773 (0.689)	<0.001 ***
(Student)	2.493 (0.984)	0.013 *	2.531 (1.001)	0.013 *	2.521 (0.336)	<0.001 ***
Anxiety level (Ref = clinical)						
(non-clinical)	- 2.068 (1.074)	0.057 †	- 1.987 (1.092)	0.072 †	- 2.560 (0.368)	<.001 ***
Symptomatology level (Ref = clinical)						
(non-clinical)	- 6.960 (1.246)	<0.001 ***	- 7.050 (1.261)	<0.001 ***	- 6.371 (0.438)	<0.001 ***
Log Likelihood	-3113.028		-3123.179		-3663.512	
AIC	6246.057		6266.359		7341.025	

Note. †Indicates statistical marginal significance at $p < 0.1$; * Indicates statistical significance at $p < 0.05$; *** Indicates statistical significance at $p < 0.001$.

Since LMEE longitudinal model presented the higher Log Likelihood and the lower AIC values, we selected this as the model to describe the progression of the symptomatology level over time.

The intercept component of the LMEE model, in this particular case, means that clients who are employed, and pertain to the clinical group for both anxiety and symptomatology level at intake will start the progression regarding treatment outcome with an expected value of 22.491 points. Additionally, unemployed and students will start the mean progression with an expected increase of 5.074 and 2.493 points, respectively, compared to employed clients (the reference category). Regarding anxiety level and symptomatology level at intake, clients in the non-clinical group will start the mean progression with an expected

decrease of -2.068 and -6.960 points, respectively, compared to clinical group (the reference category). Overall, there is a decrease of -0.489 points per session of the symptomatology level. .

The correlation structure chosen to represent the variability of the data is the one that incorporates random effects at individual level with $\hat{v} \approx 1.709 \times 10^{-4}$, an exponential correlation structure to describe the variability within clients with $\hat{\rho}(u) \approx exp(-\frac{1}{17.827}|u|)$ and $\hat{\sigma} \approx 24.952$, and a measurement error with variance $\hat{\epsilon} \approx 5.222$.

Figure 17 shows the superposition of the theoretical fitted variogram of both exponential and Gaussian correlation structure with the empirical variogram (of an OLS model considering time, professional status, anxiety level and symptomatology level at intake as independent variables). This plot supports the choice of the LMEE model, since LMEE curve (green dashed line) is the one that best approaches the empirical curve.

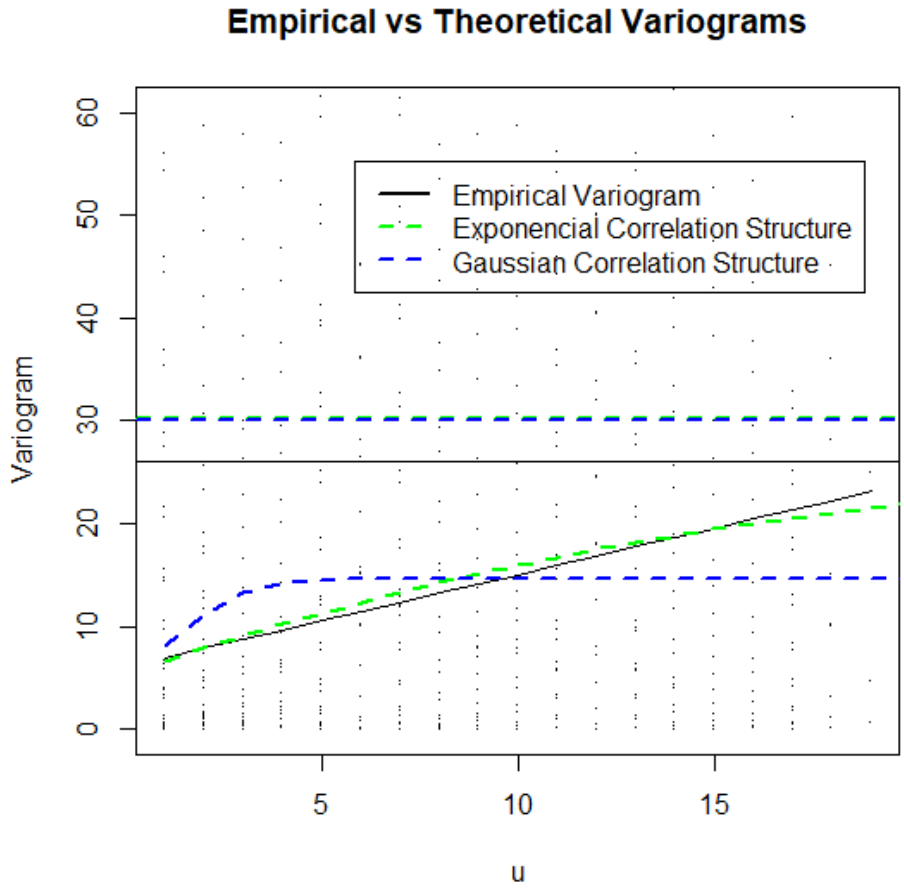


Figure 17: Superposition of empirical and theorical variograms considering symptomatology level as dependent variable, and time, professional status, anxiety level, and symptomatology level at intake as independent variables

In order to check assumptions of the final model fitted to the data, namely the normality of the random effect term and the residuals, normal probability plots are presented in Figure 18. Despite some deviations from the expected normal lines towards the tails, overall the lines look straight and therefore pretty normal suggesting that the assumption is not violated.

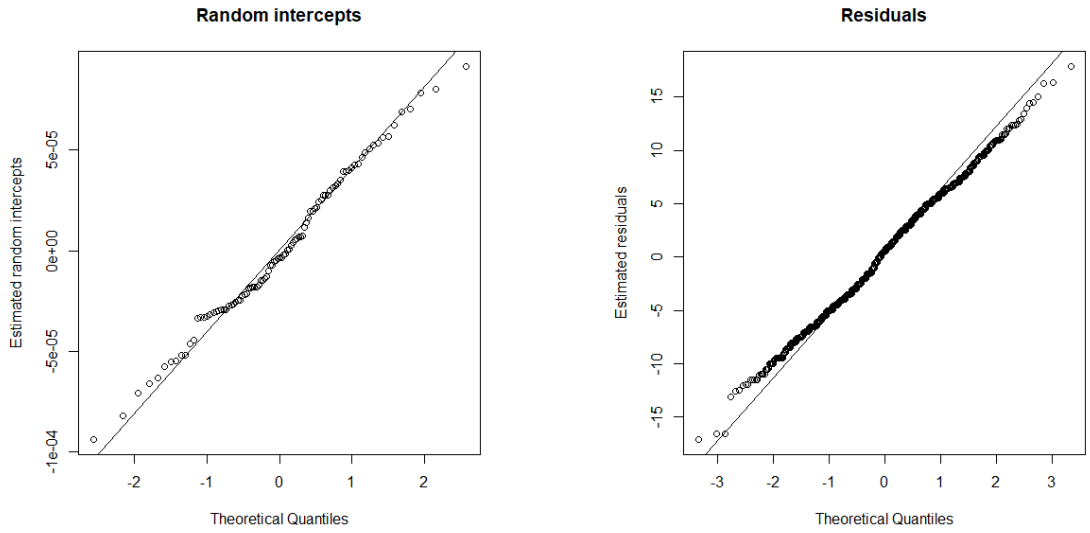


Figure 18: Quantile-quantile plots of the predicted random intercepts and residuals for the LMEE model fitted to the data

Figure 19, by turn, presents the plot of the subject specific residuals versus the fitted responses. This residual plot does not indicate any deviations from a linear form. It also shows relatively constant variance across the fitted range.

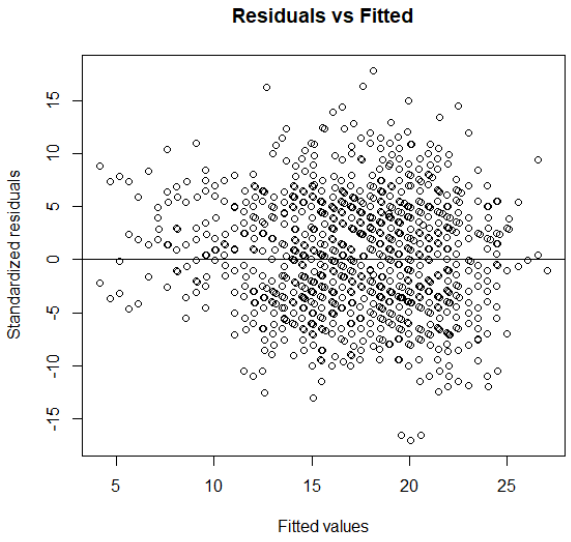


Figure 19: Residuals versus the fitted responses of OQ-10.2 values

Group differences in explanatory variables effect

In order to investigate if explanatory variables effect on mean progression of symptomatology level differ for censored and premature termination cases, we added the variable status to the LMEE previous model and tested several interaction. Thus, the final model include one significant interaction between professional status and status. Table 18 presents the estimated parameters (Estimate) of the fitted longitudinal model with a random intercept an individual level, and an exponential correlation structure.

Table 18: Estimated parameters (Est), standard errors (SE) and p-values for Longitudinal Model comparing groups for explanatory variables effect. The response variable is symptomatology level

	Est (SE)	p-value
Intercept	23.473 (1.093)	<0.001 ***
Time (Session)	- 0.503 (0.045)	<0.001 ***
Status (Ref = Censored)		
PT	- 3.149 (1.513)	0.041 *
Professional status (Ref = Employed)		
(Unemployed)	1.944 (2.281)	0.396
(Student)	2.048 (1.105)	0.067†
Unemployed * Status (= PT)	10.581 (4.107)	0.012 *
Student * Status (= PT)	1.072 (2.024)	0.598
Depressive severity at intake (Ref = clinical)		
(non-clinical)	- 2.047 (1.031)	0.05†
Symptomatology level at intake (Ref = clinical)		
(non-clinical)	- 6.544 (1.207)	<0.001 ***
Log Likelihood	-3108.056	
AIC	6242.112	

Note. PT = Premature termination. †Indicates statistical marginal significance at $p < 0.1$; * Indicates statistical significance at $p < 0.05$; *** Indicates statistical significance at $p < 0.001$.

Results show that, there is a significant difference between censored and PT cases, regarding professional status effect. So, considering the unemployed clients, those who discontinue treatment (i.e., premature terminators) will start the mean progression of symptomatology level with an expected increase of 10.581 points, compared to censored cases.

4.4. JOINT MODELLING

In the previous longitudinal analysis (section 4.3) we verify that the mean progression of therapeutic alliance quality and symptomatology level (i.e., the response variables) differ slightly according, not only to the final event observed (i.e., PT vs. censored), but also to the moment when this event happened. In fact, the results of the time effect, specially when a negative scale was considered (i.e., time since last session observed), indicate a likely association between survival endpoint and both longitudinal therapeutic alliance and treatment response variables evolution.

Joint modelling approach allow us to understand two processes of interest simultaneously, longitudinal and survival, given that there is an association between them. Thus, with the purpose of evaluating the relationship between PT in psychotherapy and both longitudinal therapeutic alliance and treatment outcome evolution, a joint model specification, implemented in the software R with *joineR* package, was analysed. Furthermore, the parameter estimates and their standard errors, using the joint modelling specification were, compared to those obtained with the independent models, namely a linear mixed model for longitudinal outcome and a Cox's regression model for the survival outcome.

The joint model of the therapeutic alliance evolution and the hazard of PT was adjusted and estimated by "step-wise backwards", starting with the satured model with all baseline covariables, allocated to both submodels, and then eliminating one-by-one the variables with lower significance (with a limit of 0.1 for the p-value, for inclusion of the variable). After adjusting several joint models with multiple covariables, we end up with a shared random-effects joint model that incorporate three covariables with a significant effect on mean progression of the therapeutic alliance quality, and three covariables with a significant effect on the client's risk of PT.

The final model implemented was:

$$\begin{aligned}\text{Longitudinal submodel}: \text{WAI}_{ij} &= \text{Intercept} + \boldsymbol{\beta}_1\, t_{ij} + \boldsymbol{\beta}_2 \text{ if } (\text{PS}_i = \text{unemployed}) \\ &+ \boldsymbol{\beta}_3 \text{ if } (\text{PS}_i = \text{student}) + \boldsymbol{\beta}_4\, \text{BDI-II}_i + \boldsymbol{\beta}_5\, \text{OQ-45.2}_i + U_{0i} + U_{1i} t_{ij} + \boldsymbol{\epsilon}_i(t_{ij})\end{aligned}$$

$$\begin{aligned}\text{Survival submodel}: \text{Hazard}_i &= \boldsymbol{\beta}_6 \text{ if } (\text{Diagnoses = Mood})_i + \boldsymbol{\beta}_7 \text{ if } (\text{Diagnoses = Other})_i \\ &+ \boldsymbol{\beta}_8\, \text{OQ-45.2}_i + \boldsymbol{\beta}_9\, \text{OQ-10.2 last session}_i + \gamma_0\, U_{0i} + \gamma_1\, U_{1i} \text{Time}_i\end{aligned}$$

Table 19 presents the results from joint modelling, namely the parameter estimates (Est) and respective standard errors (SE) and p-values.

Table 19: Estimated parameters (Est), standard errors (SE) and p-values for joint model fitted to therapeutic alliance quality (longitudinal outcome) and time to premature termination (survival outcome)

	Est(SE)	p-value
	Longitudinal Process	
Intercept	54.394 (1.981)	<0.001 ***
Time (Session)	0.391 (0.059	<0.001***
Professional status (Ref = Employed)		
(Unemployed)	- 3.014 (2.241)	0.179
(Student)	- 1.856 (1.124)	0.099†
Depressive severity at intake		
(BDI-II score)	0.205 (0.088)	0.021 *
Symptomatology level at intake		
(OQ-45.2 score)	- 0.108 (0.037)	0.003 **
	Survival Process	
Diagnoses (Ref = Anxiety)		
(Mood)	0.869 (0.461)	0.059†
(Other)	1.017 (3.863)	0.792
Symptomatology level at intake		
(OQ-45.2 score)	- 0.034 (0.013)	0.006 **
Symptomatology level at last session		
(OQ-10.2 score)	0.094 (0.031)	0.002 **
	Latent Association	
U_0	- 0.066 (0.034)	0.057 †
U_1	0.154 (0.047)	<0.001 ***
Log Likelihood	- 3316.328	

Note. Ref = Reference Category; †Indicates statistical marginal significance at $p < 0.1$; * Indicates statistical significance at $p < 0.05$; ** Indicates statistical significance at $p < 0.01$; *** Indicates statistical significance at $p < 0.001$.

The significance of the latent random effects (U_0 and U_1) indicates the relationship between the longitudinal and survival processes, verifying what was shown in previous exploratory longitudinal analysis (section 4.3.1) and highlighting the need for a joint modelling of this type of data.

By inspection of the parameter estimates, namely the latent random effect U_0, we can infer that clients with initial therapeutic alliance scores that are lower than the population average tend to have higher hazard of PT (HR = exp(- 0.066) = 0.936). Specifically, a unit decrease in WAI score, corresponds to a 6.387% fold increase in the risk for PT. Overall, there is an increase rate of 0.391 points per session of the therapeutic alliance quality. Besides that, from parameter U_1, we can infer a significant association between the therapeutic alliance mean progression over time and the survival of clients. Namely, those who have a therapeutic alliance mean progression (i.e., slope) that is higher than the population average tend to have a better survival (HR = exp(0.154) = 1.166).

5

DISCUSSION

Over the last decades, joint models of longitudinal and survival data have received much attention in literature. Widely recognized for their gain in efficiency, as well as reduction in bias compared to naive methods, joint models are often desirable in the following situations: (i) survival models with measurement errors or missing data in time-dependent covariates; (ii) longitudinal models with informative dropouts; and, (iii) survival and longitudinal processes are associated via latent variables.

The literature about this theme is quite extensive, and some review papers (Rizopoulos, 2012; Sousa, 2011; Tsiatis & Davidian, 2004) present and discuss various types of joint models. Basically, depending on the focus of the analysis, as also on the selected submodels to handle either survival and longitudinal processes, different formulations of joint models can be considered. A typical setup in the literature, adopted in the present work, is the link of a Cox Proportional Hazards survival model with a Linear Mixed-Effects longitudinal model through unobserved shared latent random effects.

Random shared effects joint models have been used in a wide range of healthcare applications. These include immune deficiency disease (HIV/AIDS) (e.g.,Tsiatis et al., 1995; Wu et al., 2010), leukemia and cancer (Ibrahim et al., 2010), cardiovascular and kidney disease or transplant data (e.g., Andrinopoulou et al., 2012; McCrink et al., 2013; Teixeira et al., 2019), and cognitive decline (e.g., Henderson et al., 2000; Proust-Lima et al., 2019), to name a few. Although the majority of applications have a medical focus, there is no reason to maintain such exclusivity, as other areas could heavily benefit from joint modelling approaches.

This work represents, as far as we known, the first study on PT in psychotherapy area using random shared effects joint modelling approach. The main motivation was to evaluate the association between PT risk and two process variables, namely the progression of therapeutic alliance quality and treatment outcome. First, ignoring a possible association, separate analysis were conducted in order to infer: i) which predictors affected significantly the PT hazard (survival analysis); and ii) the effect of significant predictors on the mean progression of process variables, therapeutic alliance quality and treatment outcome (lon-

gitudinal analysis). Then, joint modelling analysis was conducted taking into consideration the expected association between progression in time of each of the referred process variables with client's PT hazard.

The results obtained from separate survival analysis, namely from multivariate Cox Proportional Hazard model, show that diagnostic areas (i.e., anxiety, mood and others), as also factors related to client initial and final impairment, namely symptomatology level (as measured by OQ-45.2 at intake and by OQ-10.2 at last session observed), as well as therapeutic alliance quality (as measured by WAI at last session observed) are associated with clients ending treatment. Specifically, the hazard to PT is 3.6 times higher for mood disorders and 4.1 times higher for other disorders, when compared to anxiety diagnoses areas. Client's reporting less symptomatology level at intake are more likely to discontinue treatment (i.e., an unit decrease in the OQ-45.2 corresponds to a 4% fold increase in the risk to PT). However, at last session observed, those reporting higher symptomatology level and lower therapeutic alliance quality are more likely to dropout (i.e., an unit increase in OQ-10.2 corresponds to a 8% fold increase in risk to PT and an unit decrease in WAI corresponds to a 9.2% fold increase in the risk to PT). Additionally, univariate analysis regarding depressive severity show that the hazard to PT is two times higher for non-clinical group, specially after session 5 (accordingly to Kaplan-Meier survival curves). The extended Cox model shed to light one time-varying significant predictor, namely the treatment outcome (as measured by OQ-10.2 at beginning of each session). As expected, for any given session, a unit decrease in the OQ-10.2 scores corresponds to a 7.3% fold increase in the risk for PT.

This findings may indicate that mildly impaired clients have not been as motivated, since they were less severely affected, and consequently had less intense need of treatment. Possibly, these participants would benefit from short-term help or a treatment more specifically tailored to their needs (e.g., a brief treatment to cope with specific symptoms, like loneliness and isolation). Besides that, potentially, UP transdiagnostic treatment of emotional disorders deal more effectively with anxiety disorders than other diagnoses areas, given the nature of their strategies (e.g., exposure to both interoceptive and situational cues associated with intense emotional experiences). Notably, results show that hazard to PT is higher for clients with mood disorders and the greatest proportion of dropout occurred after session 4 and 9, where accordingly to UP protocol begins psycho-education about emotional experience and exposure modules, respectively. On the other hand, higher hazard to PT seems to be linked to progress decline for both therapeutic alliance quality and therapeutic outcome, which may, by turn, reflect an underling dissatisfaction with the treatment received, leading clients to discontinue their participation.

Regarding separate longitudinal analysis, the resulting multivariate linear mixed effect model fitted to the data, reveals two predictors with a significant effect on the linear progression of the therapeutic alliance

quality (as measured by WAI at the end of each session), namely: depressive severity (as measured by BDI-II at intake) and symptomatology level (as measured by OQ-45.2 at intake). Specifically, lower depressive severity scores and higher symptomatology level scores both at intake have a decreasing effect on the starting point for the average therapeutic quality progression (i.e., worse therapeutic alliance quality at beginning of treatment). Besides that, the effect of symptomatology level on intercept component of the therapeutic alliance mean progression is 1.775 times higher for PT cases, compared to censored cases. Additionally, there are other variables that, by themselves (i.e., fitting a longitudinal model considering only that specific variable alone) have a statistical effect on the mean progression of the therapeutic alliance. Namely, regarding professional status and diagnoses areas, students present a lower starting point (-2.097) of the mean progression, compared to employed clients, and those with a mood disorder present also a lower starting point (-2.431) of the mean progression, compared to clients with an anxiety disorder. However, despite the overall increasing effect of diagnoses areas predictor on the average therapeutic alliance progression over time, this rate is 1.681 times higher for mood disorders, when compared to a anxiety disorders. Moreover, comparing PT and censored cases, we observe that PT have an expected decrease of -3.277 points on the intercept component at beginning of treatment, as well as an expected decrease of -5.990 points on the intercept at last session observed, compared to censored cases, which means that, at least for the beginning and the end of treatment, PT cases show a worse therapeutic alliance quality.

The findings from longitudinal analysis suggest that PT is linked to worse therapeutic alliance quality mean progression, which by turn is negatively affected by initial impairment (OQ-45.2 scores), and positively affected by depressive severity (BDI-II scores). Accordingly, clients with mood disorder present a lower starting point of the mean progression (compared to clients with anxiety disorders), which once again might reflect motivation problems concerning participation in UP treatment. Unexpectedly, the increasing effect of time on mean therapeutic alliance progression is 1.681 higher for clients with mood disorders, compared to clients with anxiety disorders. However, it is possible that lower starting points coupled to higher hazard to PT, for clients with mood disorders, have resulted in a large average rate of therapeutic alliance mean progression over time. Therefore, the longitudinal profile, in this case, probably reflects more an artefact caused by selective dropout than genuine change over time, which justifies, by turn, the need to adopt models that consider non-ignorable dropout, as random shared effects joint modelling models.

Moving to separate longitudinal analysis of therapeutic progress, the resulting multivariate linear mixed effect model fitted to the data, reveals three predictors with a significant effect on the linear mean progression of the treatment outcome (as measured by OQ-10.2 at the beginning of each session), namely:

professional status, anxiety level (as measured by BAI at intake), and symptomatology level (as measured by OQ-45.2 at intake). Specifically, unemployed and students, when compared to employed cases, start the mean progression of treatment outcome with an expected increase of 5.074 and 2.493 points, respectively , which means a high impairment at beginning of treatment. Regarding anxiety and symptomatology level, both measured at intake, non-clinical groups start the mean progression of treatment outcome with an expected decrease of -2.068 and -6.968 points, respectively, compared to clinical group. Comparing PT an censored cases, regarding the effect of professional status on the mean progression of treatment outcome, we verify that, within the unemployed category, those who discontinue treatment (i.e., PT cases) have an expected increase of 10.581 points at the intercept component of the mean progression. Furthermore, there are other variables that, by themselves (i.e., fitting a longitudinal model considering only that specific variable alone) have a statistical effect on the mean progression of the treatment outcome. Namely, as expected, non-clinical group for depressive severity shows a lower starting point (-7.982), and a lower decline rate (-0.235), compared to clinical group. Regarding relationship status, clients who were married present a lower starting point (-3.697), compared to those who were single, which reflects lower impairment at beginning of treatment. On the other hand, as happened for longitudinal analysis regarding therapeutic alliance progression, clients with mood disorders show an increased starting point (3.108 ; i.e., high impairment) of the mean progression, but surprisingly the decline rate is higher (-3.670), when compared to clients with an anxiety disorder (vs. -3.361), which suggests a greatest improvement for therapeutic progress. Moreover, comparing PT and censored cases, we observe that PT have an expected decrease of -3.710 points in the intercept component of treatment outcome progression. Regarding the longitudinal model considering time since last session observed, no meaningful differences were found. However, premature terminators have a substantially higher estimated value for intercept component, compared to censored cases (15.760 vs. 13.951).

The findings from longitudinal analysis, regarding treatment outcome progress suggest that, even showing a lower starting point for treatment outcome mean progression (which suggest a lower impairment at beginning of treatment), dropout cases will untimely terminate the progression with higher impairment (compared to censored cases). This fact is in line with previous studies that show poor treatment outcomes and continued impairment for clients who prematurely terminate, compared to those who complete treatment (Swift & Greenberg, 2012). The results of our study also suggest that some baseline variables may promote worse starting points, as also the risk for deterioration, regarding treatment outcome mean progression over time, namely: high impairment, more anxiety or depressive symptoms, and being unemployed or student. Within unemployed cases, those who untimely dropped out, show a larger impairment

at starting point of mean progression, when compared to the remaining cases (e.g., ≈ 12.53 points more, on average, compared to employed and censored cases). Nevertheless, recall from sample description (see Table 3) that, in this study, only two clients were unemployed and dropped out, so no generalization should be applied. Regarding relationship status, being married or in a significant relationship (compared to a single status), seems to be a protective factor regarding the intercept component of treatment outcome mean progression, as the impairment is, on average, lower. Finally, along with what happened for longitudinal analysis regarding therapeutic alliance progression, clients with mood disorders, show an increased starting point of the treatment outcome mean progression, but the decline rate is 1.092 times higher, compared to clients with anxiety disorders, which suggest a greats improvement for therapeutic progress. In a similar vein, the same reasoning applies regarding the need to adopt models that consider non-ignorable dropout.

Notably, for all linear mixed effect models fitted, the estimated variance of the measurement error is quite lower than the estimated variance of the ordinary least square (OLS) model, meaning that the earliest explain the variability of the data mainly by means of variability between and within individuals assigning, by turn, a very low value for measurement error (also called white noise).

Joint modelling analysis, as the major strength of this study, revealed that, the association between the longitudinal and survival processes is significant and it is essential its recognition.

REFERENCES

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

Anderson, E., & Lambert, M. (2001). A survival analysis of clinically significant change in outpatient psychotherapy. *Journal of Clinical Psychology*, *57*(7), 875-888. Retrieved from http://doi.wiley.com/10.1002/jclp.1056

Anderson, K. (2015). *Psychotherapy: Predictors, Reasons, and Outcomes* (Doctoral dissertation, University of Nebraska). Retrieved from http://digitalcommons.unl.edu/psychdiss/74

Andrinopoulou, E.-R., Rizopoulos, D., Jin, R., Bogers, A., Lesaffre, E., & Takkenberg, J. J. (2012). An introduction to mixed models and joint modeling: Analysis of valve function over time. *The Annals of Thoracic Surgery*, *93*(6), 1765-1772. Retrieved from http://www.sciencedirect.com/science/article/pii/S0003497512004092

Asar, O., Ritchie, J., Kalra, P. A., & Diggle, P. J. (2015). Joint modelling of repeated measurement and time-to-event data: an introductory tutorial. *International Journal of Epidemiology*, *44*(1), 334-344. Retrieved from https://doi.org/10.1093/ije/dyu262

Barlow, D., Farchione, T., Fairholme, C., Ellard, K., Boisseau, C., Payne, L., & Ehrenreich-May, J. (2011). *Unified protocol for transdiagnostic treatment of emotional disorders*. doi: 10.1093/med:psych/9780199772674.001.0001

Beck, A., Brown, N., & Steer, R. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology*, *56*(6), 893-897. doi: 10.1037//0022-006x.56.6.893

Beck, A., Steer, R., Ball, R., & Ranieri, W. (1996). Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *Journal of Personality Assessment*, *67*(3), 588-597. Retrieved from https://doi.org/10.1207/s15327752jpa6703_13

Björk, T., Björck, C., Clinton, D., Sohlberg, S., & Norring, C. (2009). What happened to the ones who dropped out? Outcome in eating disorder patients who complete or prematurely terminate treatment. *European Eating Disorders Review*, *17*(2), 109-119. Retrieved from http://doi.wiley.com/10.1002/erv.911

Bordin, E. (1979). The generalizability of psychoanalytic concept of working alliance. *Psychotherapy Theory Research Practice Training*, *16*(3), 252-260. Retrieved from https://doi.org/10.1037/h0085885

Brown, E. R., & Ibrahim, J. G. (2003). Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics*, *59*(3), 686-693. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/1541-0420.00079

Bullis, R., Boettcher, H., Sauer-Zavala, S., Farchione, J., & Barlow, D. (2019). What is an emotional disorder? a transdiagnostic mechanistic definition with implications for assessment, treatment, and prevention. *Clinical Psychology: Science and Practice*, *26*(2), 1-19. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/cpsp.12278

Cahill, J., Barkham, M., Hardy, G., Rees, A., Shapiro, D. A., Stiles, W. B., & Macaskill, N. (2003). Outcomes of patients completing and not completing cognitive therapy for depression. *British Journal of Clinical Psychology*, *42*(2), 133-143. Retrieved from http://doi.crossref.org/10.1348/014466503321903553

Campos, B., R., & Gonçalves. (2011). The Portuguese version of the Beck Depression Inventory-II (BDI-II): Preliminary psychometric data with two nonclinical samples. *European Journal of Psychological Assessment*, *27*, 258-264. Retrieved from https://doi.org/10.1027/1015-5759/a000072

Carpenter, J., Del Gaudio, C., & Morrow, R. (1979, dec). Dropouts and terminators from a community mental health center: Their use of other psychiatric services. *Psychiatric Quarterly*, *51*(4), 271-279. Retrieved from http//link.springer.com/10.1007/BF01082830

Cnaan, A., & Laird, P., N.and Slasor. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, *16*(20), 2349-2380. doi: 10.1002/(SICI)1097-0258(19971030)16:20<2349::AID-SIM667>3.0.CO;2-E

Coelho, M. A., R., & Barros, H. (2002). Clinical profiles relating gender and depressive symptoms among adolescents ascertained by the Beck Depression Inventory II. *European Psychiatry*, *17*(4), 222-226. Retrieved from https://doi.org/10.1016/S0924-9338(02)00663-6

Collett, D. (2015). *Modelling survival data in medical research* (third edition ed.).

Corning, A. F., & Malofeeva, E. V. (2004). The Application of Survival Analysis to the Study of Psychotherapy Termination. *Journal of Counseling Psychology*, *51*(3), 354-367. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-0167.51.3.354

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187-220. Retrieved from http://www.jstor.org/stable/2985181

Cox, D. (1975). Partial likelihood. *Biometrika*, *62*(2), 269-276. Retrieved from https://doi.org/10.1093/biomet/62.2.269

Cox, D., & Snell, E. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, *30*(2), 248-265. doi: 10.1111/j.2517-6161.1968.tb00724.x

Diggle, P. (2005). Variogram. In *Encyclopedia of biostatistics*. American Cancer Society. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/0470011815.b2a12078

Diggle, P., Heagertry, P., Liang, K.-Y., & Zeger, S. (2002). *Analysis of longitudinal data*. Oxford University Press.

Diggle, P., Sousa, I., & Chetwynd, A. (2008). Joint modelling of repeated measurements and time-to-event outcomes: The fourth armitage lecture. *Statistics in Medicine*, *27*(16), 2981-2998. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3131

Fitzmaurice, G., Laird, N., & Ware, J. (2004). *Applied longitudinal analysis*. Wiley-Interscience Hoboken, N.J.

Garre, F. G., Zwinderman, A. H., Geskus, R. B., & Sijpkens, Y. W. J. (2008). A joint latent class changepoint model to improve the prediction of time to graft failure. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(1), 299-308. Retrieved from https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-985X.2007.00514.x

Goates-Jones, M., & Hill, C. (2008). Treatment preference, treatment-preference match, and psychotherapist credibility: Influence on session outcome and preference shift. *Psychotherapy: Theory, Research, Practice, Training*, *45*, 61-74. Retrieved from https://doi.org/10.1037/0033-3204.45.1.61

Henderson, R., Diggle, P., & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data . *Biostatistics*, *1*(4), 465-480. Retrieved from https://doi.org/10.1093/biostatistics/1.4.465

Hogan, J., Roy, J., & Korkontzelou, C. (2004). Handling drop-out in longitudinal studies. *Statistics in medicine*, *9*(23), 1455-1497. Retrieved from https://doi.org/10.1002/sim.1728

Horvath, O., & Greenberg, L. (1989). Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology*, *36*(2), 223-233. Retrieved from https://doi.org/10.1037/0022-0167.36.2.223

Ibrahim, J., Chu, H., & Chen, L. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, *28*(16), 2796-2801. Retrieved from https://doi.org/10.1200/JCO.2009.25.0654

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457-481. Retrieved from http://www.jstor.org/stable/2281868

Klein, J., & Moeschberger, M. (2011). *Survival analysis: Techniques for censored and truncated data*. New York: Springer.

Kleinbaum, D., & Klein, M. (2012). Evaluating the proportional hazards assumption. In *Survival analysis: A self-learning text* (3rd ed. ed., p. 161-200). New York, NY: Springer New York. doi: 10.1007/978-1-4419-6646-9_4

Lambert, M., Burlingame, G., Umphress, V., Hansen, N., Vermeersch, D., Clouse, G., & Yanchar, S. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology & Psychotherapy*, *3*, 249-258. doi: 10.1002/(SICI)1099-0879(199612)3:4<249::AID-CPP106>3.0.CO;2-S

Lambert, M., Finch, A., Okiishi, J., Burlingame, G., McKelvey, C., & Reisinger, C. (1998). *Administration and Scoring Manual for the OQ-10.2: An adult Outcome Questionnaire for Screening*

Individuals and Population Outcome Monitoring. American Professional Credentialing Services LLC.

Little, R. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*(431), 1112-1121. Retrieved from http://www.jstor.org/stable/2291350

Little, R., & Rubin, D. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley.

Liu, X. (2016). Chapter 1 - introduction. In X. Liu (Ed.), *Methods and applications of longitudinal data analysis* (p. 1-18). Oxford: Academic Press. Retrieved from https://doi.org/10.1016/B978-0-12-801342-7.00001-0

Machado, P., & Fassnacht, D. (2015). The Portuguese version of the Outcome Questionnaire (OQ-45): Normative data, reliability and clinical significance cut-offs scores. *Psychology and Psychotherapy: Theory, Research and Practice*, *88*, 427-437. doi: 10.1111/papt.12048

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719-748. Retrieved from https://doi.org/10.1093/jnci/22.4.719

McCrink, L. M., Marshall, A. H., & Cairns, K. J. (2013). Advances in joint modelling: A review of recent developments with application to the survival of end stage renal disease patients. *International Statistical Review*, *81*(2), 249-269. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12018

Pekarik, G. (1992). Posttreatment adjustment of clients who drop out early vs. late in treatment. *Journal of Clinical Psychology*, *48*(3), 379-387. doi: 10.1002/1097-4679(199205)48:3<379::AID-JCLP2270480317>3.0.CO;2-P

Philipson, P., Diggle, P., Sousa, I., Kolamunnage-Dona, R., Williamson, P. R., & Henderson, R. (2012). joineR: Joint modelling of repeated measurements and time-to-event data. Comprehensive R Archive Network.

Proust-Lima, C., Philipps, V., & Dartigues, J.-F. (2019). A joint model for multiple dynamic processes and clinical endpoints: Application to alzheimer's disease. *Statistics in Medicine*, *38*(23), 4702-4717. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8328

Quintão, S., Delgado, A., & Prieto, G. (2013). Validity study of the Beck Anxiety Inventory (Portuguese version) by the Rasch Rating Scale model. , *26*, 305-310. Retrieved from https://doi.org/10.1590/S0102-79722013000200010

Ramos, M. (2008). *Análise das características psicométricas da versão portuguesa do Working Aliance Inventory - short revised.*

Reis, B. F., & Brown, L. G. (1999). Reducing psychotherapy dropouts: Maximizing perspective convergence in the psychotherapy dyad. *Psychotherapy: Theory, Research, Practice, Training*, *36*(2), 123-136. doi: 10.1037/h0087822

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data : With applications in r*. Taylor & Francis Group.

Rubin, D. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592. Retrieved from http://www.jstor.org/stable/2335739

Samstag, L. W., Batchelder, S. T., Muran, J. C., Safran, J. D., & Winston, A. (1998). Early identification of treatment failures in short-term psychotherapy: An assessment of therapeutic alliance and interpersonal behavior. , *7*(2), 126-143.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, *69*(1), 239-241. Retrieved from http://www.jstor.org/stable/2335876

Sharf, P. L. H., J., & Diener, M. J. (2010). Dropout and therapeutic alliance: A meta-analysis of adult individual psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, *47*(4), 637-645. Retrieved from https://doi.org/10.1037/a0021175

Sousa, I. (2011). A review on joint modelling of longitudinal measurements and time-to-event. *REVSTAT - Statistical Journal*, *9*(1), 57-81.

Swift, J., & Greenberg, R. (2012). Premature discontinuation in adult psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology*, *80*(4), 547-559. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/a0028226

Swift, J., & Greenberg, R. (2015). *Premature termination in psychotherapy: Strategies for engaging clients and improving outcomes*. APA Books. doi: 10.1037/14469-000

Teixeira, L., Sousa, I., Rodrigues, A., & Mendonça, D. (2019). Joint modelling of longitudinal and competing risks data in clinical research. *REVSTAT - Statistical Journal*, *17*(2), 245-264.

Tolin, D., Gilliam, C., Wootton, B., Bowe, W., Bragdon, L., Davis, E., ... Hallion, L. (2018). Psychometric Properties of a Structured Diagnostic Interview for DSM-5 Anxiety, Mood, and Obsessive-Compulsive and Related Disorders. *Assessment*, *25*(1), 3-13. Retrieved from https://doi.org/10.1177/1073191116638410

Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, *14*(3), 809-834. Retrieved from http://www.jstor.org/stable/24307417

Tsiatis, A. A., Degruttola, V., & Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error: Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, *90*(429), 27-37. Retrieved from https://www.jstor.org/stable/2291126

Willett, J., & Singes, J. (1993). Investigating onset, cessation, relapse, and recovery: Why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *Journal of Consulting and Clinical Psychology*, *61*(6), 952-965. Retrieved from https://doi.org/10.1037/0022-006X.61.6.952

Wolke, D., Waylen, A., Samara, M., Steer, C., Goodman, R., Ford, T., & Lamberts, K. (2009). Selective drop-out in longitudinal studies and non-biased prediction of behaviour disorders. *British Journal of Psychiatry*, *195*(3), 249–256. Retrieved from https://doi.org/10.1192/bjp.bp.108.053751

Wu, L., Liu, W., & Hu, X. J. (2010). Joint inference on HIV viral dynamics and immune suppression in presence of measurement errors. *Biometrics*, *66*(2), 327-335. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2009.01308.x

Wu, L., Liu, W., Yi, G., & Huang, Y. (2012). Analysis of longitudinal and survival data: Joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, *2012*, 1-17. Retrieved from https://doi.org/10.1155/2012/640153

Wulfsohn, M. S., & Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, *53*(1), 330-339. Retrieved from http://www.jstor.org/stable/2533118