Joint Models for Survival and Longitudinal Data, Missing Data, and Sensitivity Analysis, with Applications in Medical Research

Edmund Njeru Njagi

Promoter: Prof. dr. Geert Molenberghs
Co-Promoter: Prof. dr. Geert Verbeke
Co-Promoter: Prof. dr. Marc Aerts

Jury Members:

Prof. dr. Geert Molenberghs
(promoter, Universiteit Hasselt & Katholieke Universiteit Leuven)

Prof. dr. Marc Aerts
(co-promoter, Universiteit Hasselt)

Prof. dr. Geert Verbeke
(co-promoter, Katholieke Universiteit Leuven & Universiteit Hasselt)

Prof. dr. Christel Faes
(adv. committee member, Universiteit Hasselt)

Prof. dr. Michael G. Kenward
(London School of Hygiene and Tropical Medicine, United Kingdom)

Prof. dr. Dimitris Rizopoulos
(Erasmus Universiteit, Rotterdam, the Netherlands)

Prof. dr. Paul Dendale
(Jessa Ziekenhuis & Universiteit Hasselt)

September 26, 2013

Samenvatting

Het is gebruikelijk van verscheidene types respons tegelijk op te meten, eventueel herhaald in de tijd. In kankerstudies worden verscheidene biologische merkers opgetekend, herhaald in de tijd, met daarnaast ook overlevingstijd of tijd tot metastase. Een typisch voorbeeld is prostaatkanker waar, na behandeling, prostaat-specifieke antigenen wordne gemeten, samen met de tijd tot herval (Law *et al.*, 2002; Yu *et al.*, 2004, 2008). In HIV/AIDS studies meet men, naast tijd tot onset van AIDS of overlijden, de virusdruk en T4 celaantallen. Dendale *et al.* (2011) en Njagi *et al.* (2013a,b) beschrijven een cardiologische studie, waarin onderzoekers gebruik maken van telemonitoring (een techniek waarmee patiënten vanop afstand worden gevolgd), niet alleen om bloeddruk op dagelijkse basis te meten, ook hartslag en gewicht worden opgetekend, naast tijd tot heropname. Het gaat hierbij om patiënten met chronische hartaandoeningen. Andersen *et al.* (1993) beschrijven een studie in levercirrhose, waar naast overlevingstijd verscheidene biochemische waarden werden opgetekend, zoals bilirubine, albumine, en prothrombine.

Daarnaast kan de overlevingstijd zelf herhaald gemeten worden. Dit kan bijvoorbeeld het geval zijn omdat het event ter studie verscheidene keren kan optreden, of omdat de tijd gemeten wordt aan alle leden van een familie of cluster. Bijvoorbeeld, de tijd tot heropname in Dendale *et al.* (2011) en Njagi *et al.* (2013a,b) was van het zogenaamde recurrente type, omdat een ontslagen patiënt natuurlijk meer dan één keer kan heropgenomen worden. Duchateau and Janssen (2008) beschrijven een veterinaire studie waar proteïne en ureum concentraties herhaald worden gemeten bij melkkoeien; daarnaast werk ook de tijd tot de eerste inseminatie opgetekend. Omdat koeien clustering vertonen binnen veestapel, is de overlevingstijd hier van het herhaalde type.

In dergelijke studies is het niet ongebruikelijk van de verschillende (herhaalde) metingen gelijk te modelleren, in een zogenaamd *joint model*. Enkele redenen waarom we dit doen zijn als volgt. Ten eerste kan men interesse hebben aan de overlevingstijd,

na correctie voor de longitudinale covariaat. Ten tweede kan men, omgekeerd, het longitudinale profiel corrigeren voor eventuele niet-random uitval, veroorzaakt door het event. Ten derde kan er belangstelling zijn voor de associatiestructuur in dit type van gegevens (Tsiatis and Davidian, 2004; Rizopoulos *et al.*, 2009; Verbeke *et al.*, 2010; Rizopoulos, 2012a; Njagi *et al.*, 2013b). Uiteraard is de onderzoeksvraag niet noodzakelijk beperkt tot de respons in hun originnele vorm. Bijvoorbeeld, het kan wetenschappelijk nodig zijn van een continue respons te dichtomiseren alvorens de analyse uit te voeren. Dit betekent dat ongeveer alle combinaties van mogelijke responsen samen kunnen voorkomen.

Als een respons van het niet-Gaussische type is, doen er zich problemen voor. Bijvoorbeeld, nog buiten de context van joint modelling, hebben Molenberghs *et al.* (2007) en Molenberghs *et al.* (2010) de nadruk gelegd op het modeleren van binomiale gegevens, aantallen, en overlevingstijden. In modellen voor dergelijke gegevens is er een relatie tussen gemiddelde en variantie. Deze relatie wordt niet noodzakelijk gevolgd door de gegevens. De auteurs voerden daarom zogenaamd geconjugeerde random effecten in om deze relatie te versoepelen. We verwijzen naar dit fenomeen als overdispersie. Bij binomiale gegevens en aantallen leidt dat traditioneel tot het beta-binomiale en negatief-binomiale model. Als we dit soort gegevens ook hiërarchisch verzamelen, worden er normale random effecten toegevoegd om de assocatie tussen de herhaalde metingen te vatten. Dit leidt tot het veralgemeend linear gemengd model (GLMM; generalized linear mixed model). Als beide fenomenen zich tegelijk voordoen, worden ook beide types random effect tegelijk gebruikt. Ze verwijzen naar dit soort modellen als *combined model*, waar dus op flexibele manier het gemiddelde en de variantie van de metingen, naast de correlatie tussen metingen, word gemodelleerd. Via data analyse werd aangetoond dat dit uitgebreide model het vaak gevoelig beter doet dan standaard modellen. Zelfs wanneer de klemtoon ligt op eenvoudiger modellen, kan het *combined model* gebruikt worden als goodness-of-fit instrument.

De klemtoon in Hoofdstuk 3 ligt op het overbrengen van het combined model naar de context van joint modeling, waarbij minstens één respons niet-Gaussisch is. Speciale aaandacht gaat uit naar de situatie waarbij ook minstens één van de responsen een overlevingstijd is. Voor dergelijke gegevens maakt men vaak gebruik van *shared-parameter modellen*. Een model voor de overlevingstijd wordt dan gelinkt aan het longitudinale proces via een gemeenschappelijk random effect; conditioneel op dit effect wordt verder onafhankelijkheid verondersteld (Tsiatis and Davidian, 2004; Verbeke *et al.*, 2010; Rizopoulos, 2011, 2012a). De latente structuur kan parametrisch zijn, maar dat hoeft niet. Traditioneel werd de overlevingstijd hoogstens één keer gemeten; vaak werd het longitudinale proces als continu veronderstelt. De random

effecten worden traditioneel meestal als normaal verdeeld beschouwd (Verbeke *et al.*, 2010). Overdispersie wordt daarbij niet in rekening gebracht. Al deze beperkingen worden aangepakt in dit hoofdstuk. Verschillende combinaties zijn mogelijk: overlevingstijd en continu; overlevingstijd en binair; overlevingstijd en aantal. Daarnaast wordt ook het samen voorkomen van een continue en een binaire maat beschouwd. Alle respons kan herhaald gemeten zijn. Ook al werd het niet in detail bekeken, het is mogelijk van meer dan twee responsen te bestuderen. In alle gevallen kan men tegelijk normaal verdeelde en geconjugeerde random effecten invoeren. Via data analyse wordt aangetoond dat een dergelijk uitgebreid model tot betere resultaten kan leiden, daarbij inbegrepen impact op significantie. Integratie over het geconjugeerd random effect is eenvoudig; dit wordt aangegrepen om een efficiënte methode van schatten voor te stellen, gebaseerd op Molenberghs *et al.* (2007). Het is daarnaast ook mogelijk gesloten vormen af te leiden voor de gemeenschappelijke verdeling.

In Hoofdstuk 4 gaan we hierop verder door joint modeling te koppelen aan predictie en onderscheiding van gevallen. We kijken in het bijzonder naar het probleem van dynamische predicties voor heropname bij patiënten met chronisch hartfalen (Njagi *et al.*, 2013a). Hartfalen wordt vaak gecompenseerd door mechanismen in het hart die op den duur zelf tot problemen kunnen leiden. Dit noemt decompensatie. Heropname kan oplopen tot 50% per jaar; het is dus een belangrijk probleem dat veel aandacht kreeg in de literatuur (Chin and Goldman, 1997; Lewin *et al.*, 2005; Chaudhry *et al.*, 2007; Zhang *et al.*, 2009; Dendale *et al.*, 2011). We formuleren het kader van Rizopoulos (2012b, 2011) zodanig dat het een oplossing van dit probleem mogelijk maakt. Eerst wordt er een shared random effect model beschouwd voor tijd tot heropname, gecombineerd met longitudinale merkers (Rizopoulos, 2011; Verbeke *et al.*, 2010; Tsiatis and Davidian, 2004). Meetfout in de merkers wordt meeegenomen (Rizopoulos, 2011). Er wordt een dynamische predictie geformuleerd, die dus evolueert naarmate meer gegevens beschikbaar worden. Naast een statistisch correcte oplossing van dit probleem dat belangrijk is maar voordien nog niet werd aangepakt, krijgt de clinicus een belangrijk tool in handen om de beslissingen in de praktijk te ondersteunen (Njagi *et al.*, 2013a).

Ontbrekende gegevens is een vaak voorkomend probleem. Men heeft drie grote kaders beschikbaar: selectiemodellen (SeM), pattern-mixture modellen (PMM), en shared-parameter modellen (SPM) (Molenberghs and Kenward, 2007). Men onderscheidt ook dire belangrijke mechanismen (Rubin, 1976; Molenberghs *et al.*, 1998; Creemers *et al.*, 2011): MCAR, MAR, en MNAR. Omdat dergelijke modellen per definitie niet verifieerbare aannames maken, is het nodig van sensitiviteitsanalyses uit te voeren (Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005; Creemers

et al., 2010). Door verontderstellingen te variëren krijgt men een idee van de stabiliteit van de conclusies. Er is een verband tussen ontbrekende gegevens in longitudinale studies, en het gemeenschappelijk modelleren van longitudinale gegevens en overlevingstijden. De tweede setting is daarbij ingewikkelder dan de eerste (Njagi *et al.*, 2013c), omdat onvolledigheid op verschieidene manieren kan voorkomen: de overlevingstijd kan gecensureerd zijn, de longitindale sequentie kan onvolledig zijn, of beide kunnen voorkomen. We spreken dan van *coarsening* in de zin dat de gegevens op een minder fijn niveau worden opgemeten dan men in principe zou willen. Deze correspondentie wordt bestudeerd in Hoofdstuk 5. Hierbij wordt een perspectief op joint modeling gehanteerd, verschillend van wat traditioneel wordt gedaan, en gebaseerd op het uitgebreide SPM van Creemers *et al.* (2011). Binnen dit kader kunnen we MAR karakteriseren en makkelijk sensitiviteitsanalyse uitvoeren.

Coarsening is één van twee klassen die samen *verrijkte gegevens* uitmaken. De andere is *augmentatie* waar aan gegevens structuren worden toegevoegd die niet worden geobserveerd doch het modelleren faciliteren. Voorbeelden zijn random effecten, latente klassen, latente veranderlijken en mengverdelingen. Coarsening verwijst naar onvolledige gegevens en censurering bij overlevingstijden. In alle gevallen is een deel van het model aangestuurd enkel door veronderstellingen en niet door gegevens. De gevaren daarvan dienen ten volle onderkend (Verbeke and Molenberghs, 2010; Molenberghs *et al.*, 2012). Waar dit probleem werd aangepakt in Verbeke and Molenberghs (2010) voor onvolledige gegevens en random effecten, beschouwden Molenberghs *et al.* (2012) het in een bredere context. Verscheidene settings komen aan bod in Hoofdstuk Chapter 6. Er is meerbepaald aandacht voor latente klassen en latente veranderlijken, factor analyse, eindige mengverdelingen, en frailty modellen. In alle gevallen is er naast goodness-of-fit ook nood aan sensitviteitsanalyse. Data analyse en niet verifieerbare aannames moeten zorgvuldig geschieden worden (Molenberghs *et al.*, 2012).

Dit werk sluit af met algemene conclusies en aanbevelingen (Hoofstuk 7). Analyse programma's en bijkomende resultaten worden ondergebracht in een appendix.

Acknowledgements

The realization of this work has definitely been through the direct or indirect input of a number of people.

Prof. dr. Geert Molenberghs, my promoter, has provided wonderful guidance, and mentorship, throughout the doctoral period. Geert, the regular meetings that we have had, during the last four years, have been illuminating, and insightful. Throughout that period, the literature that you have recommended, and provided, has been extremely helpful. The so-called "Big Note" is an excellent example in this respect. Geert, you have also been always eager to hear my reflections on the various research issues, and you have always been ready to provide your considered thoughts. It has been a great pleasure working with you.

I would also like to acknowledge my co-promoters, Prof. dr. Geert Verbeke, and Prof. dr. Marc Aerts, as well as all the other members of my thesis jury, Prof. dr. Christel Faes, Prof. dr. Michael G. Kenward, Prof. dr. Dimitris Rizopoulus, and Prof. dr. Paul Dendale. I have had direct, and/or indirect interactions with all of you in various forms, including research collaboration, meetings, seminars and conferences, and even in the review of my thesis. These interactions have not only been cordial, but also very helpful in my research career. I greatly acknowledge all of you.

I would like to thank Prof. dr. Geert Opsomer, of the *Universiteit Gent*, for kindly providing the "time-to-insemination" data, which is analyzed in a section of this work.

I greatly acknowledge the *Universiteit Hasselt*, and the *Bijzonder onderzoeksfonds (BOF)*, for the kind funding of my doctoral research.

It has been a pleasure working within the Interuniversity Institute for Biostatistics and statistical Bioinformatics, and my day-to-day interaction with colleagues at the Center for Statistics side, has been cordial.

Elasma, it has been a great pleasure working with you in the same office for the

last four years. To all my friends at the institute: Elasma Milanzi, Leacky Kamau Muchene, Jürgen Claesen, to name but a few; I greatly acknowledge you.

In a special way I wish to mention the late Arthur Gichuki Gitome; a great friend, and a one-time member of the institute.

Finally, I wish to extend my gratitude to my family. I am grateful to my mum and dad for nurturing me, and providing their unwavering support over the years. Mum and dad, I am greatly indebted to you. I cannot forget my siblings; they have also played important roles in various ways.

Edmund Njeru Njagi
Diepenbeek, September 26, 2013

List of Publications

This thesis text is based on the first four publications in the publication list below.

- **Njagi, E.N.**, Molenberghs, G., Rizopoulus, D., Verbeke, G., Kenward, M.G., Dendale, P., and Willekens, K. (2013). 'A flexible joint-modeling framework for longitudinal and time-to-event data with overdispersion. *Statistical Methods in Medical Research.* Published online before print July 18, 2013, doi: 10.1177/0962280213495994.

- **Njagi, E. N.**, Rizopoulos, D., Molenberghs, G., Dendale, P., and Willekens, K. (2013). 'A joint survival-longitudinal modelling approach for the dynamic prediction of rehospitalization in telemonitored chronic heart failure patients. *Statistical Modelling*, **13**, 179–198

- **Njagi, E.N.**, Molenberghs, G., Kenward, M.G., Verbeke, G., and Rizopoulus, D. (2013c). 'A Characterization of Missingness at Random in a Generalized Shared-parameter Joint modelling Framework for Longitudinal and Time-to-Event Data, and Sensitivity Analysis. *Biometrical Journal*. Under Revision.

- Molenberghs, G., **Njagi, E.N.**, Kenward, M.G., and Verbeke, G. (2012). 'Enriched-data problems and essential non-identifiability. *International Journal of Statistics in Medical Research*, **1**, 16–44.

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

- Tremmery, S., **Njagi, E.N.**, Wampers, M., Buntinx, W., Dirix, N., Correll, C.U., de Decker, A., Molenberghs, G., and De Hert, M. 'Seclusion and restraints in children and adolescents: an 8 year study on prevalence and predictors. (2013). Submitted for publication.

Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike Information Criterion
AUC	Area under the receiver operating characteristic curve
CHF	Chronic Heart Failure
DDI	Dynamic Discrimination Index
LC	Latent Class
LV	Latent Variable
LVEF	Left Ventricular Ejection Fraction
MAR	Missing At Random
MCAR	Missing Completely At Random
MLE	Maximum Likelihood Estimate
MNAR	Missing Not At Random
NTproBNP	N-terminal pro brain natriuretic peptide
NYHA	New York Heart Association
PMM	Pattern Mixture Model
ROC	Receiver Operating Characteristic
SEM	Selection Model
SPM	Shared-parameter Model

Chapter 1

Introduction

The collection of outcomes of different types, especially with one of them being of survival type, is common in research. In cancer studies, several biological markers may be collected over time, together with the time to death or metastasis. An example is in prostate cancer, where, after treatment for cancer, prostate-specific antigen measurements may be collected over time, and the time to disease recurrence recorded (Law *et al.*, 2002; Yu *et al.*, 2004, 2008). In HIV/AIDS studies, apart from the time to onset of AIDS or death, viral load and CD4 cell count may be recorded repeatedly over time (DeGruttola and Tu, 1994; Rizopoulos, 2011). Dendale *et al.* (2011) and Njagi *et al.* (2013a,b) describe a study in cardiology, in which researchers, through telemonitoring (a process through which patients are remotely monitored), not only repeatedly measured daily blood pressure, heart rate and weight from initially discharged chronic heart failure patients, but also recorded the time-to-rehospitalization. Andersen *et al.* (1993) describe a study in liver cirrhosis in which apart from the survival time, several biochemical variables were collected at follow-up, among them bilirubin, albumin, and prothrombin.

Moreover, the time-to-event outcome may also be a repeated outcome. This is the case when it is recurrent, or when it is collected from subjects within clusters. The time-to-rehospitalization in Dendale *et al.* (2011) and Njagi *et al.* (2013a,b) was a recurrent survival outcome, since a discharged patient could be rehospitalized more than once over time. Duchateau and Janssen (2008) describe a study, in veterinary research, where the protein and ureum concentrations were repeatedly measured over time in dairy cows, and the time to first insemination was also recorded. Since cows were clustered within herds, the time-to-event was a repeated outcome.

The objectives thereof may be varied, and some may require a joint modelling

approach. Such objectives include studying the survival outcome, accounting for the longitudinal covariate; studying the longitudinal outcome, accounting for possibly non-random drop-out caused by the occurrence of events; and to examine the association structure between the two outcomes (Tsiatis and Davidian, 2004; Rizopoulos *et al.*, 2009; Verbeke *et al.*, 2010; Rizopoulos, 2012a; Njagi *et al.*, 2013b). Furthermore, research questions do not necessarily have to be restricted to the outcomes in their original form; rather, they may be directed at modified versions of the outcomes, where, say, objectives require a joint analysis of a dichotomized version of an originally continuous longitudinal outcome, and, say, a survival outcome. In this respect, there therefore is no limit as to the possible combinations of different types of outcomes that the statistician is bound to encounter.

When one of the outcomes is non-Gaussian, there are complexities that need to be taken into account. Outside the context of joint modelling, Molenberghs *et al.* (2010), in building on the work by Molenberghs *et al.* (2007), paid emphasis on models for binomial, count, and survival data. It is known that such models exhibit a mean-variance prescription. Whenever data at hand violate this prescription, random effects with a conjugate distribution are usually incorporated, in order to relax the prescription. Note that such violation is referred to as overdispersion. This leads, for instance, to the beta-binomial and the negative-binomial models for binomial and count data, respectively. When these data are obtained within a hierarchical setting, common practice is to use models incorporating normally distributed random effects, in the form of generalized linear mixed models, to account for correlation among repeated measurements from the same experimental unit. The above-mentioned authors highlight that since both overdispersion and hierarchies may occur simultaneously, it would be befitting to consider a model in which, rather than make a choice for either the conjugate or the normal random effects, the strength of both is combined in a single model. In this way, the "combined" model explicitly pays attention to the three structures that exist in the context of hierarchical non-Gaussian outcomes: the mean structure, the variance structure, and the correlation structure. Therefore, simultaneously, the often-restrictive mean-variance relationship is relaxed through conjugate random effects, and the correlation induced by hierarchies is accounted for through normal random effects. Through case studies, the authors show that model fit can be substantially improved by switching to the "combined" framework. This may even have impact on hypothesis tests for covariate effects. Even when interest is on simpler (less flexible) models, the extended framework can still be used as a goodness-of-fit tool.

The focus of **Chapter** 3 will be to transpose such an extended framework to the

context of joint modelling of outcomes of different types, at least one of which is non-Gaussian, with particular emphasis on cases in which one of the outcomes is of survival type. A common approach in the field of joint modelling of longitudinal and time-to-event data is the shared-parameter framework. A sub-model for the time-to-event is linked to one for the longitudinal process using a shared latent structure, conditional on which independence is assumed (Tsiatis and Davidian, 2004; Verbeke *et al.*, 2010; Rizopoulos, 2011, 2012a). The latent structure can but does not have to be parametric. In the parametric case, for instance, shared normal random effects are usually considered; these play the role of taking into account the correlation between the longitudinal measurements, and the association between the time-to-event and longitudinal outcomes (Verbeke *et al.*, 2010). It is evident that most of the work in this field has focussed on cases in which the time-to-event is univariate. It is also evident that most of the focus has been on applications in which the time-to-event is analyzed jointly with a continuous longitudinal outcome. Finally, it is evident that whenever a parametric choice for the shared latent structure is made, in which case it is mostly of the Normal type, focus is only on accounting for the correlation induced by repeated measures in the longitudinal outcome, and the association between the longitudinal and the time-to-event outcome. It is instructive to reiterate here that even univariate survival outcomes can exhibit overdispersion (Molenberghs *et al.*, 2010).

In the said chapter, we will present an extended shared random-effects framework for the joint analysis of pairs of outcomes of different types: time-to-event and continuous; time-to-event and binary; time-to-event and count. We will also consider the additional case of continuous and binary outcomes. We will pay attention to the possibility that each of the different outcomes may be obtained as repeated measures. The framework can easily be extended to analyze any number of different types of outcomes. The flexibility of our proposed framework stems from the fact that in formulating the shared-parameter model, each submodel for a non-Gaussian outcome incorporates both conjugate and normal random effects. Therefore, each submodel for a non-Gaussian outcome reflects all the structures exhibited by repeated measures non-Gaussian outcomes, as highlighted earlier. Through a case study, we will show that the additional flexibility can provide improvement to model fit, and even have impact on significance tests. We will exploit the ease of analytical integration over conjugate random effects to easily estimate the framework via partial marginalization (Molenberghs *et al.*, 2007). Nevertheless, we will show that it is possible to derive closed form-expressions for the fully marginalized joint probabilities; indeed, estimation could proceed by specifying the marginal likelihood contributions, though

existence of infinite series may render the same cumbersome.

In **Chapter** 4, we will delve into an important theme within joint modelling of survival and longitudinal data, namely, dynamic prediction and discrimination. We will address the practical and important problem of dynamically predicting rehospitalization using longitudinal markers such as blood pressure and heart rate, in telemonitored chronic heart failure (CHF) patients (Njagi *et al.*, 2013a). Heart failure is a condition in which the heart fails to pump enough blood for the needs of the body. The body then initiates mechanisms to compensate for the heart's failure. Over time, these mechanisms may overshoot and by themselves cause problems, and this is referred to as cardiac decompensation. Since heart failure patients are known to have a very high rate of rehospitalization, which reaches 50% per year in the most severe cases, the prediction of threatening decompensation is very important in clinical practice. In the heart failure literature, Chin and Goldman (1997), Lewin *et al.* (2005), Chaudhry *et al.* (2007), Zhang *et al.* (2009), and Dendale *et al.* (2011), have addressed various aspects, including the challenges in determining a patient's rehospitalization risk.

We recast the framework in Rizopoulos (2012b, 2011) to propose a solution in this area. The first step will be to fit a shared random effects joint model for the time-to-rehospitalization and the longitudinal marker, with the latter as a time-varying and error-contaminated covariate (Rizopoulos, 2011; Verbeke *et al.*, 2010; Tsiatis and Davidian, 2004). Note that in this step, we will take into account the fact that the marker may be measured with error, an aspect that is clearly ignored in the existing (non-statistical) approaches. The second step involves calculating (patient-specific) conditional survival probabilities, and their respective confidence intervals, based on the fitted model (Rizopoulos, 2011). These probabilities, which take into account the patient's available marker measurements, and get (dynamically) updated as more measurements become available, are a potential additional tool to physicians, in making their intervention decisions. The third step will be to quantify the discriminative ability of the marker (Rizopoulos, 2012b, 2011), hence aiding physicians in assessing the performance of the predictions as provided by the statistical model. This approach does not only provide a sound statistical modelling approach to the substantive problem, a problem which to the best of our knowledge has not previously been addressed using a statistical modelling approach, it also provides clinicians with a valuable additional tool on which to base their intervention decisions, and thus provides immense contribution to heart failure management (Njagi *et al.*, 2013a).

A common problem in research is that of missing data. This has led to the development of selection, pattern-mixture, and shared-parameter missing data modelling frameworks; see for example Molenberghs and Kenward (2007). These frameworks

have further been supplemented with characterizations of missing value mechanisms (Rubin, 1976; Molenberghs *et al.*, 1998; Creemers *et al.*, 2011), under the missing completely at random (MCAR), missing at random (MAR), and the missing not at random taxonomy (MNAR). Given that models for missing data often make unverifiable assumptions about the missing value mechanism, a recurring theme is that of sensitivity analysis (Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005; Creemers *et al.*, 2010). As assumptions regarding the missing value mechanism are varied, the stability of inferences, or lack thereof, provides a guide on the caution with which the inferences need to be embraced.

Undeniably, there is a strong connection between the missing data setting in a longitudinal context, on the one hand, and the joint longitudinal and time-to-event setting, on the other. Conceptually, the two settings actually correspond, but with an added layer of complexity in the latter setting (Njagi *et al.*, 2013c). The additional complexity stems from the fact that data can now be coarsened in various ways: the longitudinal sequence can be incomplete; the time-to-event outcome can be censored; both of these can occur simultaneously. Coarsening refers to the phenomenon that data observed are less refined than the, possibly counterfactual, full data.

This conceptual correspondence will be the focus of **Chapter** 5. We will take a slightly different perspective on joint models than is prevalent in the literature, and argue that conceptually, the two settings actually correspond. Based on this, we will build an extended shared random effects survival-longitudinal joint model, similar in spirit to that of Creemers *et al.* (2011) in the context of longitudinal data subject to missing observations, but now transposed to the current more complex setting. Within the extended framework, we will provide a characterization of MAR, consistent to the one in the missing data setting. We will then provide some reflections on the complexity of model formulation in the extended setting. The extended random-effects structure will then be utilized for sensitivity analysis.

Coarsening, mentioned above, is one of two different phenomena which, together, constitute enriched data. The other is augmentation, which entails supplementing the observed data with latent or unobserved quantities. Augmentation occurs, for instance, in random-effects models, latent class and latent variable models, and finite-mixture models. Coarsening, on the other hand, is encounted in incomplete data and censored survival data settings. The fitting of models for enriched data combines evidence arising from empirical data with non-verifiable model components, *i.e.,* that are purely assumption driven. The user should be fully aware of the potential dangers and pitfalls that follow from this. This is because in general, to any given model an entire class of models can be assigned, with all of its members producing the same fit

to the observed data but arbitrary regarding the unobservable parts of the enriched data. Non-identified parts can be replaced arbitrarily, without altering the fit to the observed data but with potentially non-trivial consequences for inferences and substantive conclusions (Verbeke and Molenberghs, 2010; Molenberghs *et al.*, 2012).

While Verbeke and Molenberghs (2010) addressed this issue in the context of incomplete-data and random effects models, Molenberghs *et al.* (2012) considered it in a broader sense, bringing together a wider range of seemingly disparate settings. In **Chapter** 6, we will focus on latent classes and latent variables, factor analysis, finite mixture models, and frailty models. We will illustrate how non-identified parts can be replaced arbitrarily. It will be made clear that this can be dangerous and the user must carefully reflect on the arbitrary components. For example, they should be supported by substantive considerations or be made part of a sensitivity analysis. Therefore, acceptable goodness-of-fit to the observed data cannot be used as the sole justification for the analysis. In the absence of external corroborating knowledge or information, two alternative routes can be followed. First, it can be made clear that the conclusions drawn have meaning only under the external assumptions built into the analysis. For example, a researcher can choose to draw inferences given a set of scientifically plausible but otherwise non-verifiable causal relationships. It is then important not to divorce the data analysis from the assumptions made. Second, an appropriate sensitivity analysis can be conducted to augment the conclusions (Molenberghs *et al.*, 2012).

General conclusions and recommendations for future research regarding the work presented in this thesis will be presented in **Chapter** 7. Analysis programs, and the calculations and results excluded from the main text, will be provided in the Appendix.

Chapter 2

Analyzed Data Sets

This chapter introduces the data sets that will be considered in this thesis. The Chronic Heart Failure Data is introduced in Section 2.1. Data from a randomized clinical trial in liver cirrhosis are introduced in Section 2.2. A prevention trial, where children who are at a high risk of developing asthma are involved is considered in Section 2.3. Data on the number of insurance policies reporting a certain number of claims in a certain year are introduced in Section 2.4. In Section 2.5, data on the time-to-insemination, for dairy cows, are introduced. We then introduce data related to record times, in track events, in Section 2.6. Finally, in Section 2.7, data on youth risk behavior, are introduced. These data sets will be used in the subsequent chapters.

2.1 The Chronic Heart Failure Data

These data originated from a study conducted in Belgium between 2008 and 2010, and whose objective was to study whether follow-up of chronic heart failure (CHF) patients, by means of a telemonitoring program, reduced mortality and rehospitalization rates (Dendale *et al.*, 2011). Heart failure is a condition in which the heart has difficulty pumping enough blood for the needs of the body. Daily measurements of systolic and diastolic blood pressure, heart rate and weight, were remotely collected from 80 patients. This was through a set of apparatuses availed to the patients at hospital discharge, and through which they not only made the measurements, but through which the measurements were remotely availed to the medical personnel. These longitudinal measurements were recorded each day for a period of about 6 months. 16 patients had rehospitalizations; 13 of these once, 2 twice, and 1 thrice. In addition, the following patient characteristics were also collected at baseline: patient's

sex, age, heart rhythm, NTproBNP as a measure of cardiac muscle fiber stretch, patient fitness indicator (given using the New York Heart Association, NYHA, class indication), and the Left Ventricular Ejection Fraction (LVEF), which is a measure of heart performance.

The data are analyzed in Chapters 3 and 4.

2.2 The Liver Cirrhosis Data

These data are from a randomized clinical trial, conducted in Copenhagen, of which the goal was to assess whether Prednisone prolonged survival in patients with liver cirrhosis (Andersen *et al.*, 1993). Liver cirrhosis is a disease in which the liver function deteriorates due to injury. Patients were enrolled between 1962–1969, randomized to either Prednisone or placebo, and followed-up until 1974. Follow-up visits were planned at 3, 6, and 12 months after treatment, and once a year thereafter. We use data on 251 and 237 patients from the Prednisone and placebo groups, respectively. Among other variables recorded at follow-up were several biochemical variables, including bilirubin, albumin, and prothrombin. In this analysis, we consider the quasi-continuous prothrombin index, viewed a marker for severity of liver fibrosis.

These data are considered in Chapter 5.

2.3 Data on Recurrent Asthma Attacks in Children

These data have also been used by Molenberghs *et al.* (2010) and Duchateau and Janssen (2008). The setting is a prevention trial, where children, who are between 6 and 24 months, and who are at a high risk of developing asthma, are involved. They are randomized, before they experience asthmatic attacks, to the study drug and placebo, and the attacks that occur are recorded. Since a patient will typically experience more than one event, there is clustering. Additionally, during the entire observation period, a patient will have different at risk times, separated by a period of attack or a period of no observation. We present part of the data in Table 2.1, in calendar-time format, where the time at risk is the time from the end of previous to the start of the next event. The end of each period will correspond to either an event or no event.

Table 2.1: *Asthma data: The first four data points for the first two patients.*

Patient ID	Drug	Begin	End	Status
1	0	0	15	1
1	0	22	90	1
1	0	96	325	1
1	0	329	332	1
2	1	0	180	1
2	1	189	267	1
2	1	273	581	1
2	1	582	600	0

These data are considered in Chapter 6, in the context of frailty models in survival.

2.4 Accident Insurance Policies Data

Böhning (2000) analyzes data on claims made by 9461 accident insurance policies issued by La Royale Belge Insurance Company. These data have been attributed to Thyrion (1960), and have also been used by Simar (1976), as well as by Carlin and Louis (1996). The data are on the number of policies reporting a certain number of claims in a certain year.

Table 2.2: *Accident insurance policies data of Thyrion (1960).*

Count (No. of claims)	0	1	2	3	4	5	6	7
Frequency (No. of policies)	7840	1317	239	42	14	4	4	1

We use these data in the context of finite mixture models, within Chapter 6.

2.5 Time-to-insemination Data

These data are collected to assess factors associated with time to insemination in dairy heifer cows (Duchateau *et al.*, 2005). Dairy farmers aim for a calving interval between 12 and 13 months. The time from parturition to first insemination is a main factor determining this interval. Duchateau *et al.* (2005) analyze data on the time-to-insemination for dairy cows, which were clustered within herds (farms). Some cows failed to get inseminated, and some were culled before insemination, thus there was censoring. We will focus on the covariate "parity," which is the number of times the

cow has already calved, and which is dichotomized into "primiparous" and "multiparous" cows. Duchateau and Janssen (2008) have also analyzed the data in terms of this covariate.

These data are considered in Chapter 6, in the context of survival frailty models.

2.6 National Track Records for Women

These data are on records for 7 women track events (Johnson and Wichern, 2007). For each of the seven events (100, 200, 400, 800, 1500, and 3000 metres, and the marathon), the record times are provided for $n = 54$ countries.

These data are considered in Chapter 6, in the context of factor analysis.

2.7 The 2005 United States' National Youth Risk Behavior Survey Data

This survey, conducted by the US Centers for Disease Control and Prevention, targets youths in grades 9–12, and the questions of interest are on various health-risk behaviors. These include alcohol and drug use, sexual behavior, dietary habits, and physical activity. The Youth Risk Behavior Surveillance System aims at among others to monitor the trends of health-risk behavior and to assess the impact of efforts to combat the same. We pay attention to 12 questions relating to smoking, alcohol consumption, consumption of other drugs, and sexual behavior. Collins and Lanza (2009) have previously extensively analyzed these variables, in the context of latent class models.

These data are considered in Chapter 6, in the context of latent class models.

Chapter 3

A Flexible Joint Modeling Framework for Longitudinal and Time-to-event Data With Overdispersion

In this chapter, we combine conjugate and normal random effects in a joint model for outcomes, at least one of which is non-Gaussian, with particular emphasis on cases in which one of the outcomes is of survival type. Conjugate random effects are used to relax the often-restrictive mean-variance prescription in the non-Gaussian outcome, while normal random effects account for not only the correlation induced by repeated measurements from the same subject, but the association between the different outcomes. Though the normal random effects also account for some amount of overdispersion, it is not in a sufficiently flexible manner; this is often overlooked. Using a case study in chronic heart failure, we show that model fit can be improved, even resulting in impact on significance tests, by switching to our extended framework. By first taking advantage of the ease of analytical integration over conjugate random effects, we easily implement our framework, using maximum likelihood, in standard software.

3.1 Introduction

The collection of outcomes of different types from study subjects is common in research. In HIV/AIDS studies, apart from the time to onset of AIDS or death, viral

load and CD4 cell count may be recorded repeatedly over time. In cancer studies, several biological markers may be collected over time, together with the time to death or metastasis. An example is in prostate cancer, where, after treatment for cancer, prostate-specific antigen measurements may be collected over time, and the time to disease recurrence recorded (Yu *et al.*, 2004, 2008). Andersen *et al.* (1993) describe a study in liver cirrhosis in which apart from the survival time, several biochemical variables were collected at follow-up, among them bilirubin, albumin, and prothrombin. Moreover, the time-to-event outcome may also be a repeated outcome. This is the case when it is recurrent, or when it is collected from subjects within clusters. Dendale *et al.* (2011) and Njagi *et al.* (2013a) describe a study in cardiology, in which researchers, through telemonitoring (a process through which patients are remotely monitored), not only repeatedly measured daily blood pressure, heart rate and weight from initially discharged chronic heart failure patients, but also recorded the time-to-rehospitalization. Time-to-rehospitalization in this case was a recurrent survival outcome, since a discharged patient could be rehospitalized more than once over time. Duchateau and Janssen (2008) describe a study, in veterinary research, where the protein and ureum concentrations were repeatedly measured over time in dairy cows, and the time to first insemination was also recorded. Since cows were clustered within herds, the time-to-event was a repeated outcome.

The objectives thereof may be varied, and some may require a joint modelling approach. Interest may be on the distribution of one of the outcomes, conditional on the other. For instance, the distribution of the time-to-event, conditional on the history of the "true" longitudinal biomarker process (Tsiatis and Davidian, 2004; Verbeke *et al.*, 2010; Rizopoulos, 2011). It may also be of interest whether covariates have an effect on all the outcomes jointly. Furthermore, research questions do not necessarily have to be restricted to the outcomes in their original form; rather, they may be directed at modified versions of the outcomes, where, say, objectives require a joint analysis of a dichotomized version of an originally continuous longitudinal outcome, and, say, a survival outcome. In this respect, there therefore is no limit as to the possible combinations of different types of outcomes that the statistician is bound to encounter.

Now, on the one hand, though not in the context of joint modelling, Molenberghs *et al.* (2010) in building on the work by Molenberghs *et al.* (2007) paid emphasis on models for binomial, count, and survival data. It is known that such models exhibit a mean-variance prescription. Random effects with a conjugate distribution are usually incorporated in order to relax this prescription, whenever data at hand violate it. Note that such violation is referred to as overdispersion. This leads, for instance,

to the beta-binomial and the negative-binomial models for binomial and count data respectively. Whenever these data are obtained within a hierarchical setting, common practice is to use models incorporating normally distributed random effects, in the form of generalized linear mixed models, to account for correlation among repeated measurements from the same experimental unit. The above-mentioned authors highlight that since both overdispersion and hierarchies may occur simultaneously, it would be befitting to consider a model in which, rather than make a choice for either the conjugate or the normal random effects, the strength of both is combined in a single model. In this way, the "combined" model explicitly pays attention to the three structures that exist in the context of hierarchical non-Gaussian outcomes: the mean structure, the variance structure, and the correlation structure. Therefore, simultaneously, the often-restrictive mean-variance relationship is relaxed through conjugate random effects, and the correlation induced by hierarchies is accounted for through normal random effects. Through case studies, the authors show that model fit can be substantially improved by switching to the "combined" framework. This may even have impact on hypothesis tests for covariate effects. Even when interest is on simpler (less flexible) models, the extended framework can still be used as a goodness-of-fit tool.

Turning attention to joint modelling, on the other hand, the shared-parameter modelling framework (Tsiatis and Davidian, 2004; Verbeke *et al.*, 2010) is a popular approach in the so-called field of joint modelling of longitudinal and time-to-event data. It entails postulating a submodel for the longitudinal outcome, a submodel for the time-to-event, and linking the two submodels through a shared latent structure. The latent structure can but does not have to be parametric. In the parametric case, for instance, shared normal random effects are usually considered; these play the role of taking into account the correlation between the longitudinal measurements, and the association between the time-to-event and longitudinal outcomes (Verbeke *et al.*, 2010). The field has, and is currently, experiencing rapid growth, and excellent reviews are provided by Tsiatis and Davidian (2004) and Yu *et al.* (2004) It is evident that most of the work in this field has focussed on cases in which the time-to-event is univariate. It is also evident that most of the focus has been on applications in which the time-to-event is analyzed jointly with a continuous longitudinal outcome. Finally, it is evident that whenever a parametric choice for the shared latent structure is made, in which case it is mostly of the Normal type, focus is only on accounting for the correlation induced by repeated measures in the longitudinal outcome, and the association between the longitudinal and the time-to-event outcome. It is instructive to reiterate here that even univariate survival outcomes can exhibit overdispersion

(Molenberghs *et al.*, 2010).

In this chapter, we present an extended shared random effects framework for the joint analysis of pairs of outcomes of different types: time-to-event and continuous; time-to-event and binary; time-to-event and count. We also consider an additional case of continuous and binary outcomes. We pay attention to the possibility that each of the different outcomes may be obtained as repeated measures. The framework can easily be extended to analyze any number of different types of outcomes. The flexibility of our framework stems from the fact that in formulating the shared-parameter model, each submodel for a non-Gaussian outcome incorporates both conjugate and normal random effects. Therefore, each submodel for a non-Gaussian outcome reflects all the structures exhibited by repeated measures non-Gaussian outcomes, as highlighted earlier. This work therefore builds on Molenberghs *et al.* (2010) to make shared-parameter joint modelling more flexible. Through a case study, we show that the additional flexibility can provide improvement to model fit, and even have impact on significance tests. Our choice for conjugate random effects is based on the attendant ease of analytical integration, which we exploit to easily estimate the model via partial marginalization (Molenberghs *et al.*, 2007). Nevertheless, we show that it is possible to derive closed form-expressions for the fully marginalized joint probabilities; indeed, estimation could proceed by specifying the marginal likelihood contributions, though existence of infinite series may render the same cumbersome.

In the following section, we briefly review the *Weibull-gamma-normal*, *probit-beta-normal*, and the *Poisson-gamma-normal* models (Molenberghs *et al.*, 2010), which constitute the ingredients for our framework. In Section 3.3, we introduce our modeling framework, followed by a discussion on estimation in Section 3.4. The data are analyzed in Section 3.5.

3.2 Review of Ingredients

In this section, we briefly review the extended generalized linear modelling framework for repeated measures, discussed by Molenberghs *et al.* (2010), and which combines conjugate and normal random effects. We only highlight the specific cases of the *Weibull-gamma-normal*, *probit-beta-normal*, and the *Poisson-gamma-normal* models. We also mention the linear mixed model, for completeness purposes, and since it will, together with the aforementioned, be an ingredient for our framework.

3.2.1 *Weibull-gamma-normal Model*

For repeated measures time-to-event outcomes, the above-mentioned authors present the following, for the k^{th} survival time in cluster i, $k = 1, \ldots, p_i$:

$$\begin{aligned} f(t_{ik}|\psi_{ik}, \boldsymbol{b}_i) &= \lambda\rho {t_{ik}}^{\rho_k - 1}\psi_{ik}e^{\mu_{ik}+d_{ik}}e^{-\lambda_k {t_{ik}}^{\rho_k}\psi_{ik}e^{\mu_{ik}+d_{ik}}}, && (3.1)\\ \psi_{ik} &\sim \text{Gamma}(\alpha_k, \beta_k), && (3.2)\\ d_{ik} &= \boldsymbol{w}'_{ik}\boldsymbol{b}_i, \boldsymbol{b}_i \sim N(0, D), && (3.3) \end{aligned}$$

where $\mu_{ik} = \boldsymbol{x}'_{ij}\boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is a vector of fixed-effects parameters. Forthwith, whenever $\boldsymbol{b}_i$ is used, the above distributional assumption will apply. The baseline hazard function is therefore assumed to follow the parametric form leading to Weibull distribution for the event times. The gamma random effects, of conjugate distribution to the Weibull, are assumed independent. This is not too restrictive, as correlation between the repeated measures is induced by normal random effects. However, dependence between the gamma random effects can also be assumed. One could also consider a common gamma random effect for a cluster:

$$\psi_i \sim \text{Gamma}(\alpha, \beta).$$

3.2.2 *Probit-beta-normal Model*

For repeated measures binary outcomes, the authors consider both the logistic-beta-normal and the probit-beta-normal. The distinction is in usage of the logit link for the former, and the probit for the latter. Here, however, we only review the latter. Closed form expressions are possible with the latter, and in case interest is on the former, there exists back-transformation approximations. With Y_{ij} the j^{th} measurement in cluster i, $j = 1, \ldots, n_i$, they consider the following:

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(\pi_{ij}), && (3.4)\\ \pi_{ij} &\sim \theta_{ij}\Phi(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i), && (3.5)\\ \theta_{ij} &\sim \text{Beta}(\alpha_1, \beta_1), && (3.6) \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the normal distribution. $\boldsymbol{\gamma}$ is a vector of fixed-effects parameters. The model combines both normal ($\boldsymbol{b}_i$) and conjugate (beta) random effects. The vectors $\boldsymbol{x}'_{ij}$ and $\boldsymbol{z}'_{ij}$ are rows stacked into the design matrices X_i and Z_i , corresponding to the fixed and random effects respectively. This will hold whenever these vectors are used.

3.2.3 *Poisson-gamma-normal Model*

For repeated measures count data, the authors consider, with Y_{ij}, $j = 1, \ldots, n_i$ the j^{th} count in the i^{th} cluster, the following model:

$$\begin{aligned} P(Y_{ij} = y_{ij}|\theta_{ij}, \boldsymbol{b}_i) &= \frac{1}{y_{ij}!}\left(\theta_{ij}e^{\tau_{ij}+\boldsymbol{z}'_{ij}\boldsymbol{b}_i}\right)^{y_{ij}} e^{-\theta_{ij}e^{\tau_{ij}+\boldsymbol{z}'_{ij}\boldsymbol{b}_i}}, \\ \theta_{ij} &\sim \text{Gamma}(\alpha_j, \beta_j), \\ \tau_{ij} &= \boldsymbol{x}'_{ij}\boldsymbol{\zeta}, \end{aligned} \tag{3.7}$$

where $\boldsymbol{\zeta}$ is a vector of fixed-effects parameters. Both normal ($\boldsymbol{b}_i$) and conjugate (gamma) random effects are incorporated; as in the Weibull-gamma-normal case, various settings for the gamma random effect could be considered.

3.2.4 *Linear Mixed Model*

Models for continuous outcomes, just like those for survival, binary, and count data, are not only brought together in the generalized linear modelling framework, but also in the exponential family mathematical framework. The mean and variance are separate parameters in the case of continuous normally distributed outcomes. Random effects with a normal distribution, which provide a conjugate distribution to the normal linear model, are incorporated to induce intracluster correlation in repeated measures continuous outcomes. For the j^{th} measurement in cluster i, $j = 1, \ldots, n_i$,

$$\boldsymbol{Y}_{ij}|\boldsymbol{b}_i \sim N(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i, \sigma^2) \tag{3.8}$$

is the hierarchically specified linear mixed model, where $\boldsymbol{\beta}$ is a vector of fixed-effects parameters.

3.3 Flexible Joint Modelling Framework

In the foregoing section, we have highlighted the *Weibull-gamma-normal*, *probit-beta-normal*, and the *Poisson-gamma-normal* framework, discussed by Molenberghs *et al.* (2010). The framework explicitly addresses each of the structures characteristic of hierarchical non-Gaussian outcomes: the mean, the variance, and the correlation structure. The mean-variance prescription is relaxed, and correlation is accounted for, using two different sets of random effects. One of the advantages of switching to this extended framework is the gain in terms of model fit. This can also have impact on tests for covariate effects. The ease of analytical integration over conjugate random effects also makes the framework easy to estimate by maximum likelihood, via partial

marginalization (Molenberghs *et al.*, 2007). Partial marginalization involves analytical integration over the conjugate random effects, leaving the normal random effects for numerical integration, using such a tool as the SAS procedure NLMIXED. Finally, even when interest is on simpler (less flexible) models, the extended framework can still be used as a goodness-of-fit tool.

We use this framework to develop an extended and flexible shared random effects joint modelling framework. By allowing the submodels for non-Gaussian outcomes to reflect all the structures exhibited by repeated measures non-Gaussian outcomes, through the incorporation of both conjugate and normal random effects, our work aims at providing a framework which provides improvement to model fit. In the following, we will consider specific cases of pairs of different types of repeated measures outcomes, though extension to any number of different types of outcomes is straightforward, as mentioned earlier.

3.3.1 Case 1: Repeated Time-to-event and Repeated Continuous Outcomes

For the continuous outcome submodel, we formulate the linear mixed model, as in (3.8), for the j^{th} measurement in cluster i, $j = 1, \ldots, n_i$. We reiterate here that $\boldsymbol{b_i}$ follows the distribution as stated in Section 3.2.1, which will hold throughout.

For the time-to-event outcome submodel, we formulate the Weibull-gamma-normal model, specified by (3.1), (3.2), and (3.3), for the k^{th} survival time in cluster i, $k = 1, \ldots, p_i$. The k gamma random effects are assumed independent, with parameters α and β. This could be relaxed, for example, to allow for serial association; however, this possibility will not be pursued here. Moreover, $d_{ik} = \boldsymbol{w}'_{ik}\boldsymbol{b}_i$, where $\boldsymbol{w}'_{ik}$ is a vector of scale factors. For the scale and shape parameters in the baseline hazard, we consider a more general case, where both λ and ρ are allowed to vary between members of a cluster.

The continuous and survival processes are assumed independent, conditional on the shared normal random effects. The joint model, conditional on both the normal and gamma random effects, thus takes the form

$$\begin{aligned} f(\boldsymbol{t}_i, \boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\psi}_i) &= \prod_k \lambda_k \rho_k {t_{ik}}^{\rho_k - 1} \psi_{ik} e^{\mu_{ik} + d_{ik}} e^{-\lambda_k {t_{ik}}^{\rho_k} \psi_{ik} e^{\mu_{ik} + d_{ik}}} \\ &\quad \times \frac{1}{(2\pi)^{\frac{n_i}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{\frac{-1}{2}(\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b}_i)' \Sigma_i^{-1} (\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b}_i)}, \quad (3.9) \end{aligned}$$

with Σ_i an n_i by n_i diagonal covariance matrix with diagonal elements σ^2. Note that the shared random effect in the way considered here is generic. For example,

one can choose $\boldsymbol{z}'_{ij}$ and $\boldsymbol{w}'_{ij}$ such that some random effects are present in the normal-outcome linear predictors, with others influencing the Weibull predictor, and a third set influencing both. As such, our paradigm encompasses both shared as well as correlated random effects.

3.3.2 Case 2: Repeated Time-to-event and Repeated Binary Outcomes

For the binary outcome submodel, we formulate the probit-beta-normal model, given in (3.4), (3.5), and (3.6), for the j^{th} measurement in cluster i, $j = 1, \ldots, n_i$. For the time-to-event outcome submodel, we formulate the Weibull-gamma-normal model, with all the specifications as in Section 3.3.1. Conditional independence of the binary and survival processes, given the shared normal random effects, is again assumed. By first noting that integrating out the beta random effects from the model specified by (3.4), (3.5), and (3.6) can be shown to imply that

$$f(y_{ij}|\boldsymbol{b}_i) = \frac{1}{\alpha_1 + \beta_1}(K_{ij}\alpha_1)^{y_{ij}} \left[(1 - K_{ij})\alpha_1 + \beta_1\right]^{1-y_{ij}}, \tag{3.10}$$

where

$$K_{ij} = \Phi\left(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i\right), \tag{3.11}$$

the form of the joint conditional model can then be seen as

$$\begin{aligned} f(\boldsymbol{y}_i, \boldsymbol{t}_i|\boldsymbol{b}_i, \boldsymbol{\psi}_i) &= \prod_j \frac{1}{\alpha_1 + \beta_1}(K_{ij}\alpha_1)^{y_{ij}} \left[(1 - K_{ij})\alpha_1 + \beta_1\right]^{1-y_{ij}} \\ &\quad \times \prod_k \lambda_k \rho_k t_{ik}{}^{\rho_k - 1} \psi_{ik} e^{\mu_{ik} + d_{ik}} e^{-\lambda_k t_{ik}{}^{\rho_k} \psi_{ik} e^{\mu_{ik} + d_{ik}}}. \end{aligned} \tag{3.12}$$

The conditioning here is only on the gamma and normal random effects, given that the beta random effects in the binary outcome submodel has been integrated out.

3.3.3 Case 3: Repeated Time-to-event and Repeated Count Outcomes

For the repeated counts submodel, we formulate the Poisson-gamma-normal model, given in (3.7), while for the time-to-event outcome submodel, we consider the Weibull-gamma-normal model, with specifications once again as in Section 3.3.1. Conditional independence of the two processes is again assumed, given the shared normal random

effects. The joint model, conditional on the random effects, then takes the form

$$\begin{aligned} & P(\boldsymbol{Y}_i = \boldsymbol{y}_i, \boldsymbol{T}_i = \boldsymbol{t}_i | \boldsymbol{b}_i, \boldsymbol{\Theta}_i, \boldsymbol{\Psi}_i) \\ = & \prod_j \frac{1}{y_{ij}!} \left(\theta_{ij} e^{\tau_{ij} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i}\right)^{y_{ij}} e^{-\theta_{ij} e^{\tau_{ij}} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i} \\ & \times \prod_k \lambda_k \rho_k {t_{ik}}^{\rho_k - 1} \psi_{ik} e^{\mu_{ik} + d_{ik}} e^{-\lambda_k {t_{ik}}^{\rho_k} \psi_{ik} e^{\mu_{ik} + d_{ik}}}. \end{aligned} \tag{3.13}$$

Note that here we have two sets of gamma random effects; one set in the survival, and the other in the count process.

3.3.4 Case 4: Repeated Binary and Repeated Continuous Outcomes

Finally, we consider this commonly encountered setting. For the continuous outcome submodel, we consider the linear mixed model, as specified in Section 3.3.1. For the binary outcome submodel, we consider the probit-beta-normal model specified in (3.4), (3.5), and (3.6). Conditional independence, as described in the foregoing sections, is once again invoked. For notational distinction between the binary and the continuous outcome, $\boldsymbol{Y_{1}}_i$ is used for the continuous outcome, and $\boldsymbol{Y_{2}}_i$ is used for the binary outcome. Precisely, for the binary outcome, letting Y_{2ik} be the k^{th} measurement in cluster i, $k = 1, \ldots, p_i$, we assume:

$$\begin{aligned} Y_{2ik} &\sim \text{Bernoulli}(\pi_{ik}), \\ \pi_{ik} &\sim \theta_{ik} \Phi(\boldsymbol{x}'_{ik}\boldsymbol{\gamma} + \boldsymbol{z}'_{ik}\boldsymbol{b}_i), \\ \theta_{ik} &\sim \text{Beta}(\alpha_1, \beta_1). \end{aligned}$$

Again, note that by integrating out the beta random effects, these model specifications for the binary process can be shown to imply that

$$f(y_{2ik}|\boldsymbol{b}_i) = \frac{1}{\alpha_1 + \beta_1} (M_{ik}\alpha_1)^{y_{2ik}} \left[(1 - M_{ik})\alpha_1 + \beta_1\right]^{1 - y_{2ik}}, \tag{3.14}$$

where

$$M_{ik} = \Phi\left(\boldsymbol{x}'_{ik}\boldsymbol{\gamma} + \boldsymbol{z}'_{ik}\boldsymbol{b}_i\right).$$

The joint model for the continuous and the binary process, this time only conditional on the normal random effects, is then

$$\begin{aligned} & f(\boldsymbol{y}_{2i}, \boldsymbol{y}_{1i} | \boldsymbol{b}_i) \\ = & \prod_k \frac{1}{\alpha_1 + \beta_1} (M_{ik}\alpha_1)^{y_{2ik}} \left[(1 - M_{ik})\alpha_1 + \beta_1\right]^{1 - y_{2ik}} \\ & \times \frac{1}{(2\pi)^{\frac{n_i}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{\frac{-1}{2}[\boldsymbol{y}_{1i} - (X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i)]' \Sigma_i^{-1} [\boldsymbol{y}_{1i} - (X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i)]}. \end{aligned} \tag{3.15}$$

3.4 Estimation and Inference

It is possible to derive the fully marginalized joint model probabilities, and therefore parameter estimation for our framework can proceed through maximum likelihood, by directly specifying the marginal likelihood contributions. As an illustration, we derive this for both Case 1 and Case 3 above. The relevant calculations are highlighted in Appendix A.1. While the marginal probabilities could be directly specified for estimation, the existence of infinite series, as can be seen from the expressions, may make the approach intractable. In all apart from the binary-continuous case, they contain infinite series. However, given the ease of analytical integration over the conjugate random effects, and the availability of software that can numerically integrate over normal random effects, analytical integration and numerical integration can be combined to provide a convenient estimation route. This is the idea behind the partial marginalization approach of Molenberghs *et al.* (2007). It involves first analytically integrating out the conjugate random effects, and leaving the normal random effects for numerical integration, such as implemented in the SAS procedure NLMIXED. Therefore, for estimation, we require only the expressions for the joint distribution marginal over the conjugate but conditional on the normal random effects. In the following, we provide expressions for the partially marginalized models, for our four cases presented above.

Case 1: Repeated Time-to-event and Repeated Continuous Outcomes:

We consider model (3.9), and integrate over the gamma random effects. We then have the following as the joint distribution conditional on the normal random effects:

$$\begin{aligned} f(\boldsymbol{t}_i, \boldsymbol{y}_i|\boldsymbol{b}_i) &= \frac{1}{(2\pi)^{\frac{n_i}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{\frac{-1}{2}[\boldsymbol{y}_i-(X_i\boldsymbol{\beta}+Z_i\boldsymbol{b}_i)]'\Sigma_i^{-1}[\boldsymbol{y}_i-(X_i\boldsymbol{\beta}+Z_i\boldsymbol{b}_i)]} \\ &\quad \times \prod_k \frac{\lambda_k\rho_k t_{ik}^{\rho_k-1}e^{\mu_{ik}+d_{ik}} \times \alpha}{\left(\lambda_k t_{ik}^{\rho_k}e^{\mu_{ik}+d_{ik}} + \frac{1}{\beta}\right)^{\alpha+1} \times \beta^{\alpha}}. \end{aligned} \tag{3.16}$$

Case 2: Repeated Time-to-event and Repeated Binary Outcomes:

Considering model (3.12), and integrating over gamma random effects, leads to the joint distribution conditional on the normal random effects:

$$\begin{aligned} f(\boldsymbol{t}_i, \boldsymbol{y}_i|\boldsymbol{b}_i) &= \prod_j \frac{1}{\alpha_1+\beta_1}(K_{ij}\alpha_1)^{y_{ij}}\left[(1-K_{ij})\alpha_1+\beta_1\right]^{1-y_{ij}} \\ &\quad \times \prod_k \frac{\lambda_k\rho_k t_{ik}^{\rho_k-1}e^{\mu_{ik}+d_{ik}} \times \alpha}{\left(\lambda_k t_{ik}^{\rho_k}e^{\mu_{ik}+d_{ik}} + \frac{1}{\beta}\right)^{\alpha+1} \times \beta^{\alpha}}, \end{aligned}$$

with K_{ij} as in (3.11).

Case 3: Repeated Time-to-event and Repeated Count Outcomes:

We integrate over both $\boldsymbol{\Theta}_i$ and $\boldsymbol{\Phi}_i$ from model (3.13), having

$$\begin{aligned} f(\boldsymbol{t}_i, \boldsymbol{y}_i|\boldsymbol{b}_i) &= \prod_j (\frac{1}{y_{ij}!}) \left(e^{\tau_{ij}+\boldsymbol{z}'_{ij}\boldsymbol{b}_i}\right)^{y_{ij}} \frac{\Gamma(y_{ij}+\alpha_j)}{\Gamma(\alpha_j)\beta_j^{\alpha_j}} \left(e^{\tau_{ij}+\boldsymbol{z}'_{ij}\boldsymbol{b}_i} + \frac{1}{\beta_j}\right)^{y_{ij}+\alpha_j} \\ &\quad \times \prod_k \lambda_k \rho_k t_{ik}^{\rho_k - 1} e^{\mu_{ik}+d_{ik}} \frac{\alpha_k}{\beta_k^{\alpha_k} \left(\lambda_k t_{ik}^{\rho_k} e^{\mu_{ik}+d_{ik}} + \frac{1}{\beta_k}\right)^{\alpha_k+1}} \end{aligned} \quad (3.17)$$

as the joint model, conditional on normal random effects.

Case 4: Repeated Binary and Repeated Continuous Outcomes:

For this case, (3.15) already provides the required expression, given that the beta random effect has already been marginalized out of the model for the binary outcome.

These expressions are all that is required to use, for example, the SAS procedure NLMIXED for estimation and inference.

Note that though we here focus on maximum likelihood estimation, other estimation methods can also be considered; for instance, a Bayesian approach.

3.5 Analysis of the Chronic Heart Failure Data

In this section, we consider the data introduced in Section 2.1. Interest will be on jointly modelling the risk of rehospitalization and the mean number of times a patient's heart rate measurements could be classified as "abnormal", with LVEF as a baseline covariate. Notice that the survival outcome is repeated (recurrent). We jointly model the recurrent time-to-rehospitalization and a count version of the dichotomized longitudinal heart rate. Understanding rehospitalization is important in heart failure management. Heart rate was first dichotomized into "normal" (50–90; coded 0) and "abnormal" (values higher than 90; coded 1). Values less than 50 are not considered for this analysis. During each period in which a patient was not under hospitalization, the number of times that the patient's heart rate measurements were classified as "abnormal" was enumerated, generating a count response. Notice that patients who were rehospitalized and discharged at least once in the course of the study had at least 2 periods in which they were not under hospitalization, separated by a period of hospitalization. As such, the count outcome was also a repeated response. As a covariate, we considered the baseline left ventricular ejection fraction, LVEF.

LVEF indicates the fraction of blood being pumped out of the ventricle with each contraction. We consider two categories for this covariate; {>45% (Dendale *et al.*, 2011); coded 0}, and ($\leq$ 45%; coded 1), forthwith referred to as "preserved ejection," and "reduced ejection," respectively.

For the time-to-rehospitalization submodel, we considered

$$f_{ik}(t|\psi_i, b_i) = \lambda\rho t^{\rho-1}\psi_i e^{\mu_i+d_i} e^{-\lambda(t_{2ik}^{\rho}-t_{1ik}^{\rho})\psi_i e^{\mu_i+d_i}}, \tag{3.18}$$

for the k^{th} risk period of patient i, where $d_i = \kappa b_i$, κ a scale factor, and $b_i \sim N(0, \sigma_b^2)$. Note that t_{1ik} and t_{2ik} respectively represent the beginning and the end of the risk period, where $t_{1ik} \leq t \leq t_{2ik}$. Notice that representation of the at-risk periods was done using the calendar-time format (Duchateau and Janssen, 2008). Further, $\mu_i = X_i\xi$; and $X_i = 1$ if subject had reduced ejection, and 0 otherwise. The gamma random effect, ψ_i, was assumed to follow a one-parameter distribution:

$$\psi_i \sim \text{Gamma}\left(\alpha, \alpha^{-1}\right), \tag{3.19}$$

for identifiability purposes (Duchateau and Janssen, 2008).

For the repeated heart rate counts submodel, on the other hand, with Y_{ij} being the number of times subject i exhibited abnormal heart rate measurements during the j^{th} period in which the subject was not under hospitalization, we considered

$$P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\theta_i, b_i) = \prod_j \frac{1}{y_{ij}!}\left(\theta_i e^{\tau_{ij}+b_i}\right)^{y_{ij}} e^{-\theta_i e^{\tau_{ij}+b_i}},$$

with $\tau_{ij} = \log(\text{time}) + \zeta_0 + \zeta_1 X_i$, X_i representing the baseline left ventricular ejection status as before, and $b_i \sim N(0, \sigma_b^2)$. Notice the inclusion of an offset term, log(time), to account for differing time periods within which the counts were made. Identifiability of both parameters of the gamma distribution in this context is problematic, as discussed by Molenberghs *et al.* (2007). To circumvent this, θ_i was assumed to follow a one-parameter distribution:

$$\theta_i \sim \text{Gamma}\left(\alpha_2, \alpha_2^{-1}\right). \tag{3.20}$$

In line with (3.17), and taking our specifications into account, the joint model for the survival and count outcome, conditional upon the normal random effect, takes the form:

$$\begin{aligned} f(y_{ij}, t|b_i) \quad = \quad & \prod_j \frac{1}{y_{ij}!}\left(e^{\tau_{ij}+b_i}\right)^{y_{ij}} \frac{\Gamma(y_{ij}+\alpha_2)\times\alpha_2^{\alpha_2}}{\Gamma(\alpha_2)\times\left(e^{\tau_{ij}+b_i}+\alpha_2\right)^{y_{ij}+\alpha_2}} \\ & \times \frac{\lambda\rho t^{\rho-1}e^{\mu_i+d_i}\times\alpha^{\alpha+1}}{\left\{\lambda(t_{ik2}^{\rho}-t_{ik1}^{\rho})e^{\mu_i+d_i}+\alpha\right\}^{\alpha+1}}. \end{aligned} \tag{3.21}$$

By assuming that the likelihood contribution of censored observations is that of survival probabilities, we allow censoring of the survival outcome in the following manner:

$$\begin{aligned} f(y_{ij}, t|b_i) &= \prod_j \frac{1}{y_{ij}!}\left(e^{\tau_{ij}+b_i}\right)^{y_{ij}} \frac{\Gamma(y_{ij}+\alpha_2)\times\alpha_2^{\alpha_2}}{\Gamma(\alpha_2)\times(e^{\tau_{ij}+b_i}+\alpha_2)^{y_{ij}+\alpha_2}} \\ &\quad \times \frac{(\lambda\rho t^{\rho-1}e^{\mu_i+d_i})^{\delta_i}\times\Gamma(\alpha+\delta_i)\times\alpha^{\alpha}}{\Gamma(\alpha)\left(\lambda(t_{ij2}^{\rho}-t_{ij1}^{\rho})e^{\mu_i+d_i}+\alpha\right)^{\alpha+\delta_i}}, \end{aligned} \tag{3.22}$$

where δ_i is a censoring indicator, taking the value 1 if the observation is an event time, and 0 otherwise.

Implementation was done in the SAS procedure NLMIXED, and the analysis program is presented in Appendix A.2. Given that we need to define different likelihood contributions for the two different outcomes, we use the general likelihood feature, as provided for in the above procedure. This feature has also been utilized by Molenberghs and Verbeke (2005) in estimating joint models for continuous and discrete outcomes. Starting values were obtained by fitting a separate independent poisson and survival model to the count and survival outcome respectively. The results are summarized in Table 3.1. Alongside results from our extended joint model, results from what we refer to as the conventional model, are shown. The conventional model only incorporates the normal random effect, to account for both the correlation induced by repeated measurements from the same subject, as well as the association between the two responses, as opposed to our extended model which goes a step further to relax the mean-variance prescriptions in the respective submodels, through the inclusion of the gamma random effects, as described above. A 5% level of significance was used.

We first look at the results from our extended model. The test for a joint effect of ejection status on both processes is not statistically significant ($p = 0.1650$), and therefore we conclude that there is no statistically significant evidence of a joint effect of ejection status on both the mean number of abnormal heart rate measurements and the risk of rehospitalization. Based on exponentiation of the relevant parameter estimate, the mean number of abnormal heart rate measurements in patients with reduced ejection is found to be 3.3531 times that of patients with preserved ejection. This effect is at borderline statistical significance ($p = 0.0594$). The risk of rehospitalization for patients with reduced ejection is obtained, by also exponentiating the corresponding parameter estimate, as 5.5168 times that of patients with preserved ejection; however, this effect is clearly not statistically significant ($p = 0.6498$). Finally, the estimate of the scale factor, κ, is positive, implying that patients with a higher mean number of abnormal heart rate measurements have increased risk of re-

hospitalization. However, this should not be over-emphasized, given that it is not statistically significant ($p = 0.3201$).

We now compare results from the extended and the conventional model. Based on an AIC-based comparison, we observe that our extended model provides improvement to model fit, without compromising parsimony. There is a noticeable impact on both the point estimates and standard errors. As noted above, the effect of ejection status on the mean number of abnormal heart rate measurements is borderline significant under the extended model; however, the case is quite different under the conventional model ($p = 0.0901$). There is also a remarkable difference in the scale factor; it is highly significant under the conventional model ($p = 0.0022$), while this is clearly not the case under the extended model, as mentioned above. However, in terms of the hypothesis of a joint effect of ejection status on both processes, the two models provide close results; ($p = 0.1650, 0.1648$) for the extended and the conventional model respectively. It is important to recall that in univariate generalized linear models for non-Gaussian outcomes, overly restrictive variance functions pose the risk of incorrect standard errors (Agresti, 2002). Though the joint modelling case is different since more outcomes are involved, it would be expected that too parsimonious variance structures may pose similar problems. This will be investigated in follow-up simulation studies, to study the impact of omitting the conjugate random effects on specific model parameters, under different scenarios.

Finally, we consider the predictive ability of the extended model as compared to that of the conventional model, with respect to the time-to-event outcome. In assessing predictive performance, we use the approach of concordance, as described by Harrell *et al.* (1996). This approach is based on first considering all pairs of patients, where at least one member of the pair has experienced the event of interest. If the event times in such a pair can be ordered, the pair is considered "usable." The logic is then that for each "usable" pair, the predicted probability of surviving up to a certain point should be larger for the member of the pair who survives longer. Pairs for which this condition holds are considered "concordant", and the proportion of "usable" pairs for which concordance holds, provides the concordance index (Harrell *et al.*, 1996).

The extended and the conventional models were both fitted to the first 90 days' data. Based on the model estimates, and assuming a time zero starting point, the predicted probabilities of remaining rehospitalization-free for 90 days were computed, for each patient. These predictions were compared to the actual first rehospitalization events which occurred between day 91 and day 180, in order to provide the concordance index. Note that in constructing the "usable" pairs, we only consider

the patient's first rehospitalization event in the concerned period (the day 91-day 180 period).

Under the extended model, and once again assuming (3.19) for the distribution of ψ_i, the survival probabilities described above are of the form

$$\begin{aligned} P(T > t|b_i) &= \int_0^\infty [\exp\{-\int_0^t \lambda\rho s^{\rho-1}\psi_i \exp(\mu_i + d_i)\, ds\}] f(\psi_i)\, d\psi_i, \\ &= \frac{\alpha^\alpha}{\{\alpha + \lambda\, t^\rho \exp(\mu_i + d_i)\}}, \end{aligned} \tag{3.23}$$

where $t = 90$.

On the other hand, under the conventional model, they are of the form

$$P(T > t|b_i) = \exp\{-\lambda\, t^\rho \exp(\mu_i + d_i)\}. \tag{3.24}$$

Estimates for the probabilities were computed based on the maximum likelihood estimates of the respective model parameters. For the normal random effects, the empirical Bayes estimates were used.

A total of 434 pairs were usable. Based on the extended model, 328 of these were concordant, 105 discordant, and 1 pair had tied predicted survival probabilities. This therefore provided an index of

$$\frac{328 + 0.5(1)}{434} = 75.6912\%.$$

For the conventional model, on the other hand, 310 were concordant, 123 discordant, and 1 pair had tied probability predictions; this gave an index of

$$\frac{310 + 0.5(1)}{434} = 71.5438\%.$$

Therefore, the extended model seems capable of better discriminating between patients who are going to be rehospitalized within the 90 days, and the ones who will not.

A useful extension that would enable a more comprehensive treatment to the subject of prediction would be to connect our work to that of dynamic prediction, (Rizopoulos, 2011, 2012a,b) where conditional patient-specific probabilities of surviving later time points are computed from a fitted joint model, given the patient's available information up to a certain time point. The so-called dynamic discrimination indices are then computed. Actually, in the estimation of the probability of being rehospitalization free provided by (3.23), we partly utilize the previously recorded information of each patient by plugging in the empirical Bayes estimates of the random effects

b_i, which are the modes of the posterior distribution of the random effects given the observed data of the patient. However, even in such a context, the calculation of discrimination indices in the context of recurrent events is not straightforward, and more work would be required to adequately address this.

Table 3.1: *Chronic Heart Failure Data. Parameter estimates (standard errors) for the extended joint repeated counts/recurrent time-to-event model and for the conventional analysis.*

		Extended model	Conventional model
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Count process			
Intercept	ζ_0	-3.9913 (0.8269)	-5.4085 (0.6501)
LVEF status	ζ_1	1.2099 (0.6326)	1.2247 (0.7138)
Variance (Gamma random effects)	$1/\alpha_2$	2.7528 (1.2335)	
Normal random effects			
Variance	${\sigma_b}^2$	1.3662 (1.6182)	5.4479 (1.2892)
Survival process			
LVEF status	ξ	1.7078 (3.7476)	0.0119 (0.5871)
Scale	λ	0.000045 (0.0004)	0.0047 (0.0056)
Shape	ρ	1.4718 (1.5600)	0.6998 (0.1937)
Variance (Gamma random effects)	$1/\alpha$	7.0263 (12.9652)	
Scale factor	κ	1.8304 (1.8294)	0.4212 (0.1333)
Joint effect of LVEF status (p-value)			
		0.1650	0.1648
Model fit [Akaike information criterion (AIC)]			
		792.7461	806.4592

3.6 Discussion

In this chapter, we have outlined a very broad and extended, hence flexible, joint modelling framework. Apart from the correlation induced by repeated measurements from the same subject and the association between the different outcomes, the often-restrictive mean-variance prescription in the model for the non-Gaussian outcome has explicitly been addressed. This has been done conveniently through the inclusion of conjugate random effects, an aspect that has been exploited to easily estimate the

framework in standard software, through partial marginalization.

Through our analyses of data from the area of chronic heart failure, we have shown that our extended framework provides improvement to model fit, while still maintaining parsimony, and also offers better prediction. We have also observed impact on significance tests.

There is need for follow-up work, involving general assessments through simulation studies, on the effects of omitting the conjugate random effects. More specifically, the impact on specific model parameters, under various scenarios, will be investigated. These scenarios include the level of overdispersion in each of the non-Gaussian outcomes, the amount of censoring, as well as the length of the longitudinal sequence.

This framework also opens avenues for further research work in related areas. It is possible to derive the marginal joint correlation functions, which may be of interest, for example, in surrogate marker evaluation and in psychometrics.

Chapter 4

A Joint Survival-Longitudinal Modelling Approach for the Dynamic Prediction of Rehospitalization in Telemonitored Chronic Heart Failure Patients

Telemonitoring in chronic heart failure involves remote monitoring, by clinicians, daily patient measurements of biomarkers such as blood pressure and heart rate. As a strategy in heart failure management, the aim is for clinicians to use these measurements to predict rehospitalization, so that intervention decisions can be made. This is important for clinical practice since heart failure patients have a very high rehospitalization rate. In this chapter, we present a dynamic prediction approach, based on calculating dynamically-updated patient-specific conditional survival probabilities, and their confidence intervals, from a joint model for the time-to-rehospitalization and the time-varying and possibly error-contaminated biomarker. We quantify the ability of the biomarker to discriminate between patients who are and those who are not going to get rehospitalized within a given time window of interest. This approach does not only provide a sound statistical modelling approach to the substantive problem, a problem which to the best of our knowledge has not previously been addressed using a statis-

tical modelling approach, it also provides clinicians with a valuable additional tool on which to base their intervention decisions, and thus provides immense contribution to heart failure management.

4.1 Introduction

Joint modelling of longitudinal and time-to-event data, in response to the fact that the longitudinal outcome is usually error-contaminated and only intermittently observed, has evolved over the past few decades. The field has developed, starting from the simple so-called Last Value Carried Forward approach, to the two-stage procedures, culminating in the current shared-parameter joint modelling approaches (Verbeke *et al.*, 2010; Tsiatis and Davidian, 2004). Indeed, a lot of work has been done in related aspects, and some key references include Pawitan and Self (1993), DeGruttola and Tu (1994), Taylor, Cumberland, and Sy (1994), Faucett and Thomas (1996), Lavalley and DeGruttola (1996), Hogan and Laird (1997), Hogan and Laird (1998), Tsiatis, DeGruttola, and Wulfsohn (1995), Henderson, Diggle, and Dobson (2000), and Xu and Zeger (2001), with excellent overviews being given by Tsiatis and Davidian (2004) and Yu *et al.* (2004).

Within classical survival analysis, there has been a lot of interest on the ability of models in discriminating between patients who will and patients who will not experience the event of interest. Harrell *et al.* (1996) discussed an index, analogous to the receiver operating curve (ROC), based on comparing survival probabilities for the so-called comparable subjects. Antolini *et al.* (2005) extended the discrimination index to include time-dependent covariates. This included a time-dependent area under the curve (AUC) approach, allowing for the evaluation of the discriminative ability at any time point, as well as overall. Among other key references in this rapidly evolving field of prospective accuracy include Zheng and Heagerty (2007) and Heagerty and Zheng (2005). Within the joint modelling framework, in contrast, there has not been as much work related to discrimination. Yu *et al.* (2008) considered a Bayesian approach to individual prediction of recurrent probabilities using a joint longitudinal survival-cure model. Rizopoulos (2011) recently focused on dynamic prediction of conditional survival probabilities. By first taking into account the endogenous nature of the time-dependent biomarkers, this author discussed a Monte Carlo based procedure for computing dynamic conditional survival predictions, as well as their confidence intervals. He also presented a general definition of prediction rules, based on a more elaborate function of the longitudinal history, as well as the consideration of different threshold values. Similar in spirit to the approach of Antolini *et al.* (2005),

the paper discussed the so-called dynamic discrimination index, following from the time-dependent AUCs, which offers an overall measure of the discriminative ability.

Existing methodology on dynamic prediction and discrimination is applicable in many areas of medical research. These include HIV research, where CD4 counts as a marker for survival may be analyzed to provide conditional survival probabilities for individual patients based on their available information (Rizopoulos, 2012b, 2011). In liver cirrhosis research, prothrombin measurements may be analyzed for similar purposes (Rizopoulos, 2012b). The vast application areas extend to research in breast cancer (Antolini *et al.*, 2005). Therefore, the area of dynamic prediction and discrimination is an important one.

In this manuscript, we explore the practical question of dynamically predicting rehospitalization in telemonitored chronic heart failure (CHF) patients. Heart failure is a condition in which the heart fails to pump enough blood for the needs of the body. The body then initiates mechanisms to compensate for the heart's failure. Over time, these mechanisms may overshoot and by themselves cause problems, and this is referred to as cardiac decompensation. Since heart failure patients are known to have a very high rate of rehospitalization, which reaches 50% per year in the most severe cases, the prediction of threatening decompensation is very important in clinical practice. In the heart failure literature, Chin and Goldman (1997), Lewin *et al.* (2005), Chaudhry *et al.* (2007), Zhang *et al.* (2009), and Dendale *et al.* (2011), have addressed various aspects, including the challenges in determining a patient's rehospitalization risk.

To the best of our knowledge, a statistical modelling approach has not previously been used to predict rehospitalizations in telemonitored chronic heart failure patients. Given that these data comprise time-to-event (time-to-rehospitalization) and longitudinal outcomes (the various longitudinal biomarkers), and given that interest is on predicting rehospitalization within a given time window of interest so that intervention can be done as appropriate, the practice of heart failure management through telemonitoring would immensely benefit from dynamic prediction as developed within the joint modelling framework. In this chapter, we aim to recast the framework in Rizopoulos (2012b, 2011) to propose a solution in this area. The first step is to fit a shared random effects joint model for the time-to-rehospitalization and the biomarker, with the latter as a time-varying and error-contaminated covariate (Rizopoulos, 2011; Verbeke *et al.*, 2010; Tsiatis and Davidian, 2004). Note that in this step, we take into account the fact that the biomarker may be measured with error, an aspect that is clearly ignored in the existing (non-statistical) approaches. The second step involves calculating (patient-specific) conditional survival probabilities, and their respective

confidence intervals, based on the fitted model (Rizopoulos, 2011). These probabilities, which take into account the patient's available biomarker measurements, and get (dynamically) updated as more measurements become available, would clearly provide physicians with an additional tool on which to base their intervention decisions. The third step is to quantify the discriminative ability of the biomarker (Rizopoulos, 2012b, 2011). This would aid physicians in assessing the performance of the predictions as provided by the statistical model.

4.2 The Joint Model

4.2.1 Specification, Assumptions, and Estimation

Models in which a sub-model for the time-to-event outcome is linked to that for the longitudinal outcome through a shared latent structure have received considerable focus. Distributional assumptions may be placed on the latent structure, or the same may be relaxed. Tsiatis and Davidian (2004), for instance, review both the parametric and conditional score approach. Verbeke *et al.* (2010) and Rizopoulos (2011) fit shared random effects joint models, with normality assumptions on the random effects.

For the i^{th} subject, we let T_i be the observed event time, $T_i = \min(T_i^*, C_i)$, with T_i^* the true event time, and C_i the censoring time. We assume the following hazard model:

$$h_i(t|\mathcal{M}_i(t), \boldsymbol{w}_i) = h_0(t)\exp\{\boldsymbol{\gamma}'\boldsymbol{w}_i + \alpha m_i(t)\}, \tag{4.1}$$

where $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ is the history of the true, unobserved longitudinal process up to time point t, α the parameter representing the effect of the longitudinal process on the hazard, and $h_0(\cdot)$ the baseline risk function. In addition, $\boldsymbol{w}_i$ is a baseline covariates' vector with associated parameter vector $\boldsymbol{\gamma}$. The hazard for an event at time t therefore depends on the baseline hazard, baseline covariates, and the true value of the longitudinal covariate at that time. The risk ratio associated with unit changes in the baseline covariates is given by $\exp(\boldsymbol{\gamma})$, while the relative change in the risk for a unit change in the true value of the longitudinal covariate is $\exp(\alpha)$.

For the longitudinal outcome, we assume that the unobserved true value for the i^{th} subject at any time point t is related with the observed value $y_i(t)$ via the model:

$$y_i(t) = m_i(t) + \varepsilon_i(t) = \boldsymbol{x}_i'(t)\boldsymbol{\beta} + \boldsymbol{z}_i'(t)\boldsymbol{b}_i + \varepsilon_i(t), \tag{4.2}$$

with $\varepsilon_i(t) \sim N(0, \sigma^2)$, $\boldsymbol{\beta}$ the parameter vector, $\boldsymbol{x}_i'$ and $\boldsymbol{z}_i'$ the vectors of the fixed effects and random effects design matrices, respectively. The measurement error $\varepsilon_i(t)$ is assumed independent of the random effects $\boldsymbol{b}_i$, with $\boldsymbol{b}_i \sim N(0, D)$. The longitudinal

outcome at time t therefore comprises of the true value, $m_i(t)$, contaminated by a random error term, $\varepsilon_i(t)$. The true value, as shown, is represented by a mixed model.

Note that the formulation so far assumes that the longitudinal outcome is observed at any time t. This is normally not the case, since measurements are only observed intermittently, at the time points t_{ij}. Therefore, as discussed in Verbeke *et al.* (2010), the aim is to estimate $m_i(t)$ using the available measurements, $y_i(t_{ij}), j = 1, \ldots, n_i$, combined with model (4.2).

As discussed by the above-mentioned authors, the likelihood contribution for the i^{th} patient is:

$$f(T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta}) = \int f(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta}) \left[\prod_j f\{\boldsymbol{y}_i(t_{ij}) | \boldsymbol{b}_i; \boldsymbol{\theta}\} \right] f(\boldsymbol{b}_i; \boldsymbol{\theta})\, d\boldsymbol{b}_i,$$

with $\boldsymbol{\theta}$ the parameter vector, $\boldsymbol{y}_i$ the longitudinal information for the i^{th} subject, δ_i the event indicator, and

$$\begin{aligned} f(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta}) &= [h_0(T_i) \exp\{\boldsymbol{\gamma}' \boldsymbol{w}_i + \alpha m_i(T_i)\}]^{\delta_i} \\ &\quad \times \exp\left(- \int_0^{T_i} [h_0(t) \exp\{\boldsymbol{\gamma}' \boldsymbol{w}_i + \alpha m_i(t)\}]\, dt \right). \end{aligned}$$

Note that the definition evokes the assumption of non-informativeness of the censoring mechanism and the so-called visiting process. Tsiatis and Davidian (2004) offer an excellent overview on the usually made assumptions.

We mention that different parameterizations can be considered for the true marker, $m_i(t)$, in (4.1). Instead of the true value, the true trajectory, as well as both the true value and the true trajectory, can easily be considered. In addition, in case of highly non-linear longitudinal profiles, splines, or higher order polynomials, could be considered (Rizopoulos, 2012a). As such, the model allows for a lot of flexibility.

4.2.2 Predicted Conditional Survival Probabilities

Based on a fitted joint model, interest here lies in calculating survival probabilities for a new subject who has provided a set of longitudinal measurements $\mathcal{Y}_i(t) = \{y_i(s); 0 \leq s \leq t\}$. Rizopoulos (2011) takes into consideration the fact that a subject providing longitudinal measurements up to time t implies they have survived up to that point; in other words, the longitudinal process provides a time-dependent endogenous covariate. Therefore, it is of relevance to consider the conditional probability that the subject survives time $u > t$, given survival up to t:

$$\pi_i(u|t) = \Pr\{T_i^* \geq u | T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_n; \boldsymbol{\theta}\}, \tag{4.3}$$

with $\mathcal{D}_n = \{T_i, \delta_i, y_i; i = 1, \ldots, n\}$ being the sample on which the model was fitted, and on which we wish to base our predictions (Rizopoulos, 2011). The problem is written in a Bayesian formulation to facilitate the calculation of standard errors. In particular, the subject-specific survival probabilities (4.3) are decomposed into three factors. The first is the posterior of the parameters $\boldsymbol{\theta}$ given the data, which is approximated by a normal distribution with the MLEs and variance-covariance the asymptotic covariance matrix of the MLEs. The second factor is the posterior of the random effects for subject i given his oberved data $\{\mathcal{Y}_i(t), T_i^* > t\}$. We sample from this distribution using a Metropolis-Hastings algorithms with multivariate t proposals. The last factor is the ratio of the conditional survival probabilties $S_i(u|\mathcal{M}_i(u))/S_i(t|\mathcal{M}_i(t))$, which is calculated using realization of the parameters $\boldsymbol{\theta}$ from the asymptotic normal distribution and of the random effects from factor 2.

Repeating these steps a desired number of times we obtain a Monte Carlo sample of $\pi_i(u|t)$ based on which can derive standard errors and calculate confidence intervals. For the interested reader, technical details are provided by Rizopoulos (2011).

4.2.3 Prospective Accuracy: Time-dependent AUCs and the Dynamic Discrimination Index

It is usually of interest to assess the predictive performance of a joint model. This is commonly done in the context of calibration, i.e., how well the model predicts what is observed, as well as in the setting of discrimination, i.e., how well the model creates a distinction between patients who experience the event of interest and those who do not. We focus on the latter approach, and use ROC methodology.

Given that conditional survival probabilities are progressively updated as more measurements become available, of medical interest is then to distinguish between patients who are going to experience the event within a given time window from those who are not. This can then aid making appropriate intervention decisions. Formally, given longitudinal measurements $\mathcal{Y}_i(t)$, interest is on the time window $(t, t + \Delta t]$. A prediction rule is then defined using $\pi_i(t + \Delta t|t)$, where, for c in $[0, 1]$, $\pi_i(t + \Delta t|t) \leq c$ is termed a success (occurrence of the event), and $\pi_i(t + \Delta t|t) > c$ a failure. Sensitivity and specificity are thus defined as

$$\begin{aligned} Se(c, t) &= \Pr\{\pi_i(t + \Delta t|t) \leq c | T_i^* \in (t + \Delta t]\}, \\ Sp(c, t) &= \Pr\{\pi_i(t + \Delta t|t) > c | T_i^* > t + \Delta t\}. \end{aligned}$$

The AUC at time t, $AUC(t)$, is obtained by varying c, as

$$AUC(t, \Delta t) = \Pr[\pi_i(t + \Delta t|t) < \pi_j(t + \Delta t|t) | \{T_i^* \in (t, t + \Delta t]\} \cap \{T_j^* > t + \Delta t\}],$$

(Rizopoulos, 2012b), where i and j represent a pair of comparable subjects (comparable pairs) (Antolini *et al.*, 2005; Harrell *et al.*, 1996). Here, the logic is that at each time point, and for a given time period of interest, if we consider two subjects, one of whom experiences the event within the time window, and the other who survives the time window, the calculated conditional survival probability for the first patient should be lower. Predictive accuracy can be assessed at certain time points and for given time windows, using the time-dependent AUC, and the overall performance can be assessed using a summary of the AUCs, in the form of the dynamic discrimination index, $C_{dyn}^{\Delta t}$. Rizopoulos (2012b) uses the following:

$$C_{dyn}^{\Delta t} = \frac{\int AUC(t, \Delta t) \Pr\{\varepsilon(t, \Delta t)\}\, dt}{\int \Pr\{\varepsilon(t, \Delta t)\}\, dt}, \tag{4.4}$$

where

$$\varepsilon(t, \Delta t) = [\{T_i^* \in (t, t + \Delta t]\} \cap \{T_j^* > t + \Delta t\}].$$

$\Pr\{\varepsilon(t, \Delta t)\}$ is the probability that a random pair of subjects is comparable at time t. Technical details regarding the estimation of these quantities are provided in Rizopoulos (2011). It should be mentioned that as with Harrell's C-index, the dynamic discrimination index (DDI) does not fully account for censoring; nonetheless, just as the Harrell's C-index has found routine use, we believe that this DDI has a place in practice.

4.3 Analysis of the Chronic Heart Failure Data

We now consider the data introduced in Section 2.1. Our analysis of these data will have two objectives. One, we will explore prediction of patient-specific survival probabilities, given the available biomarker measurements, and assess how well the biomarkers discriminate between patients who are and patients who are not going to get hospitalized. Secondly, we will evaluate the added value of correcting for each of the baseline covariates on the discriminative ability of the biomarkers. Four patients with missing baseline cardiac muscle fiber stretch (NTproBNP) are not included in the analysis. In Figure 4.1, we present the Kaplan-Meier survival estimate, with the 95% confidence interval, for the time to first hospitalization.

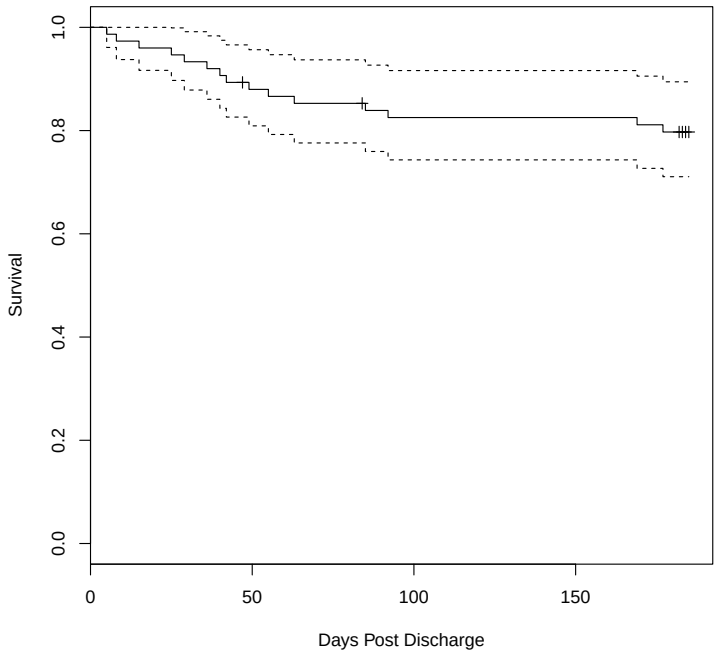


Figure 4.1: *Kaplan-Meier survival estimate, for time to first hospitalization.*

4.3.1 Model Formulation

For the time to first hospitalization and for each of the longitudinal markers separately, the joint model consisting of (4.1) and (4.2) was considered. At a first step, baseline covariates were not included into the joint model. In this case, the linear predictor of the survival sub-model only contained the effect of the biomarker, while the fixed-effects structure of the longitudinal sub-model only included the linear time evolution. We will forthwith refer to this as the first step model. At a second step, each of the baseline covariates were considered in turn. In this case, the concerned baseline covariate was not only included in the survival but also in the longitudinal sub-model. This will be referred to as the second step model. The random-effects structure in longitudinal sub-model, and hence the shared random-effects structure, comprised of a random intercept. The Weibull baseline risk function was assumed for $h_0(t)$. For convenience, (4.1) was re-parameterized as:

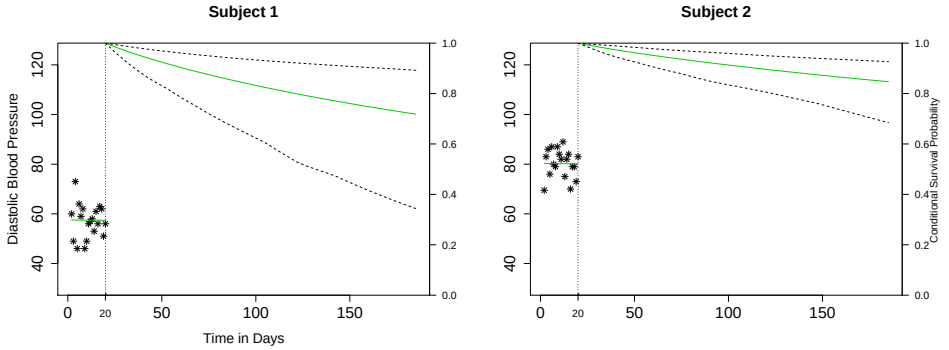
$$h_i(t|\mathcal{M}_i(t), w_i) = \rho\, t^{\rho-1} \exp\{\gamma_0 + \boldsymbol{\gamma}' \boldsymbol{w}_i + \alpha m_i(t)\}, \tag{4.5}$$

with ρ the shape parameter, and scale parameter $\exp(\gamma_0)$.

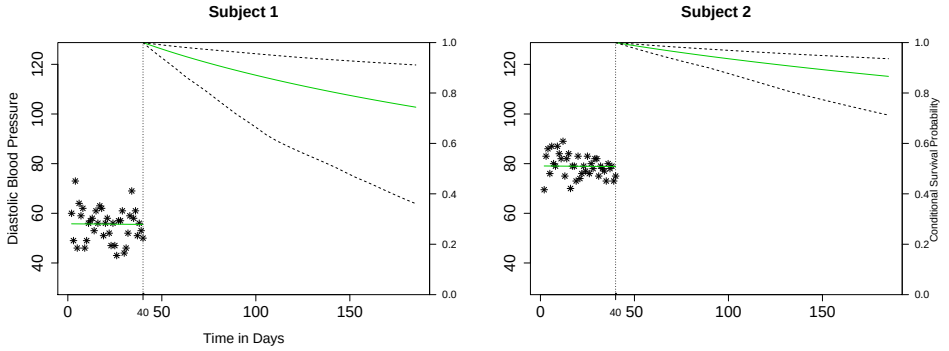
4.3.2 Diastolic Blood Pressure

Given the model fit, we now consider dynamic prediction of conditional survival probabilities. For illustrative purposes, we will consider two patients, who provided diastolic blood pressure measurements in a similar pattern for the first 100 days, where 'pattern' refers to missingness. For each of these patients, we consider their measurements during the first 20, 40, 80, and 100 days and, given that they had survived up to each of these time points, we compute the predicted conditional survival probabilities at each of the remaining time points until the study end. Two hundred Monte Carlo samples are generated and their median considered. We plot these probabilities, for both patients, in Figure 4.2 (considering measurements during the first 20 and 40 days), and Figure 4.3 (considering measurements during the first 60 and 80 days). The scatter points appearing before the vertical dashed line represent a plot of the longitudinal measurements up to that particular time point. To the right of this line, the conditional survival probability curve is shown, with the solid line representing the median over the Monte Carlo replications, and the dashed curves representing the confidence intervals. We notice how the diastolic blood pressure profile of the patient reflects in the conditional survival probability profile. For the first patient, the diastolic blood pressure measurements show a general clustering below 60, in comparison to the second patient, where the same seems to generally remain between 60 and 90. We notice that the conditional survival probability profile for the first patient declines rapidly, in comparison to the one for the second patient. This seems logical given that diastolic blood pressure measurements below 60 may be indicative of hypotension, hence it may be expected that the first patient is less likely to survive later times, in comparison to the second.

To show more clearly how the conditional survival probabilities are updated as more measurements become available, Figure 4.4 shows, for the same couple of patients as above, how the conditional probabilities of surviving an extra 20, 40, 60, and 80 days are updated, with each additional 20 days of measurements, starting from day 20 to the 100^{th} day. The same number of Monte Carlo samples as above is considered. The large dots represent the estimates, as given by the median, and the lines indicate the confidence intervals. We now consider how well the model performs in terms of discriminating between subjects who are going to experience hospitalization, and those who are not. We pre-specify that AUCs will be calculated every 2 weeks (14 days), and the time windows of interest will be 2, 4, 8 and, 16 days. Hence, DDIs will be calculated for the windows of interest Δt=2, 4, 8 and 16 days. In Table 4.1, we present the results. From the AUCs, we notice varying degrees of discriminative



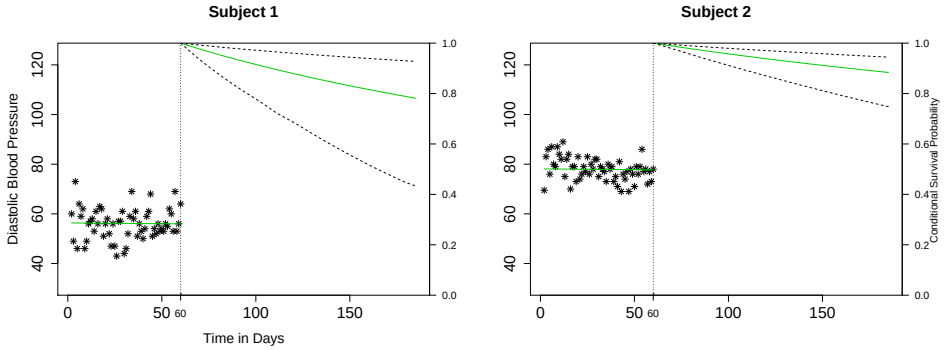
(a) Considering measurements during the first 20 days



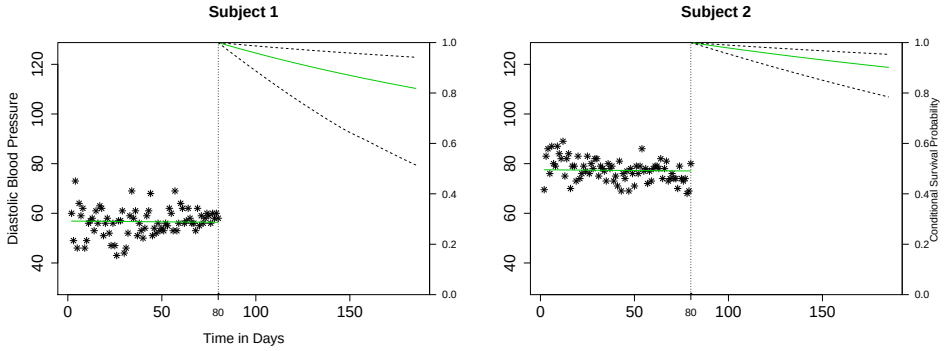
(b) Considering measurements during the first 40 days

Figure 4.2: *Conditional survival probabilities at each of the remaining time points until study end.*

ability for different time windows at different time points, from a high of 0.9552 for Δt=2 at 6 weeks (42 days), to a low of 0.0517 for Δt=16 at day 154. This means, for instance, that if interest was on predicting rehospitalization within a 2-day window, and such a prediction was done at 6 weeks, then the probabilty that the model would allocate a lower conditional survival probability to a patient who was going to be rehospitalized within the next 2 days as compared to one that was not, would be 0.9552. To get a summarized measure of the discriminative ability over the follow-up period, we look at the DDIs, which provide a weighted average of the AUCs, with weights accounting for how many patients are still at risk. The indices range from 0.4875 for a time window of 2 days to 0.5814 for a time window of 8 days.



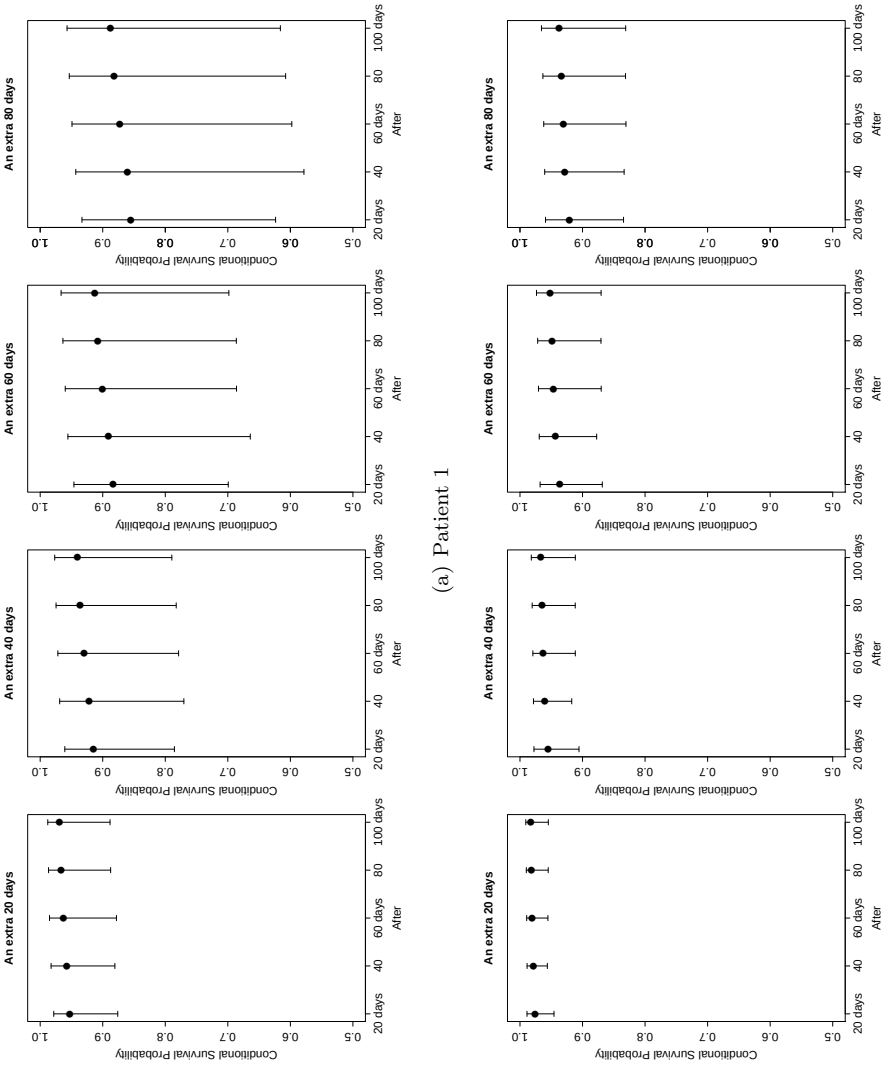
(a) Considering measurements during the first 60 days



(b) Considering measurements during the first 80 days

Figure 4.3: *Conditional survival probabilities at each of the remaining time points until study end.*

Second step Here, interest will only be on assessing the added overall discriminative value of correcting for each of the baseline covariates. In Table 4.2, we provide DDIs, calculated under the same time-point and time-window specifications as above. We also include DDI results from the first step model as above. We note that correction for NTproBNP has the highest positive effect on the overall discriminative ability under all the time windows, with the index improving to above 0.7 for an 8-day window. Patient age has the second-highest positive effect, with the index reaching above 0.6 for 8-day and 16-day windows.



(a) Patient 1

(b) Patient 2

Figure 4.4: *Conditional probabilities of surviving an extra 20, 40, 60 and 80 days, with each additional 20 days of measurement.*

Table 4.1: *Diastolic Blood Pressure, First Step Model. AUCs and DDIs.*

Time window Δt	Time point t	AUC(t)	DDI
2	14	0.6944	0.4875
	28	0.7429	
	42	0.9552	
	84	0.3770	
	168	0.0862	
4	14	0.6944	0.4949
	28	0.7714	
	42	0.8955	
	84	0.3770	
	168	0.0862	
8	14	0.7500	0.5814
	28	0.8696	
	42	0.4615	
	56	0.7619	
	84	0.3934	
	168	0.0862	
16	14	0.7810	0.5745
	28	0.6493	
	42	0.4688	
	56	0.7619	
	70	0.4590	
	84	0.7119	
	154	0.0517	
	168	0.2692	

Table 4.2: *Diastolic Blood Pressure, DDIs for different time windows.*

		Second Step. Covariate controlled for:					
Δt	First Step	NTproBNP	Heart Rhythm	NYHA	Sex	LVEF	Age
2	0.4875	0.6054	0.4842	0.5087	0.5156	0.4725	0.5425
4	0.4949	0.6167	0.4956	0.5274	0.5267	0.4838	0.5535
8	0.5814	0.7206	0.5819	0.6126	0.5909	0.5702	0.6140
16	0.5745	0.6260	0.5874	0.5933	0.5967	0.5864	0.6254

4.3.3 Systolic Blood Pressure, Heart Rate, and Weight

In the foregoing elaborate analysis and discussion on diastolic blood pressure, we have looked at the dynamic prediction of conditional survival probabilities, how these probabilities get updated as more measurements become available, seen how the longitudinal profiles reflect in these probabilities, and assessed the discriminative ability under the first step and the second step model. Clearly, such a detailed procedure can be repeated for each of the three remaining markers (systolic blood pressure, heart rate and weight). For each of these three, however, we will focus on assessing the discriminative ability. We will consider the DDIs computed following the same time-point and time-window specifications as above.

Systolic Blood Pressure For numerical stability, various variables were transformed or rescaled. In the first step model, systolic blood pressure measurements were rescaled to within unit magnitude, by dividing with the largest value. This was also done in the second step model that controlled for each of the following: NTproBNP, NYHA, sex, LVEF and age, with, in addition, NTproBNP values being transformed through the square root, and age values being rescaled by dividing with the minimum, in the respective models. In the second step model controlling for heart rhythm, however, systolic blood pressure was transformed through the cube root. In Appendix B.1, where we present the results, both for the first and the second step, we notice discriminative power of at least 0.6 for all of the time windows in the first step model. Patient age is seen to provide the highest positive impact on discriminative power.

Heart Rate Heart rate values were rescaled to within unit magnitude, by dividing with the largest value, for numerical stability. In Table 4.3, DDIs for the various time windows are given. The discriminative ability in the first step model was above 0.65. Controlling for NTproBNP provided the highest positive enhancement on the discriminative ability for the time windows of 2, 4 and 16, while for the 8-day window it was NYHA.

Weight Weight values were rescaled in a similar manner as above. For the second step model containing age, age values were rescaled as in the systolic blood pressure case. As shown in Appendix B.2, the first step model showed discriminative indices

Table 4.3: *Heart Rate. DDIs.*

	Time window Δt	DDI
First Step Model	2	0.6698
	4	0.6698
	8	0.7433
	16	0.6576
Second Step Model		
NTproBNP	2	0.6983
	4	0.7061
	8	0.7524
	16	0.6706
Heart Rhythm	2	0.6700
	4	0.6700
	8	0.7434
	16	0.6532
NYHA	2	0.6801
	4	0.6842
	8	0.7554
	16	0.6670
Sex	2	0.6837
	4	0.6917
	8	0.7459
	16	0.6671
LVEF	2	0.6622
	4	0.6698
	8	0.7452
	16	0.6585
Age	2	0.6644
	4	0.6685
	8	0.7341
	16	0.6459

of between 0.3877 and 0.5020, while patient age had the highest positive effect on the discriminative ability.

4.3.4 Overall Findings

Diastolic blood pressure measurements on their own showed generally poor discriminative ability, with patient's cardiac muscle fiber stretch and patient age being seen to provide, in that order, the best enhancement to this ability. Systolic blood pressure measurements had fairly moderate discriminative ability, and controlling for patient age provided the best impact. Heart rate measurements showed moderate to good discriminative power, with the DDI in the first step model for an 8-day time window being above 0.7. In this case, controlling for patient's cardiac muscle fiber stretch had the best impact. Patient's age provided the best enhancement to the discriminative ability of the longitudinal weight measurements. Finally, in our particular case of sensitivity analysis, consideration of a more elaborate random effects structure comprising of both random intercepts and slopes, as opposed to intercepts alone, was seen to produce noticeable improvement in discriminative ability.

4.4 Sensitivity Analysis

Considerable flexibility exists when specifying the joint model. The hazard for an event can be taken to depend on the true value of the biomarker, its true trajectory, or both, or even on its previous values. The true biomarker can be estimated using elaborate fixed and random effects structures, through higher order polynomials or splines. Finally, piecewise-constant and spline representations can be used for the survival baseline, in addition to the common parametric choices.

We briefly explore a number of these options, in terms of the discriminative ability of the model, and its fit. For our illustration, we consider the first step model for heart rate, and focus on the baseline hazard function, the lag of the true value of the longitudinal covariate, and the random effects structure. The linear predictor in the survival sub-model, (4.1), was expressed as

$$\gamma' w_i + \alpha m_i\{\max(t-k,0)\},$$

where k represents the lag. Lag values of zero and one, respectively denoting the assumption that the hazard depended on the true current value or the true previous value of the longitudinal covariate, were considered. Two forms for the baseline hazard function were considered: the Weibull and the piecewise constant baseline risk function, with the knots in the latter being equally-spaced in the percentiles of the observed event times. Both random intercepts only as well as random intercepts and slopes were considered. In Table 4.4, DDIs under the various combinations of

assumptions are provided. In this particular case, while for a given random effects structure we do not notice substantial impact as assumptions regarding the lag of the true value of the longitudinal covariate and the baseline hazard function are varied, there is noticeable improvement in discriminative ability when the random effects structure is extended to include random slopes. Therefore, in this particular case, consideration of a more elaborate random effects structure comprising of both random intercepts and slopes, as opposed to intercepts alone, produces noticeable improvement in discriminative ability.

Table 4.4: *Heart Rate. DDI under various assumptions.*

Random effects	Lag	Baseline hazard	Δt	DDI	AIC
Intercept	0	Weibull	2	0.6698	-26413.15
			4	0.6698	
			8	0.7433	
			16	0.6576	
		Piecewise Constant	2	0.6698	-26408.43
			4	0.6698	
			8	0.7433	
			16	0.6595	
	1	Weibull	2	0.6698	-26413.14
			4	0.6698	
			8	0.7433	
			16	0.6576	
		Piecewise Constant	2	0.6698	-26408.42
			4	0.6698	
			8	0.7433	
			16	0.6595	
Intercept and slope	0	Weibull	2	0.7427	-27995.78
			4	0.7508	
			8	0.7902	
			16	0.6878	
		Piecewise Constant	2	0.7427	-27991.15
			4	0.7508	
			8	0.7902	
			16	0.6865	
	1	Weibull	2	0.7427	-27995.72
			4	0.7508	
			8	0.7941	
			16	0.6884	
		Piecewise Constant	2	0.7427	-27991.09
			4	0.7508	
			8	0.7941	
			16	0.6865	

The Akaike Information Criterion (AIC) values for the various models are also given, for assessment of model fit. In line with model selection practices, this can be used, alongside the DDI, to settle on a model. For the set of models considered, the AIC is the smallest for the Weibull baseline hazard model incorporating both random intercepts and slopes, and with dependence of the hazard on the current value of the

true longitudinal marker.

4.5 Discussion

In this chapter, we have explored how dynamic prediction can assist physicians make intervention decisions in telemonitored CHF patients. The dynamically updated conditional survival probabilities, and their confidence intervals, can provide physicians with additional information on which to base such decisions. We have also explored how well each of the available biomarkers discriminate between patients who are and those who are not going to get rehospitalized. This approach does not only provide a sound statistical modelling approach to predicting rehospitalizations in telemonitored chronic heart failure patients, it also provides a practical solution in heart failure management.

We have only addressed dynamic prediction in relation to the time to first hospitalization. There is therefore need for methodological extension to cope with dynamic prediction in the context of recurrent events . We have also analyzed each biomarker separately. It would be important to consider them jointly, taking their association structure into account. More software implementation work is needed to allow dynamic prediction and calculation of accuracy measures when multiple longitudinal biomarkers are available. The main challenge in this setting is that when marker-specific random effects are incorporated, computation will become prohibitive as the number of random effects increases.

Chapter 5

A Characterization of Missingness at Random in a Generalized Shared-parameter Joint Modelling Framework for Longitudinal and Time-to-Event Data, and Sensitivity Analysis

In this chapter, we consider a conceptual correspondence between the missing data setting, and joint modeling of longitudinal and time-to-event outcomes. Based on this, we formulate an extended shared random effects joint model, and provide a characterization of missing at random, coherent with that in the missing data setting. The additional complexity in the joint longitudinal and time-to-event setting is highlighted, and a sensitivity analysis is presented within the extended random effects framework. This is illustrated using data from a study on liver cirrhosis.

5.1 Introduction

In the missing data setting, three main classes of models have been developed: the so-called selection (SEM), pattern-mixture (PMM), and shared-parameter (SPM) frameworks; see for example Molenberghs and Kenward (2007). The SEM and PMM approaches are defined by different factorizations of the joint distribution of the data and the missing value processes. In the SPM, on the other hand, both data and the missing value process are assumed to depend on latent variables, conditional upon which independence is assumed.

Originally due to Rubin (1976), a classification of missing value processes within the SEM has been developed. In a frequentist paradigm, with outcomes only missing, this classification can be expressed as follows: Missing completely at random (MCAR), implying that, conditional upon covariates, the missing value mechanism does not depend on outcomes; missing at random (MAR), implying that conditional on covariates and observed outcomes, the mechanism does not further depend on missing outcomes; and, finally, when MCAR and MAR do not hold, we have a missing not at random (MNAR) process, in which, conditional on covariates and observed outcomes, the missing value mechanism does depend on unobserved outcomes. The taxonomy has also been transposed to the PMM (Molenberghs *et al.*, 1998) and SPM (Creemers *et al.*, 2011) frameworks.

Given that models for missing data often make unverifiable assumptions about the missing value mechanism, a recurring theme is that of sensitivity analysis (Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005). As assumptions regarding the missing value mechanism are varied, the stability of inferences, or lack thereof, provides a guide on the caution with which the inferences need to be embraced. Though sensitivity analysis has primarily been done under the SEM and PMM frameworks, Creemers *et al.* (2010) considered it in the context the SPM.

The joint longitudinal and time-to-event setting is slightly different, given that a time-to-event is also collected. An example is in prostate cancer, where, after treatment for cancer, prostate-specific antigen measurements are collected over time, alongside the time-to-disease-recurrence (Law *et al.*, 2002; Yu *et al.*, 2004, 2008). In HIV/AIDS studies, apart from the time to onset of AIDS or death, viral load and CD4 cell count may be recorded repeatedly over time (DeGruttola and Tu, 1994; Rizopoulos, 2011). The objectives may be three-fold: first, to study the survival outcome, accounting for the longitudinal covariate; second, to study the longitudinal outcome, accounting for possibly non-random drop-out caused by the occurrence of events; and third, to examine the association structure between the two outcomes

(Tsiatis and Davidian, 2004; Rizopoulos *et al.*, 2009; Verbeke *et al.*, 2010; Rizopoulos, 2012a).

Among the three objectives, the first one is arguably the most common. This objective is usually achieved within the SPM framework. A sub-model for the time-to-event is linked to one for the longitudinal process using a shared latent structure, say a normal random effect, conditional on which independence is assumed (Tsiatis and Davidian, 2004; Verbeke *et al.*, 2010; Rizopoulos, 2011, 2012a). Like in the missing data setting, there are challenges. The longitudinal covariate may be measured with error, its values are likewise only available at the specific time points at that the patient appears at the clinic for longitudinal measurements, and the time-to-event may also be censored (Tsiatis and Davidian, 2004).

As such, though interest is on the time-to-event accounting for the "true" longitudinal process, the joint density incorporates not only the censoring but also the visiting and measurement probabilities (Tsiatis and Davidian, 2004). The visiting probabilities represent the process which generates the time points at which measurements are available (Rizopoulos, 2012a). To identify the relationship of interest, under likelihood inference, it is assumed that the probabilities of censoring and visiting can depend on past visit times and longitudinal measurements, but not further on the future longitudinal measurements and event time (Tsiatis and Davidian, 2004; Verbeke *et al.*, 2010; Rizopoulos, 2012a). These conditions mirror the MAR assumption mentioned earlier. The assumptions are again unverifiable based on the data, raising sensitivity issues. Rizopoulos (2012a) for instance considers the second objective mentioned above, and highlights the consideration of different parameterizations of the longitudinal process in the survival sub-model, as a possible route for sensitivity analysis.

Undeniably, there is a strong connection between the missing data and the joint longitudinal and time-to-event settings. In this chapter, we take a slightly different perspective on joint models than is prevalent in the literature, and argue that conceptually, the two settings actually correspond. Based on this, we build an extended shared random effects joint model, similar in spirit to that of Creemers *et al.* (2011) in the context of longitudinal data subject to missing observations, but now transposed to the current more complex setting. The added layer of complexity stems from the fact that data can now be coarsened in various ways: the longitudinal sequence can be incomplete; the time-to-event outcome can be censored; both of these can occur simultaneously. Coarsening refers to the phenomenon that data observed are less refined than the, possibly counterfactual, full data.

Within the extended framework, we provide a characterization of MAR, consistent

to the one in the missing data setting. The complexity of model formulation in the extended setting is then considered. The extended random-effects structure is then utilized for sensitivity analysis.

The organization of the chapter is as follows. In the following section, we briefly review missing data concepts, the various modeling frameworks, and the characterization of MAR in each of these frameworks. We review the generalized shared-parameter modeling (GSPM) framework of Creemers *et al.* (2011), and its MAR characterization. In Section 5.3, we introduce the problem of joint modeling of longitudinal and time-to-event data and illustrate the correspondence between this problem and that of missing data. Based on this, we introduce our extended framework, and derive its MAR characterization. In Section 5.4, we reflect on the complexity of model formulation in the current setting. We consider a sensitivity analysis in Section 5.5, with an illustrative application, and finally provide some closing remarks in Section 5.6.

5.2 Background on Missing Data

5.2.1 Notation

Let Y_{ij} denote the outcome for the i^{th} subject measured at the j^{th} occasion, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. Additionally, define a missingness indicator, R_{ij}, which takes the value 1 if Y_{ij} is observed, and 0 otherwise. We then have $\boldsymbol{Y_i}$ and $\boldsymbol{R}_i$, representing the measurement and missingness process vectors for subject i, respectively. Further partition $\boldsymbol{Y_i}$ into its observed and unobserved components $\boldsymbol{Y_i}^o$ and $\boldsymbol{Y_i}^m$, respectively. We also define $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ to be the parameter vectors for the measurement and missingness processes, respectively. We suppress the covariate vector $\boldsymbol{x}_i$ in the notation. Hence, we write $f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi})$ for the full data density.

5.2.2 Modeling Frameworks

The *selection model* (SEM) starts from the factorization $f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{\theta})f(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{\psi})$, in contrast to the PMM representation which is based upon $f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{r}_i, \boldsymbol{\theta})f(\boldsymbol{r}_i|\boldsymbol{\psi})$. The conventional SPM incorporates a vector of shared latent variables $\boldsymbol{b_i}$, conditional upon which independence of the measurement and missingness processes is assumed, i.e.,

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = \int f(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})f(\boldsymbol{r}_i|\boldsymbol{b}_i, \boldsymbol{\psi})f(\boldsymbol{b_i})\, d\boldsymbol{b_i}. \tag{5.1}$$

5.2.3 Characterization of Missing at Random

First, we consider MAR in the various frameworks. By definition, under the SEM framework, missingness is MAR if $f(\boldsymbol{r}_i|\boldsymbol{y}_i,\boldsymbol{\psi}) = f(\boldsymbol{r}_i|\boldsymbol{y}_i^o,\boldsymbol{\psi})$. Under the PMM framework, the missing-data mechanism is MAR if

$$f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i,\boldsymbol{\theta}) = f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{\theta}) \tag{5.2}$$

(Molenberghs *et al.*, 1998). In (5.2), the so-called predictive distribution of the unobserved outcomes, given the observed ones, is made explicit. Thus, in the PMM setting, MAR can be seen to mean that the unobserved outcomes can be predicted from the observed outcomes and covariates, without further reference to the missingness mechanism.

Under conventional SPM in (5.1), MAR cannot hold without reducing to MCAR, in which case $\boldsymbol{b_i}$ drops from at least one of the factors in the integrand of (5.1). Creemers *et al.* (2011) generalized the SPM (GSPM) by expanding the random-effects structure:

$$\begin{aligned} & f(\boldsymbol{y}_i,\boldsymbol{r}_i|\boldsymbol{g}_i,\boldsymbol{h}_i,\boldsymbol{j}_i,\boldsymbol{k}_i,\boldsymbol{\ell}_i,\boldsymbol{m}_i,\boldsymbol{q}_i) \\ &\quad = \quad f(\boldsymbol{y}_i^o|\boldsymbol{g}_i,\boldsymbol{h}_i,\boldsymbol{j}_i,\boldsymbol{\ell}_i) f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{g}_i,\boldsymbol{h}_i,\boldsymbol{k}_i,\boldsymbol{m}_i) f(\boldsymbol{r}_i|\boldsymbol{g}_i,\boldsymbol{j}_i,\boldsymbol{k}_i,\boldsymbol{q}_i). \end{aligned} \tag{5.3}$$

The random-effects structure is more general; the random effects $\boldsymbol{g}_i$ are shared among all processes, $\boldsymbol{h}_i$, $\boldsymbol{j}_i$, and $\boldsymbol{k}_i$ are shared between two processes only, while $\boldsymbol{\ell}_i$, $\boldsymbol{m}_i$, and $\boldsymbol{q}_i$ are specific to one. Using this general formulation, the above-mentioned authors established that GSPM (5.3) is MAR if and only if

$$\begin{aligned} & \frac{\int f(\boldsymbol{y}_i^o|\boldsymbol{g}_i,\boldsymbol{h}_i,\boldsymbol{j}_i) f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{g}_i,\boldsymbol{h}_i,\boldsymbol{k}_i) f(\boldsymbol{r}_i|\boldsymbol{g}_i,\boldsymbol{j}_i,\boldsymbol{k}_i) f(\boldsymbol{b_i})\, d\boldsymbol{b_i}}{\int f(\boldsymbol{y}_i^o|\boldsymbol{g}_i,\boldsymbol{j}_i) f(\boldsymbol{r}_i|\boldsymbol{g}_i,\boldsymbol{j}_i) f(\boldsymbol{b_i})\, d\boldsymbol{b_i}} \\ & \qquad = \quad \frac{\int f(\boldsymbol{y}_i^o|\boldsymbol{g}_i,\boldsymbol{h}_i) f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{g}_i,\boldsymbol{h}_i) f(\boldsymbol{b_i})\, d\boldsymbol{b_i}}{f(\boldsymbol{y}_i^o)}. \end{aligned} \tag{5.4}$$

A convenient proper sub-class of GSPM (5.3) that satisfies MAR:

$$\begin{aligned} f(\boldsymbol{y}_i,\boldsymbol{r}_i|\boldsymbol{g}_i,\boldsymbol{h}_i,\boldsymbol{j}_i,\boldsymbol{k}_i,\boldsymbol{\ell}_i,\boldsymbol{m}_i,\boldsymbol{q}_i) &= f(\boldsymbol{y}_i,\boldsymbol{r}_i|\boldsymbol{j}_i,\boldsymbol{\ell}_i,\boldsymbol{m}_i,\boldsymbol{q}_i) \\ &= f(\boldsymbol{y}_i^o|\boldsymbol{j}_i,\boldsymbol{\ell}_i) f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{m}_i) f(r_i|\boldsymbol{j}_i,\boldsymbol{q}_i). \end{aligned} \tag{5.5}$$

5.3 Joint Modeling of Longitudinal and Time-to-event Data

We consider now the setting in which it is intended to observe both a longitudinal and time-to-event outcome between the start and the planned end of a study. We

examine three scenarios to illustrate the correspondence between this setting and that of missing data.

Scenario 1. For subjects who drop out before the planned end of the study, we observe longitudinal information prior to drop-out, as well as the censoring time. Consequently, for these subjects, we observe neither the latter part of the longitudinal sequence nor the survival time.

Scenario 2. For subjects who experience the event within the study period, such that the event censors the longitudinal sequence, we observe longitudinal information prior to the event, as well as the survival time. For these subjects, longitudinal data after the event, as well as the censoring time, are unobserved. The latter means, of course, that no censoring occurs.

Scenario 3. Finally, for subjects who reach the end of the study without experiencing the event, full longitudinal information as well as the censoring time are observed. For these subjects, the survival time remains unobserved.

From these three scenarios, we note that this setting always entails a part of the data being observed, and a part unobserved. We must also consider the mechanism that causes the coarsening, which consists of the union of the missingness mechanism in the longitudinal outcome, and a certain *choice mechanism,* related to the time-to-event outcome, which determines whether either the event time or censoring time is observed. Note that further scenarios are possible, even though we will restrict attention to these three in the current chapter. For example, it is possible for the longitudinal process to terminate early, while the patient is still followed for the event.

5.3.1 Additional Notation

We extend the notation of Section 5.2.1 to let T_i and C_i denote the survival and censoring times, respectively. Additionally, let $D_i^o = \min(T_i, C_i)$, and $D_i^m = \max(T_i, C_i)$.

We also introduce a vector of missingness indicators, $\boldsymbol{R}_i^* = (\boldsymbol{R}_i', W_i)'$, where $W_i = 1$ if the survival time is observed and zero otherwise. The full set of stochastic components is then:

$$\boldsymbol{Q}_i = (\boldsymbol{Y_i}^{o\prime}, \boldsymbol{Y}_i^{m\prime}, D_i^o, D_i^m, \boldsymbol{R}_i^{*'})' = (\boldsymbol{Z}_i^{o\prime}, \boldsymbol{Z}_i^{m\prime}, \boldsymbol{R}_i^*)', \tag{5.6}$$

where $\boldsymbol{Z}_i^o = (\boldsymbol{Y_i}^{o\prime}, D_i^o)'$ and $\boldsymbol{Z}_i^m = (\boldsymbol{Y}_i^{m\prime}, D_i^m)'$. Hence, we can represent the information in a form that parallels that for incomplete longitudinal data, but with each of the three vectors combining both longitudinal and time-to-event information. The information in $\boldsymbol{Z}_i^o$ and $\boldsymbol{Z}_i^m$ depends on the particular scenario, which sets this notation apart from that found in the literature.

5.3.2 The Extended Framework

Define the following shared random-effects model for these data:

$$f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, d_i^o, d_i^m, \boldsymbol{r_i}^*|\boldsymbol{b_i}) = f(\boldsymbol{y}_i^o|\boldsymbol{b_i})f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{b_i})f(d_i^o|\boldsymbol{b_i})f(d_i^m|d_i^o, \boldsymbol{b_i})f(\boldsymbol{r_i}^*|\boldsymbol{b_i}). \quad (5.7)$$

Here, $\boldsymbol{b_i}$ encompasses an elaborate random effects structure, which contains the following 31 sets of random effects: 1 shared between all five components, 5 shared between four components, 10 shared between three components, another 10 shared between two components, and 5 specific to a single component. These random effects are assumed independent. It is unwieldy to spell them all out as was done in (5.3), but the nature of the decomposition is identical. As we see below, the advantage of such a framework is that appropriate subsets of random effects can be chosen so that MAR holds.

We make the following points. First, model (5.7) is the generic shared random-effects model for this setting under this factorization. Such a general structure implies, for instance, that at the time of drop-out, there are processes which may stop, while other processes may get modified. Second, specific sub-models may be considered that are deemed appropriate for the application at hand. This is important, because the full set of random effects will typically be too elaborate for practical purposes. Third, for every application, it is important to consider the implication of the corresponding simplification, especially in terms of the underlying coarsening mechanism. It is then important to carefully distinguish between the case where the coarsening does not depend on unobserved measurements, on the one hand, and the coarsening mechanism is MAR, on the other hand. Fourth, the extended model in (5.7) is obviously based on conditional independence assumptions: given the collection of random effects $\boldsymbol{b_i}$, the processes $\boldsymbol{y}_i$, $\boldsymbol{d}_i$, and $\boldsymbol{r_i}$ are independent of one another. Of course, if all 31 random effects would be present, there still would be a rich association structure present between the various outcomes, which may be simplified by omitting one or more of these components, as will be done to allow for MAR in the next section. Fifth, the model *a priori* allows for dependence between d_i^o and d_i^m, regardless of the random-effects structure adopted, stemming from the penultimate factor on the right hand side of (5.7). In other words, T_i and C_i would be allowed to depend on one another. In full generality, this would be problematic because Tsiatis (1975) has shown that there is no information available from the data on the joint distribution of (T_i, C_i). The simplest way out is to force independence, by writing the corresponding factor as $f(d_i^m|d_i^o, \boldsymbol{b_i}) = f(d_i^m|\boldsymbol{b_i})$, with perhaps even the random effect removed. However, it is not our purpose to provide a general framework of which every member is identifi-

able, but rather one that generates more identifiable special cases than conventionally considered, in particular the MAR version of the next section. In addition, classes of non-identifiable models can be considered as part of a sensitivity analysis.

5.3.3 Characterization of Missing at Random

Extended model (5.7) allows for a characterization of MAR, in the same spirit as (5.4). We can define MAR by either starting from a SEM-based or from a PMM-based factorization of the model. Under a SEM factorization, the requirement is:

$$f(\boldsymbol{r}_i^*|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, d_i^o, d_i^m) = f(\boldsymbol{r}_i^*|\boldsymbol{y}_i^o, d_i^o), \tag{5.8}$$

implying

$$\frac{f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, d_i^o, d_i^m, \boldsymbol{r}_i^*)}{f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, d_i^o, d_i^m)} = \frac{f(\boldsymbol{y}_i^o, d_i^o, \boldsymbol{r}_i^*)}{f(\boldsymbol{y}_i^o, d_i^o)}. \tag{5.9}$$

From a PMM perspective, the requirement is:

$$\frac{f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, d_i^o, d_i^m, \boldsymbol{r}_i^*)}{f(\boldsymbol{y}_i^o, d_i^o, \boldsymbol{r}_i^*)} = \frac{f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, d_i^o, d_i^m)}{f(\boldsymbol{y}_i^o, d_i^o)}. \tag{5.10}$$

Using the specific form of (5.7), MAR holds if and only if:

$$\begin{aligned}\frac{\int f(\boldsymbol{y}_i^o|\boldsymbol{b}_i) f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{b}_i) f(d_i^o|\boldsymbol{b}_i) f(d_i^m|d_i^o, \boldsymbol{b}_i) f(\boldsymbol{r}_i^*|\boldsymbol{b}_i) f(\boldsymbol{b}_i)\, d\boldsymbol{b_i}}{\int f(\boldsymbol{y}_i^o|\boldsymbol{b}_i) f(d_i^o|\boldsymbol{b}_i) f(\boldsymbol{r}_i^*|\boldsymbol{b}_i) f(\boldsymbol{b}_i)\, d\boldsymbol{b_i}} \\ = \frac{\int f(\boldsymbol{y}_i^o|\boldsymbol{b}_i) f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{b}_i) f(d_i^o|\boldsymbol{b}_i) f(d_i^m|d_i^o, \boldsymbol{b}_i) f(\boldsymbol{b}_i)\, d\boldsymbol{b_i}}{\int f(\boldsymbol{y}_i^o|\boldsymbol{b}_i) f(d_i^o|\boldsymbol{b}_i) f(\boldsymbol{b}_i)\, d\boldsymbol{b_i}}.\end{aligned} \tag{5.11}$$

Recall that $\boldsymbol{b}_i$ as used here generically refers to the set of random effects that apply to the factor concerned.

5.3.4 A Sub-class of the Generalized SPM

Consider the following sub-class of model (5.7):

$$\begin{aligned}&f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, d_i^o, d_i^m, \boldsymbol{r}_i^*|\boldsymbol{b_i}) = \\ &\quad f(\boldsymbol{y}_i^o|\boldsymbol{g}_i, \boldsymbol{h}_i, \boldsymbol{k}_i) f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{m}_i) f(d_i^o|\boldsymbol{g}_i, \boldsymbol{h}_i, \boldsymbol{\ell}_i) f(d_i^m|d_i^o, \boldsymbol{m}_i) \\ &\quad f(\boldsymbol{r}_i^*|\boldsymbol{g}_i, \boldsymbol{k}_i, \boldsymbol{\ell}_i),\end{aligned} \tag{5.12}$$

where $\boldsymbol{g}_i$, $\boldsymbol{h}_i$, $\boldsymbol{k}_i$, $\boldsymbol{\ell}_i$, and $\boldsymbol{m}_i$ are part of the 31 sets of random effects described earlier. Note that under this sub-class, the random effects driving the missing-data components ${\boldsymbol{y}_i}^m$ and d_i^m do not appear in any of the other three stochastic components.

Next, we show that this sub-class satisfies the MAR property. Let $\widetilde{\boldsymbol{b}}_i$ be shorthand for the set of random effects ($\boldsymbol{g}_i$, $\boldsymbol{h}_i$,$\boldsymbol{k}_i$, $\boldsymbol{\ell}_i$, and $\boldsymbol{m}_i$), and $\overline{\boldsymbol{b}}_i$ be shorthand for the same set but excluding $\boldsymbol{m}_i$. Then, from a SEM-based factorization,

$$\begin{aligned} f(\boldsymbol{r}_i^*|\boldsymbol{y}_i^o,\boldsymbol{y}_i^m,d_i^o,d_i^m) &= \frac{f(\boldsymbol{y}_i^o,\boldsymbol{y}_i^m,d_i^o,d_i^m,\boldsymbol{r}_i^*)}{f(\boldsymbol{y}_i^o,\boldsymbol{y}_i^m,d_i^o,d_i^m)} \\ &= \frac{\int_{\widetilde{\boldsymbol{b}}_i}\varphi_1\varphi_2\varphi_3\varphi_4\varphi_5 f(\widetilde{\boldsymbol{b}}_i)\,d\widetilde{\boldsymbol{b}}_i}{\int_{\boldsymbol{r}_i^*}\int_{\widetilde{\boldsymbol{b}}_i}\varphi_1\varphi_2\varphi_3\varphi_4\varphi_5 f(\widetilde{\boldsymbol{b}}_i)\,d\widetilde{\boldsymbol{b}}_i\,d\boldsymbol{r}_i^*} \\ &= \frac{\int_{\boldsymbol{m}_i}\varphi_2\varphi_4 f(\boldsymbol{m}_i)\,d\boldsymbol{m}_i}{\int_{\boldsymbol{m}_i}\varphi_2\varphi_4 f(\boldsymbol{m}_i)\,d\boldsymbol{m}_i} \\ &\quad\times\frac{\int_{\overline{\boldsymbol{b}}_i}\varphi_1\varphi_3\varphi_5 f(\overline{\boldsymbol{b}}_i)\,d\overline{\boldsymbol{b}}_i}{\int_{\boldsymbol{r}_i^*}\int_{\overline{\boldsymbol{b}}_i}\varphi_1\varphi_3\varphi_5 f(\overline{\boldsymbol{b}}_i)\,d\overline{\boldsymbol{b}}_i\,d\boldsymbol{r}_i^*} \\ &= \frac{f(\boldsymbol{y}_i^o,d_i^o,\boldsymbol{r}_i^*)}{f(\boldsymbol{y}_i^o,d_i^o)} = f(\boldsymbol{r}_i^*|\boldsymbol{y}_i^o,d_i^o), \end{aligned} \tag{5.13}$$

where

$$\begin{aligned} \varphi_1 &= f(\boldsymbol{y}_i^o|\boldsymbol{g}_i,\boldsymbol{h}_i,\boldsymbol{k}_i) \\ \varphi_2 &= f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{m}_i) \\ \varphi_3 &= f(d_i^o|\boldsymbol{g}_i,\boldsymbol{h}_i,\boldsymbol{\ell}_i) \\ \varphi_4 &= f(d_i^m|d_i^o,\boldsymbol{m}_i) \\ \varphi_5 &= f(\boldsymbol{r}_i^*|\boldsymbol{g}_i,\boldsymbol{k}_i,\boldsymbol{\ell}_i). \end{aligned} \tag{5.14}$$

Alternatively, one could start from a PMM-based factorization. In this case,

$$\begin{aligned} f(\boldsymbol{y}_i^m,d_i^m|\boldsymbol{y}_i^o,d_i^o,\boldsymbol{r}_i^*) &= \frac{f(\boldsymbol{y}_i^o,\boldsymbol{y}_i^m,d_i^o,d_i^m,\boldsymbol{r}_i^*)}{f(\boldsymbol{y}_i^o,d_i^o,\boldsymbol{r}_i^*)} \\ &= \frac{\int_{\widetilde{\boldsymbol{b}}_i}\varphi_1\varphi_2\varphi_3\varphi_4\varphi_5 f(\widetilde{\boldsymbol{b}}_i)\,d\widetilde{\boldsymbol{b}}_i}{\int_{d_i^m}\int_{\boldsymbol{y}_i^m}\int_{\widetilde{\boldsymbol{b}}_i}\varphi_1\varphi_2\varphi_3\varphi_4\varphi_5 f(\widetilde{\boldsymbol{b}}_i)\,d\widetilde{\boldsymbol{b}}_i\,d\boldsymbol{y}_i^m\,dd_i^m} \\ &= \frac{\int_{\boldsymbol{m}_i}\varphi_2\varphi_4 f(\boldsymbol{m}_i)\,d\boldsymbol{m}_i}{\int_{d_i^m}\int_{\boldsymbol{y}_i^m}\int_{\boldsymbol{m}_i}\varphi_2\varphi_4 f(\boldsymbol{m}_i)\,d\boldsymbol{m}_i\,d\boldsymbol{y}_i^m\,dd_i^m} \\ &\quad\times\frac{\int_{\overline{\boldsymbol{b}}_i}\varphi_1\varphi_3\varphi_5 f(\overline{\boldsymbol{b}}_i)\,d\overline{\boldsymbol{b}}_i}{\int_{\overline{\boldsymbol{b}}_i}\varphi_1\varphi_3\varphi_5 f(\overline{\boldsymbol{b}}_i)\,d\overline{\boldsymbol{b}}_i} \\ &= f(\boldsymbol{y}_i^m,d_i^m|\boldsymbol{y}_i^o,d_i^o), \end{aligned} \tag{5.15}$$

with $\varphi_1 \ldots \varphi_5$ as in (5.14). The above results show that the sub-class satisfies MAR. Therefore, a sufficient condition for our extended model to satisfy MAR is that the

random effects influencing the observed measurements and/or the coarsening mechanism do not influence the missing measurements, given the observed ones. This is equivalent to the condition that all information about the missing measurements stems from the observed measurements and covariates only.

5.3.5 An MAR Counterpart to an Extended Shared-parameter Joint Model for Longitudinal and Time-to-event Data

The developments in Sections 5.3.3 and 5.3.4 allow us to construct an MAR counterpart for any member of the extended model (5.7), with exactly the same fit to the observed data. This can be done by integrating over the distribution of the missing components given the observed ones (Molenberghs *et al.*, 2008). Therefore, $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{b_i})$ and $f(d_i^m|d_i^o, \boldsymbol{b_i})$ in (5.7) need to be replaced with, respectively,

$$\begin{aligned} h(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{m}_i^*) &= \int_{\boldsymbol{b_i}^*} f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{b_i}) d\boldsymbol{b_i}^*, \\ h(d_i^m|d_i^o, \boldsymbol{m}_i^*) &= \int_{\boldsymbol{b_i}^*} f(d_i^m|d_i^o, \boldsymbol{b_i}) d\boldsymbol{b_i}^*, \end{aligned} \tag{5.16}$$

where integration over $\boldsymbol{b_i}^*$ is over all random effects in the full set $\boldsymbol{b_i}$, except possibly those that are specific to either $\boldsymbol{y}_i^m$, or d_i^m, or both.

This non-uniqueness is an important and somewhat disconcerting point. At first, it may seem that it dismisses the use of the proposed framework. However, it is important to see that this is not particular to the model described here. Rather, it always occurs when data are incomplete, as brought forward in Molenberghs and Verbeke (2004). The same is true for models with unobservables such as random effects, as studied by Verbeke and Molenberghs (2010). More generally, whenever the model specifies more than what is available in the data, such non-uniqueness occurs. This has been brought forward in Molenberghs *et al.* (2012). It means that great care is needed when interpreting results from models with incomplete and censored observations, just as well as for random-effects models, models with latent variables and latent classes, factor-analytic models, etc. This implies that apart from goodness-of-fit, which still has a place but only addresses how well a model described the *observed* data, sensitivity analysis needs to be done, studying how assumptions about the unobservables, given the observables, influence the inferences drawn. The advantage of our framework is that this is clearly brought to the foreground, reducing the risk of a false sense of security (Creemers *et al.*, 2010).

Obviously, if for instance $\boldsymbol{b_i}$ above is a single common random effect, then the

marginalization provided above means that if patients' estimated longitudinal profiles would be extended beyong the point of censoring, then, for a given combination of baseline covariates, the MAR model would reduce the predicted post-censoring trajectories to a common profile. The MNAR model would produce different trajectories. This is of course because the subject-specific effects (random effects) would come then into play under the MNAR, while the same is not the case under the MAR. In addition, if, for censored patients, we would consider their predicted hazard for death, then, for patients having the same combination of baseline covariates, the MAR would produce a common predicted hazard curve, while the MNAR would produce different curves.

5.3.6 A Narrow Definition, and Its Limitations

We will now adopt a narrow definition of a joint model, and examine its main limitation, namely, that it defies an MAR characterization.

Before considering the narrow definition, we first set out some results on MAR in the PMM and SEM settings. We can then use a decomposition either in a PMM format:

$$f(\boldsymbol{Z}_i^o, \boldsymbol{Z}_i^m, \boldsymbol{r}_i^*|\boldsymbol{\theta}^*, \boldsymbol{\psi}^*) = f(\boldsymbol{Z}_i^o|\boldsymbol{r}_i^*, \boldsymbol{\theta}^*) f(\boldsymbol{Z}_i^m|\boldsymbol{Z}_i^o, \boldsymbol{r}_i^*, \boldsymbol{\theta}^*) f(\boldsymbol{r}_i^*|\boldsymbol{\psi}^*), \tag{5.17}$$

or in a SEM format:

$$f(\boldsymbol{Z}_i^o, \boldsymbol{Z}_i^m, \boldsymbol{r}_i^*|\boldsymbol{\theta}^{**}, \boldsymbol{\psi}^{**}) = f(\boldsymbol{Z}_i^o, \boldsymbol{Z}_i^m|\boldsymbol{\theta}^{**}) f(\boldsymbol{r}_i^*|\boldsymbol{Z}_i^o, \boldsymbol{Z}_i^m, \boldsymbol{\psi}^{**}). \tag{5.18}$$

Note that parameters are framework-specific. Further, the number of components in $\boldsymbol{Z}_i^o$ and $\boldsymbol{Z}_i^m$ varies from subject to subject, in line with what is common in the missing data literature (Rubin, 1976; Little and Rubin, 2002).

In a PMM, MAR is:

$$f(\boldsymbol{Z}_i^m|\boldsymbol{Z}_i^o, \boldsymbol{r}_i^*, \boldsymbol{\theta}^*) = f(\boldsymbol{Z}_i^m|\boldsymbol{Z}_i^o, \boldsymbol{\theta}^*), \tag{5.19}$$

whereas in a SEM it is:

$$f(\boldsymbol{r}_i^*|\boldsymbol{Z}_i^o, \boldsymbol{Z}_i^m, \boldsymbol{\psi}^{**}) = f(\boldsymbol{r}_i^*|\boldsymbol{Z}_i^o, \boldsymbol{\psi}^{**}). \tag{5.20}$$

Equation (5.19) implies that the distribution of the unobserved outcomes given the observed ones does not further depend on the coarsening mechanism.

The narrow definition assumes a single common underlying random-effects structure, as opposed to our extended approach. Decomposing this model in a SEM

fashion:

$$\begin{aligned} & f(\boldsymbol{Z}_i^o, \boldsymbol{Z}_i^m, \boldsymbol{r}_i^*, \boldsymbol{b}_i^{**} | \boldsymbol{\theta}^{**}, \boldsymbol{\psi}^{**}, \boldsymbol{\omega}^*) \\ &= f(\boldsymbol{Z}_i^o | \boldsymbol{b}_i^{**}, \boldsymbol{\theta}^{**}) f(\boldsymbol{Z}_i^m | \boldsymbol{Z}_i^o, \boldsymbol{b}_i^{**}, \boldsymbol{\theta}^{**}) f(\boldsymbol{r}_i^* | \boldsymbol{Z}_i^o, \boldsymbol{Z}_i^m, \boldsymbol{b}_i^{**}, \boldsymbol{\psi}^{**}) f(\boldsymbol{b_i}^{**} | \boldsymbol{\omega}^*), \quad (5.21) \\ &= f(\boldsymbol{Z}_i^o | \boldsymbol{b}_i^{**}, \boldsymbol{\theta}^{**}) f(\boldsymbol{Z}_i^m | \boldsymbol{Z}_i^o, \boldsymbol{b}_i^{**}, \boldsymbol{\theta}^{**}) f(\boldsymbol{r}_i^* | \boldsymbol{b}_i^{**}, \boldsymbol{\psi}^{**}) f(\boldsymbol{b_i}^{**} | \boldsymbol{\omega}^*), \quad (5.22) \end{aligned}$$

with $\boldsymbol{\omega}^*$ parameterizing the random-effects distribution, and the step from (5.21) to (5.22) following from conditional independence. Formulation (5.22) means that coarsening cannot depend on the unobserved longitudinal measurements and the future time (whether survival or censoring), *given the random effect*. Hence, by construction, we have MNAR, unless in the trivial case where the random effect drops out from either the Z-factors, or from the r-factor, or from both. In the latter case MCAR applies. As stated in Creemers *et al.* (2011) for longitudinal data, this means that such a narrow formulation cannot admit MAR. This is in contrast to our results in Sections 5.3.3 and 5.3.4, where MAR conditions are established for the extended framework.

5.4 Some Considerations

In the previous section, we have relied on the correspondence established at the beginning, to develop theory which parallels that in the missing data setting. It was also mentioned earlier that the current setting provides a more complex coarsening mechanism. It is instructive to then note that there are interrelationships between some of the components in (5.7), which make model formulation in the current setting exceedingly more complex than in the missing data setting provided by (5.3).

Restricting ourselves to drop-out, and considering the case of fatal events, then the distribution of $\boldsymbol{R}_i$ is determined by that of D_i^o and D_i^m. This is because the probability of drop-out at time t is the probability that the patient experiences the event of interest at time t, or censoring occurs at time t. On the other hand, the distribution of W_i is also determined by D_i^o and D_i^m, since $P(W_i = 1) = P(T_i < C_i)$.

The effect of these difficulties on (5.7), and hence on the ensuing results, and the possibility of expanding (5.7) to take into account such difficulties, is out of the scope of this chapter.

Notice that the case of non-fatal events in relation to drop-out is different, since occurrence of the event may or may not lead to drop-out.

5.5 Sensitivity Analysis

The above-mentioned difficulties notwithstanding, the elaborate random effects structure provides an avenue for sensitivity analyes. A series of models can be formulated, making different assumptions about the latent structure that influences the longitudinal, the time-to-event and the censoring processes, and the stability of inferences then studied. Creemers *et al.* (2010) highlight such an avenue in the missing data context. In what follows, we formulate a model where we assume a common random effect to influence the longitudinal process, survival time, and the censoring time. This is compared to a conventional analysis, where censoring is treated as non-informative.

5.5.1 The Liver Cirrhosis Data

These data were introduced in Section 2.2. As mentioned, we will consider the quasi-continuous prothrombin index, viewed a marker for severity. We formulate an extended model for these data, and compare the ensuing inferences, especially regarding the treatment effects, with those of a conventional analysis.

For the longitudinally recorded prothrombin index, we assume the following linear mixed model:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 X_i + \beta_3 X_i t_{ij} + g_i + \varepsilon_{ij}, \tag{5.23}$$

with $g_i \sim N(0, \sigma_g^2)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$, and g_i and ε_{ij} independent. Further, t_{ij} is the time at which the jth measurement for the ith individual was taken, and X_i the treatment indicator. While the main effect of treatment is expected to be zero, stemming from randomization, it is kept in the model to absorb small departures in randomization equilibrium. The distributional assumption about g_i made above will also be kept in what follows.

For the extended analysis, we assume the following, for the survival time t and censoring time c:

$$f_i(t|g_i) = \lambda_1 \rho_1 t^{\rho_1 - 1} \exp(X_i \xi_1 + \kappa_1 g_i) \exp\{-\lambda_1 t^{\rho_1} \exp(X_i \xi_1 + \kappa_1 g_i)\}, \tag{5.24}$$

$$f_i(c|g_i) = \lambda_2 \rho_2 c^{\rho_2 - 1} \exp(X_i \xi_2 + \kappa_2 g_i) \exp\{-\lambda_2 t^{\rho_2} \exp(X_i \xi_2 + \kappa_2 g_i)\}, \tag{5.25}$$

where κ_1 and κ_2 are scale factors. In other words, we formulate proportional hazards models, with Weibull baseline hazards. The extended model assumes a common normal random effect to influence the longitudinal process, survival time, and the censoring time.

For the conventional analysis, we assume a conventional shared-parameter model, where the longitudinal and survival outcomes are connected through a normal random

effect g_i, but with censoring being assumed non-informative. In this case, the contribution of the censoring times to the likelihood is that of survival probabilities; i.e., the contributions of survival times, t, and censoring times, c, are as follows, respectively:

$$f_i(t|g_i) = \lambda_1 \rho_1 t^{\rho_1 - 1} \exp(X_i\xi_1 + \kappa_1 g_i) \exp\{-\lambda_1 t^{\rho_1} \exp(X_i\xi_1 + \kappa_1 g_i)\}, \quad (5.26)$$

$$S_i(c|g_i) = \exp\{-\lambda_1 t^{\rho_1} \exp(X_i\xi_1 + \kappa_1 g_i)\}. \quad (5.27)$$

Notice that this is also based on a proportional hazards model, with a Weibull baseline hazard.

Both models were fitted using the NLMIXED procedure in SAS, version 9.3; analysis code for the extended analysis is provided in Appendix C. Both prothrombin and time were rescaled to the unit interval, for numerical stability and without statistical or substantive consequences. The results are provided in Table 5.1. First, based on AIC values, the extended analysis provides better fit. Of course, it is important to remember that, as mentioned earlier, the observed data do not contain information to verify model assumptions, hence the need for sensitivity analyses. Second, it is interesting to note that both analyses closely agree; indeed, both tests for treatment effect on the evolution of prothrombin, as well as on the hazard of death, are not statistically significant under both analyses. Therefore, in this case, there is no noticeable sensitivity. Based on the scale factor related to survival time in both analyses, we have evidence of a reduced hazard of death for a higher patient-specific prothrombin profile. In addition, given also the statistical significance of the scale factor κ_2, a higher patient-specific prothrombin profile is also seen to correspond to a reduced hazard of censoring.

5.6 Discussion

In this chapter, we have used the strong connection between the longitudinal and time-to-event setting, and the missing data one, to build an extended shared random effects joint model, similar in spirit to that of Creemers *et al.* (2011) in the context of longitudinal data subject to missing observations, but now transposed to the current more complex setting. In doing so, we have taken a slightly different perspective on joint models than is prevalent in the literature.

Within this extended framework, we have provided a characterization of MAR, consistent to the one in the missing data setting. While the framework has been built conceptually, we have used the elaborate random-effects structure as an avenue for sensitivity analysis in this context. We have illustrated such an analysis using data from a study on liver cirrhosis.

Table 5.1: *Liver Cirrhosis Data. Parameter estimates (standard errors) for a conventional and an extended analysis.*

		Extended analysis	Conventional analysis
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Longitudinal process (Prothrombin)			
Intercept	β_0	-0.4330 (0.0079)	0.4326 (0.0080)
Time	β_1	0.1027 (0.0177)	0.0967 (0.0175)
Treatment	β_2	-0.0405 (0.0113)	-0.0409 (0.0114)
Treatment-by-time interaction	β_3	0.0370 (0.0258)	0.0315 (0.0257)
Measurement error	σ^2	0.0111 (0.0003)	0.0111 (0.0003)
Survival time			
Treatment	ξ_1	0.0324 (0.1214)	0.0493 (0.1378)
Scale	λ_1	4.2064 (0.4192)	1.9817 (0.2161)
Shape	ρ_1	0.9569 (0.0455)	0.9076 (0.0443)
Scale factor	κ_1	-2.6694 (0.6496)	-6.7345 (0.6980)
Censoring time			
Treatment	ξ_2	-0.1572 (0.1477)	
Scale	λ_2	3.6254 (0.4291)	
Shape	ρ_2	1.0450 (0.0638)	
Scale factor	κ_2	-2.8537 (0.8941)	
Shared effect			
Variance of random effects	σ_g^2	0.0119 (0.0009)	0.0123 (0.0010)
Model fit [Akaike information criterion (AIC)]			
		-4361	-4014

Needless to say, the longitudinal and time-to-event setting is more complex than the missing data one, given the added levels of coarsening. It has been highlighted that model formulation under the framework studied here then becomes exceedingly complex, given the interrelationships that arise among model components. The effect of such difficulties on the studied framework, especially on the results established here, and the possibility of expanding the framework to take into account such difficulties, requires further investigation.

A sensitivity analysis has been conducted by formulating an extended model, next to a conventional one. It is still possible, within the elaborate random effects structure, to formulate additional alternative models, all making different assumptions. The

stability of inferences across such models can then be considered. This provides a rich avenue for sensitivity analyses.

A further point requires attention. In Section 5.4, we have distinguished between fatal and non-fatal events. A fatal or terminating event can take two forms. First, the event can be terminal in the sense that it forces the patient to be taken off study. An example is blood pressure crossing a pre-specified threshold. Second, it can be death, like in the liver cirrhosis data. On the one hand, one can argue that models should avoid making predictions of measurements taken past death; this view-point starts from conditioning on a patient's actual history. On the other hand, one can take a marginal perspective and consider such patients as representative of a sub-population with common history and with a certain probability of dying. In the latter view, predictions past the point of death are still interpretable. This important point is not considered further in this chapter, but the interested reader is referred to Kurland and Heagerty (2005) and Kurland *et al.* (2009).

Finally, the method proposed here comes at a computational cost. Even though it is possible to use, without too much difficulty, existing statistical software, multi-dimensional random effects structures will make the method more complicated. This is not particular to the proposed model but rather to any mixed model that is at least as complex as a generalized linear mixed model with increasing dimension of the random-effects vector. Various authors have considered ways to simplify the model fitting. This includes, for example, pseudo-likelihood or composite likelihood, an account of which is given in Molenberghs and Verbeke (2005) and Molenberghs *et al.* (2011). Further treatment of this topic is outside the scope of this chapter.

Chapter 6

Enriched-data Problems and Essential Non-identifiability

There are two principal ways in which statistical models extend beyond the data available. First, the data may be coarsened; that is, what is actually observed is less detailed than what is planned, owing to, for example, attrition, censoring, grouping, or a combination of these. Second, the data may be augmented, that is, the observed data are hypothetically but conveniently supplemented with structures such as random effects, latent variables, latent classes, or component membership in mixture distributions. These two settings together will be referred to as *enriched data*. Reasons for modelling enriched data include the incorporation of substantive information, such as the need for predictions, advantages in interpretation, and mathematical and computational convenience. The fitting of models for enriched data combine evidence arising from empirical data with non-verifiable model components, *i.e.,* that are purely assumption driven. This has important implications for the interpretation of statistical analyses in such settings. While widely known, the exploration and discussion of these issues is somewhat scattered. The user should be fully aware of the potential dangers and pitfalls that follow from this. Verbeke and Molenberghs (2010) considered the incomplete-data and random effects models case, while Molenberghs *et al.* (2012) addressed this issue in a broader framework, encompassing a bigger number of seemingly disparate enriched-data settings. In this chapter, we explore several specific settings, namely that of latent classes, finite mixtures, factor analysis, and frailty models. The results are applied to a range of relevant examples.

6.1 Introduction

It is common in statistics to use models that rely on assumptions that cannot be examined from the data under analysis. This is not a weakness, but an inevitable consequence of drawing statistical inferences in the settings in which such models are used. As a consequence, it is important that the use of these models properly reflects the implied reliance on external information. A good example of the failure to appreciate the nature of such models is provided by the now well-known historical developments surrounding factor analysis in so-called general intelligence measurement(Gould, 1981). Factor analysis dates back to Pearson and Spearman, though it is the latter that is credited with its introduction into psychology, a field in which it has held popularity for close to a century. Its arrival coincided with a time when psychologists were attempting to quantify 'mental worth' in a scientific sense. Motivated by the positive correlations exhibited by a set of mental tests, Spearman used the technique to develop the so-called "two-factor" theory. The theory implies that a set of mental tests represents an underlying general factor (g), in addition to each test's specific information. Spearman proceeded to accord g some real existence, terming it *general intelligence*. He further proceeded to identify g as an attribute, resident in the brain, which he called *general energy*. Some physical existence is also attributed to the test-specific information (s-factors): he identified them as specific engines in the brain, which are under the influence of the general energy. He also argued g as the theoretical basis of the IQ-testing, which was prevalent at the time: the IQ-test simply measures g, with each component test having a certain loading on g, and certain test-specific information, s. There was no corroborating structural neurological work to support this theory, however. The attribution of real existence to such mathematically constructed abstractions is an example of *reification*. A debate ensued, between two schools of thought: Spearman and Burt on one side and Thurstone on the other. We note that Burt, like Spearman, believed in the supremacy of g, though he also believed in the existence of *group factors*, subsidiary to g. Thurstone faulted Spearman's (and Burt's) method, and produced a solution which totally dispenses with g. The solution, which he called *simple structure*, is actually a rotation of Spearman's principal-components solution. The two solutions explain an identical amount of information, i.e., they fit the observed data equally well. Hence, they differ only in aspects of the model that cannot be verified from the data. The value of their respective solutions including their non-verifiable assumptions rests entirely on practical considerations. To show that this phenomenon is very common throughout statistical modeling, and extends across a range of common data-analytic settings,

well beyond factor analysis, is the central theme of this chapter.

At the time, rather than view this as an indication of the need to acknowledge sensitivity, and consequently refrain from reification, Thurstone proceeded to present his solution as a discovery of the correct explanation of the structure of the mind. The two schools of thought passionately advocated the validity of their model as the proper representation of the mind. Gould (1981) exposes a fundamental flaw which both parties failed to take notice of: that their respective solutions comprised of positioning of axes at locations which represented their *a priori* suppositions of the nature of the mind. Therefore, their respective models merely mirrored their prior belief.

In the following, we illustrate, through a range of settings, the common structure of problems like that of factor analysis above, and show how the practical implications of these rest on a division of information into that supplied by the data under analysis and that supplied externally. Molenberghs *et al.* (2012) distinguish two broad types of settings. The first can be termed *augmented* data, in the sense of supplementing the observed data with latent or unobserved quantities; examples include random-effects models, latent class and latent variable models, and finite-mixture models. The second, introduced by Heitjan (Heitjan and Rubin, 1991; Zhang and Heitjan, 2007) is a concept called *coarsening*, which refers broadly to situations where the observed data are coarser than the hypothetically conceived data structures, to which the models of interest apply. Examples include incomplete data and censored survival data. It is obvious that models for such augmented structures or coarsened data are identifiable only by virtue of making sometimes strong but always partially non-verifiable assumptions. Augmentation and coarsening taken together, and from now on termed *enriched data*, in line with Verbeke and Molenberghs (2010), will be treated in a unified way, such that important, common features can be illuminated and scrutinized. There is a formal distinction between the two types. In the coarse-data setting, it is understood that a part of the data would ideally be observed but is not in practice (*e.g.*, actual survival time after censoring, outcomes after dropout, etc.). Augmented data refers rather to the addition of useful but artificial constructs to the data setting, such as random effects, latent classes, latent variables, factors, and mixture component membership. These can never be observed. The focus in Molenberghs *et al.* (2012) is not so much on the distinctions between coarse and augmented data on the one hand, nor on subtle distinctions within the coarsening and augmentation families on the other; rather, they review a selected range of each and bring out commonality.

In this chapter, we focus on a number of relevant settings, and establish that there will always be a part of the model that is totally unidentifiable from the observed

data. This implies that the identification of such a part can come from assumptions only. This leads us to the main message of the chapter. First, we set how models in enriched-data settings are identified by the triple: data, design-based assumptions (such as randomization), and further unverifiable assumptions. For this we focus on the model itself and its relationship to the data through likelihood. We are not concerned with subsequent inferences; the same message holds whether we are being Bayesian or frequentist. In each setting considered we identify a part of the model for which, in the Bayesian case, the posterior depends only on the choice of prior (assuming appropriate independence relationships among components of the prior), and in the frequentist case that does not affect goodness-of-fit to the observed data. Second, while various forms of this are known in various sub-fields, to variable degrees, we emphasize the great similarity between these fields and settings; appropriate review of a number of selected areas is presented to facilitate study of the common features. We illustrate this by showing how non-identified parts can be replaced arbitrarily, without altering the fit to the observed data but with potentially non-trivial consequences for inferences and substantive conclusions. It should be clear that this can be dangerous and the user must carefully reflect on the arbitrary components. For example, they should be supported by substantive considerations or be made part of a sensitivity analysis. Therefore, acceptable goodness-of-fit to the observed data cannot be used as the sole justification for the analysis. In the absence of external corroborating knowledge or information, two alternative routes can be followed. First, it can be made clear that the conclusions drawn have meaning only under the external assumptions built into the analysis. For example, a researcher can choose to draw inferences given a set of scientifically plausible but otherwise non-verifiable causal relationships. It is then important not to divorce the data analysis from the assumptions made. Second, an appropriate sensitivity analysis can be conducted to augment the conclusions. By sensitivity analysis, we mean in this context, either a study of how unverifiable assumptions affect overall inferences, or an assessment of *traceability* (Molenaar, 2004, 2008), i.e., how unverifiable assumptions influence predictions for individual subjects. For example, analyses can be conducted under a number of alternative sets of hypothesized structures as well. This then allows the researcher to examine the sensitivity of the inferences concerning the scientific question to varying the underlying assumptions. See, for example, Part V of Molenberghs and Kenward (2007)

In the next section, we introduce our general results concerning enriched data structures, in particular showing how components of the models can be chosen in an effectively infinite number of ways without affecting the fit to the observed data. In

the subsequent sections, these general results are applied to five widely used settings, namely that of latent class models (Section 6.3), finite-mixture models (Section 6.4), factor analysis (Section 6.5), and frailty models (Section 6.6), and practical implications are illustrated using the examples.

6.2 General Result About Counterparts in Enriched-data Structures

The result in this section is based upon Verbeke and Molenberghs (2010). Assume data $\boldsymbol{Z}_i$ for an independent unit $i = 1, \ldots, N$ are augmented with $\boldsymbol{c}_i$. The $\boldsymbol{c}_i$ can take any conventional enriched-data form. For example, the vector can refer to missing measurements, random effects, or perhaps a combination of both. An example of a setting where the latter situation arises naturally is the shared-parameter framework, that will be considered in the next section.

Assume a joint model of the generic form $f(\boldsymbol{z}_i, \boldsymbol{c}_i|\boldsymbol{\theta}, \boldsymbol{\psi})$, where covariates have been suppressed for notational simplicity. We assume the parameters to be disjoint, in the sense of Rubin (1976), meaning that the parameter space of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ equals the set theoretic product of the individual parameter spaces. Consider the factorizations:

$$f(\boldsymbol{z}_i, \boldsymbol{c}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{z}_i|\boldsymbol{c}_i, \boldsymbol{\theta}) f(\boldsymbol{c}_i|\boldsymbol{\psi}), \tag{6.1}$$

$$= f(\boldsymbol{z}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) f(\boldsymbol{c}_i|\boldsymbol{z}_i, \boldsymbol{\theta}, \boldsymbol{\psi}). \tag{6.2}$$

Borrowing terminology from the hierarchical-models context, such as mixed models, which are given consideration in Verbeke and Molenberghs (2010); Molenberghs *et al.* (2012), every factor in both (6.1) and (6.2) can usefully be given a name. The left hand side is the *joint model*. Consider first the right hand sides. The first factor in (6.1) is the *hierarchical model* and the second one is the *prior density* for the enriched data. The first factor in (6.2) may be termed the *marginal model*, whereas the second one is the *posterior density* of the enriched data.

The above terminology makes clear the obvious link between (6.1)–(6.2) and the mixed-model setting. The link with incomplete data follows by setting $\boldsymbol{c}_i \equiv \boldsymbol{y}_i^m$ and $\boldsymbol{z}_i = (\boldsymbol{y}_i^o, \boldsymbol{r}_i)$.

These considerations immediately establish the following theorem.

Theorem 1 (A Family of Counterparts to a Given Model for Enriched Data.) *Let us assume that data $\boldsymbol{z}_i$ are enriched with $\boldsymbol{c}_i$. Then, any model (6.1) formulated for and fitted to such data, can be replaced by an infinite family of models, all retaining the fit to the observed data as achieved by the original model. This is*

done by preserving the marginal model $f(\boldsymbol{z}_i|\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\psi}})$ *and replacing the posterior density* $f(\boldsymbol{c}_i|\boldsymbol{z}_i,\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\psi}})$ *by an arbitrary conditional density*

$$f(\boldsymbol{d}_i|\boldsymbol{z}_i,\boldsymbol{\gamma}). \tag{6.3}$$

Here, $\boldsymbol{d}_i$ rather than $\boldsymbol{c}_i$ is used to indicate that there need not be any connection between the original and substituted enriched data. Also, the new density (6.3) can be parameterized by a completely new parameter $\boldsymbol{\gamma}$.

While it might be argued that the conventional error term in an ordinary regression equation already is an instance of enrichment, we choose to view this as different from the theme of this chapter. Broadly, in a univariate regression context (encompassing linear regression, analysis of variance, regression based on generalized linear models, etc.) the response is split into signal and noise. While this surely depends on the posited model (e.g., a linear model with a certain mean function), it is verifiable from the data, using a realm of fit and diagnostic tools. Our situation of interest is different because of two aspects. First, in augmented-data settings, the noise is split into several sources of noise, a split which cannot be verified definitively from data. Second, in coarse-data settings, models describe the unobserved outcomes, given the observed ones, and predict the same; the models, by construction, are not verifiable from the data.

It may seem that the above derivations violate the so-called extended likelihood principle (Bjørnstad, 1996), which states that the extended likelihood $f(\boldsymbol{z}_i,\boldsymbol{c}_i)$ carries all information in the data about the unobservables. Of course, this is a very sensible principle to make inferences *given the posited model*. Our main point is not to take issue with the extended likelihood principle, but rather to demonstrate how models, coinciding in $f(\boldsymbol{z}_i)$ but differing in $f(\boldsymbol{c}_i|\boldsymbol{z}_i)$, are indistinguishable in terms of the data only. In contrast, the extended likelihood principle states that, *once a particular model has been chosen,* parametric inferences about the parameters governing the joint distribution, follow through the extended likelihood function.

6.3 Case I: Latent Classes and Latent Variables

Latent class (LC) models are widely used, especially in the social and behavioral sciences (Goodman, 1974; Xu and Craig, 2009), where they are used to identify subgroups of individuals, based on phenomena defined in terms of categorical data. The observed variables are assumed to be a manifestation of some underlying categorical latent variable, the levels of which are believed to organize individuals into subgroups exhibiting distinct tendencies. In a latent variable (LV) model the unobservable is

of a continuous nature. From a statistical perspective, use of the LC model may be viewed as a way of addressing heterogeneity among observations. A qualitative mixture distribution is assumed, and the observed, also called manifest or indicator variables, are assumed to be independent, conditional on the latent class. This is termed the local independence assumption.

When considered in a broad sense, LC and LV models exhibit connections with item-response theory models (Tatsuoka, 1990), shared-parameter models for incomplete data Molenberghs *et al.* (2012), and factor analysis (Section 6.5). LC and LV models are based on unobservables. It is therefore impossible to decide, in terms of the data alone, whether there are in fact such latent classes and, if we assume that there are, how many exist, and the number of categories in each. The 'identification' of the number of latent classes bears similarity to the identification of the number of components in mixture models (Section 6.4).

Suppose we observe response variables $Y_1, Y_2, \ldots, Y_T$, each with C_t categories, $t = 1, \ldots, T$, and assume a categorical latent variable, Z, with g levels. The basic latent class (LC) model takes the following form:

$$P(Y_1 = y_1, \ldots, Y_T = y_T, Z = z) = P(Z = z) \prod_{t=1}^{T} P(Y_t = y_t | Z = z). \tag{6.4}$$

This is called the probabilistic representation of the model, the parameters of which are the conditional (item-response) probabilities: $P(Y_t = y_t | Z = z)$, and the latent class probabilities (prevalences): $P(Z = z)$. An equivalent representation of the basic LC model, called the log-linear representation, takes the form:

$$\log P(Y_1 = y_1, \ldots, Y_T = y_T, Z = z) = \lambda + \lambda_z^Z + \sum_{t=1}^{T} \lambda_{y_t}^{Y_t} + \sum_{t=1}^{T} \lambda_{z y_t}^{Z Y_t}. \tag{6.5}$$

The link between the conditional and log-linear model parameters is the following:

$$P(Y_t = y_t | Z = z) = \frac{\exp\left(\lambda_{y_t}^{Y_t} + \lambda_{z y_t}^{Z Y_t}\right)}{\sum_{i=1}^{C_t} \exp\left(\lambda_i^{Y_t} + \lambda_{z i}^{Z Y_t}\right)}. \tag{6.6}$$

An iterative procedure, such as the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977), is used for model estimation. The Akaike Information Criterion (AIC) and the likelihood ratio statistic (G^2) are typically used in evaluating the appropriate number of latent classes. The likelihood ratio statistic evaluates the proximity of the expected cell frequencies to the observed cell frequencies, whereas AIC adds penalty to this that depends on the number of parameters in the model. We now apply Theorem 1 to model (6.4), replacing the posterior distribution with two rather

different choices: (a) the normal distribution and (b) a distribution corresponding to the posterior distribution of a model with $k \neq g$ latent classes.

We first set out the components in (6.1)–(6.2) for model (6.4). The joint model is simply the exponent of model (6.5), with expression

$$P(Y_1 = y_1, \ldots, Y_T = y_T, Z = z) = \exp\left(\lambda + \lambda_z^Z + \sum_{t=1}^{T} \lambda_{y_t}^{Y_t} + \sum_{t=1}^{T} \lambda_{zy_t}^{ZY_t}\right). \tag{6.7}$$

From (6.6), and by the local independence assumption of the LC model, the hierarchical model is seen to take the form:

$$\begin{aligned} P(Y_1 = y_1, \ldots, Y_T = y_T | Z = z) &= \prod_{t=1}^{T} P(Y_t = y_t | Z = z) \\ &= \prod_{t=1}^{T} \frac{\exp\left(\lambda_{y_t}^{Y_t} + \lambda_{zy_t}^{ZY_t}\right)}{\sum_{i=1}^{C_t} \exp\left(\lambda_i^{Y_t} + \lambda_{zi}^{ZY_t}\right)} \end{aligned} \tag{6.8}$$

and the prior distribution, the ratio of the joint to the hierarchical model, is:

$$P(Z = z) = \frac{\exp\left(\lambda + \lambda_z^Z + \sum_{t=1}^{T} \lambda_{y_t}^{Y_t} + \sum_{t=1}^{T} \lambda_{zy_t}^{ZY_t}\right)}{\prod_{t=1}^{T} \frac{\exp\left(\lambda_{y_t}^{Y_t} + \lambda_{zy_t}^{ZY_t}\right)}{\sum_{i=1}^{C_t} \exp\left(\lambda_i^{Y_t} + \lambda_{zi}^{ZY_t}\right)}}. \tag{6.9}$$

The marginal model is a weighted sum of the hierarchical probabilities:

$$P(Y_1 = y_1, \ldots, Y_T = y_T) = \sum_{z=1}^{g} P(Y_1 = y_1, \ldots, Y_T = y_T, Z = z) \tag{6.10}$$

and hence, the posterior distribution follows as:

$$\begin{aligned} &P(Z = z | Y_1 = y_1, \ldots, Y_T = y_T) \\ &\quad = \frac{\exp\left(\lambda + \lambda_z^Z + \sum_{t=1}^{T} \lambda_{y_t}^{Y_t} + \sum_{t=1}^{T} \lambda_{zy_t}^{ZY_t}\right)}{\sum_{z=1}^{g} \exp\left(\lambda + \lambda_z^Z + \sum_{t=1}^{T} \lambda_{y_t}^{Y_t} + \sum_{t=1}^{T} \lambda_{zy_t}^{ZY_t}\right)}. \end{aligned} \tag{6.11}$$

We now turn to each of the posteriors.

6.3.1 Normal Posterior

We retain the marginal model (6.10), but replace the sets of probabilities given in (6.11) with a unit-variance normal density and linear mean model:

$$f(h|\boldsymbol{Y}) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}\left[h - (\alpha_0 + \sum_{t=1}^{T} \alpha_t Y_t)\right]^2}. \tag{6.12}$$

The new joint model follows as the product of (6.10) and (6.12), with now prior distribution

$$\frac{e^{\lambda}}{\sqrt{2\pi}}\sum_{y_1}\cdots\sum_{y_T}e^{\frac{1}{2}\left[h-(\alpha_0+\sum_{t=1}^{T}\alpha_t Y_t)\right]^2}\sum_{z=1}^{g}\exp\left(\lambda+\lambda_z^Z+\sum_{t=1}^{T}\lambda_{y_t}^{Y_t}+\sum_{t=1}^{T}\lambda_{zy_t}^{ZY_t}\right) \quad (6.13)$$

and hierarchical model

$$\frac{\frac{1}{\sqrt{2\pi}}e^{\frac{1}{2}\left[h-(\alpha_0+\sum_{t=1}^{T}\alpha_t Y_t)\right]^2}\Psi}{\frac{e^{\lambda}}{\sqrt{2\pi}}\sum_{y_1}\cdots\sum_{y_T}e^{\frac{1}{2}\left[h-(\alpha_0+\sum_{t=1}^{T}\alpha_t Y_t)\right]^2}\Psi}, \quad (6.14)$$

where

$$\Psi=\sum_{z=1}^{g}\exp\left(\lambda+\lambda_z^Z+\sum_{t=1}^{T}\lambda_{y_t}^{Y_t}+\sum_{t=1}^{T}\lambda_{zy_t}^{ZY_t}\right).$$

We note here that (6.13) and (6.14) complete a new hierarchical specification, which gives the same marginal fit (6.10) achieved by the initial set (6.8) and (6.9). However, as will be clear from the data analysis below, there are consequences for ensuing inferences.

6.3.2 Distribution Corresponding to the Posterior of a Model With $k \neq g$ Latent Classes

We now couple (6.10) with

$$\begin{aligned}&P(X=x|Y_1=y_1,\ldots,Y_T=y_T)\\ &=\frac{\exp\left(\beta+\beta_x^X+\sum_{t=1}^{T}\beta_{x_t}^{X_t}+\sum_{t=1}^{T}\beta_{xy_t}^{XY_t}\right)}{\sum_{x=1}^{k}\exp\left(\beta+\beta_x^X+\sum_{t=1}^{T}\beta_{y_t}^{Y_t}+\sum_{t=1}^{T}\beta_{xy_t}^{XY_t}\right)}.\end{aligned} \quad (6.15)$$

The prior distribution and the hierarchical model, respectively, can then be seen to take the following forms:

$$\begin{aligned}&\sum_{y_1}\cdots\sum_{y_T}\left[\sum_{z=1}^{g}\exp\left(\lambda+\lambda_z^Z+\sum_{t=1}^{T}\lambda_{y_t}^{Y_t}+\sum_{t=1}^{T}\lambda_{zy_t}^{ZY_t}\right)\right]\\ &\times\ \frac{\exp\left(\beta+\beta_x^X+\sum_{t=1}^{T}\beta_{x_t}^{X_t}+\sum_{t=1}^{T}\beta_{xy_t}^{XY_t}\right)}{\sum_{x=1}^{k}\exp\left(\beta+\beta_x^X+\sum_{t=1}^{T}\beta_{y_t}^{Y_t}+\sum_{t=1}^{T}\beta_{xy_t}^{XY_t}\right)}\end{aligned} \quad (6.16)$$

and

$$\frac{f(y_1,\ldots,y_T)}{\sum_{y_1}\cdots\sum_{y_T}f(y_1,\ldots,y_T)}, \quad (6.17)$$

with

$$f(y_1,\ldots,y_T) = \sum_{z=1}^{g}\exp\left(\lambda+\lambda_z^Z+\sum_{t=1}^{T}\lambda_{y_t}^{Y_t}+\sum_{t=1}^{T}\lambda_{zy_t}^{ZY_t}\right)\times \\ \times\frac{\exp\left(\beta+\beta_x^X+\sum_{t=1}^{T}\beta_{x_t}^{X_t}+\sum_{t=1}^{T}\beta_{xy_t}^{XY_t}\right)}{\sum_{x=1}^{k}\exp\left(\beta+\beta_x^X+\sum_{t=1}^{T}\beta_{y_t}^{Y_t}+\sum_{t=1}^{T}\beta_{xy_t}^{XY_t}\right)}.$$

Note that (6.16) and (6.17) complete yet another hierarchical specification, giving the same marginal fit (6.10).

6.3.3 Data Analysis

We now illustrate the above developments using the 2005 United States' National Youth Risk Behavior Survey data ($N = 13,840$), introduced in Section 2.7. We consider the 12 questions introduced earlier. Collins and Lanza (2009) have previously extensively analyzed these variables, and chosen a 5-class LC model. We use the SAS procedure LCA, an add-on procedure in SAS, to re-analyze these data. In Table 6.1, we present the model's latent class prevalence and item-response probabilities.

Table 6.1: *National Youth Risk Behavior Survey Data. Latent class model parameters.*

	Latent Class				
	1	2	3	4	5
Latent class prevalence	0.6741	0.1383	0.0910	0.0546	0.0420
	Probability "Yes"				
Driving after taking alc.	0.0058	0.4208	0.1488	0.4537	0.1098
Smoked before age 13	0.0422	0.1083	0.7584	0.6387	0.1738
Smoked daily for 30 days	0.0202	0.2670	0.3144	0.6588	0.1247
First alc. drink before age 13	0.1433	0.2075	0.7875	0.6790	0.3928
$\geq$5 alc. drinks/day, in past 30 days	0.0805	0.7421	0.4789	0.7875	0.1621
Took marijuana first before age 13	0.0074	0.0286	0.4596	0.5530	0.2173
Have ever used cocaine	0.0040	0.1919	0.0716	0.8800	0.0255
Tried glue sniffing, etc to get high	0.0550	0.1886	0.2153	0.5778	0.0420
Used methamphetamines	0.0035	0.0997	0.0245	0.7271	0.0102
Used ecstasy	0.0035	0.1093	0.0630	0.6429	0.0556
<13 years at first sexual intercourse	0.0138	0.0015	0.1753	0.2957	0.8073
Have had sex with at least 4 people	0.0639	0.2859	0.2409	0.5641	0.8348

These parameters are sufficient to calculate, by using Bayes' theorem, the posterior probability, where the latent classes and their corresponding probabilities act as (empirical) prior: the probability of belonging to a certain latent class, given a specific response pattern. Note that there is a total of 5×2^{12} posterior probabilities, where 5 represents the number of latent classes, and 2^{12} is the number of different response patterns possible. In practice, one needs to calculate only those that correspond to patterns actually occurring in the data. Classification of a respondent to any of the 5 latent classes, given his/her response pattern, is then based, for example, on the highest of the individual's set of 5 posterior probabilities: $P(Z = z|Y_1 = y_1, \ldots, Y_{12} = y_{12}), z = 1, \ldots, 5$. Alternatively, the set of posterior probabilities may be considered, which is especially instructive when a number of patterns differ only slightly in terms of posterior probability. As an example, consider the response pattern composed of a "Yes" response to all questions:

$$\begin{aligned} P(Z = z|Y_1 = \cdots = Y_{12} = \text{“Yes”}) &= \frac{P(Y_1 = \cdots = Y_{12} = \text{“Yes”}|Z = z)P(Z = z)}{P(Y_1 = \cdots = Y_{12} = \text{“Yes”})} \\ &= \frac{\prod_{t=1}^{12} P(Y_t = y_t|Z = z)P(Z = z)}{\sum_{z=1}^{5} P(Z = z)\left[\prod_{t=1}^{12} P(Y_t = y_t|Z = z)\right]}. \end{aligned}$$

Substituting the relevant parameters from Table 6.1, we obtain $(5.53E-18, 1.03E-08, 4.60E-06, 1.00, 7.33E-09)$ as the set of posterior probabilities, for latent classes 1–5 respectively. Clearly, classification of a respondent with such a response pattern would be to latent class 4, which, generally, has higher probabilities of a "Yes" response to the items than the other classes.

6.3.3.1 Normal Posterior

We now replace the posterior distribution with our first choice, the normal distribution. The parameters $\alpha_0, \ldots, \alpha_{12}$ play the role of sensitivity parameters; they can be freely specified, all without changing the marginal fit. Here, we set them to $\alpha_0 = \cdots = \alpha_{12} = 0.5$. Evidently, the concept of classifying a respondent to a particular latent class no longer exists. Notice that h in (6.12) is continuous, taking values on the whole real line, meaning that for any specific response pattern ($\boldsymbol{Y} = \boldsymbol{y}$), there exists an infinite collection of posterior densities. The prediction for the enrichment, $\widehat{h}$, is given as $E(h|\boldsymbol{Y}) = \alpha_0 + \sum_{t=1}^{12} \alpha_t y_t$. Letting y_t take the value 1 for a "Yes" response and 0 otherwise, we can calculate the prediction for h, for a specific response pattern. For example, for the response profile ($Y_1 = \cdots = Y_{12} =$ "Yes"), mentioned earlier, $\widehat{h} = 0.5 + 12 \times 0.5 = 6.5$. The point we make is that having replaced

the posterior distribution with our choice, we move to an entirely different setting, where, in contrast to the initial case where we could allocate to classes, we now work with a continuum. Once more, such manipulations are possible while the marginal fit remains unaffected.

6.3.3.2 Changing the Posterior With $k \neq 5$ Latent Classes

By choosing $k \neq 5$ in (6.15), we complete our choice. There is no information in the data about the β parameters; they can thus be specified freely. We only need to ensure that $\sum_x \beta_x^X = \sum_{y_1} \beta_{y_1}^{Y_1} = \cdots = \sum_{y_{12}} \beta_{y_{12}}^{Y_{12}} = \sum_{y_1} \beta_{y_1 x}^{Y_1 X} = \cdots = \sum_{y_{12}} \beta_{y_{12} x}^{Y_{12} X} = \sum_x \beta_{y_1 x}^{Y_1 X} = \cdots = \sum_x \beta_{y_{12} x}^{Y_{12} X} = 0$, so that the distribution is a genuine posterior from a latent class model. In so doing, we end up with completely different latent class allocations, though once more, the marginal fit remains the same.

6.4 Case II: Finite-mixture-model Component Membership

Finite mixture models (Böhning, 2000) are often used to handle heterogeneity arising from the postulation of unknown sub-populations, which are treated as latent. We assume that the response variable X follows a finite mixture distribution, formalized as

$$f(x) = \sum_{j=1}^{g} \pi_j f_j(x), \tag{6.18}$$

π_j, $j = 1, \ldots, g$ being the mixing proportion, *i.e.*, the proportion of the j^{th} sub-population in the population, and $f_j(x)$, $j = 1, \ldots, g$ the component densities, characterized by the parameters $\lambda_1, \ldots, \lambda_j$, respectively. The π_j satisfy $0 < \pi_j \leq 1$ and $\sum_j \pi_j = 1$. Sub-population membership is considered a latent variable, Z, with a discrete distribution P with values λ_j and corresponding probabilities π_j, for $j = 1, \ldots, g$. Next, we specify all components in (6.1) and (6.2), then illustrate arbitrariness of the posterior distribution. The hierarchical model is

$$f(x|Z = z, z = 1, \ldots, g) = f(\lambda_z), \tag{6.19}$$

with $f(\lambda_z)$ denoting the density characterized by the parameter λ_z. For instance, for a finite mixture of Poisson distributions, $f(x|Z = z) = f(\lambda_z)$ would be

$$\text{Poi}(\lambda_z). \tag{6.20}$$

We let

$$P(Z = z) = \pi_z, \tag{6.21}$$

be the prior distribution. The marginal distribution is obtained by summing-out Z:

$$f(x) = \sum_{z=1}^{g} f(x|Z = z)\pi_z. \tag{6.22}$$

A finite mixture of Poisson distributions, for instance, yields

$$f(x) = \pi_1 \cdot \text{Poi}(\lambda_1) + \cdots + \pi_g \cdot \text{Poi}(\lambda_g). \tag{6.23}$$

The joint model takes the form

$$f(x, z) = f(x|Z = z)\pi_z. \tag{6.24}$$

We therefore have that the posterior distribution, the ratio of (6.24) to (6.22), takes the form

$$P(Z = \ell, \ell = 1, \ldots, g|x) = \frac{f(x|Z = l) \cdot \pi_\ell}{\sum_{z=1}^{g} f(x|Z = z) \cdot \pi_z}. \tag{6.25}$$

This expression provides a channel through which data are a posteriori classified into the various sub-populations. A datum is classified into the sub-population for which $P(Z = l|x)$ is maximal. To illustrate sensitivity, we proceed as follows: retain the marginal model (6.22) but arbitrarily alter the posterior distribution (6.25). We replace the sets of probabilities in (6.25) by a continuous distribution,

$$f(g|x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2}, \tag{6.26}$$

$\mu(x) = \gamma x$. We note here that the data contains no information about γ, and we will have the liberty to set it to some value. The new joint model follows as the product of (6.22) and (6.26):

$$f(g, x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2} \sum_{z=1}^{g} f(x|Z = z) \cdot \pi_z. \tag{6.27}$$

For instance, for the Poisson mixture, we have

$$f(g, x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2} [\sum_{j=1}^{g} \pi_j \text{Poi}(\lambda_j)]. \tag{6.28}$$

The prior distribution follows by integrating or summing over X, depending on whether it is continuous or discrete. For discrete X,

$$f(g) = \sum_x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2} \sum_{z=1}^{g} f(x|Z = z) \cdot \pi_z. \tag{6.29}$$

For the Poisson mixture, where X is of course discrete, we have

$$f(g) = \sum_x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2} \left[\sum_{j=1}^{g} \pi_j \text{Poi}(\lambda_j) \right]. \tag{6.30}$$

The hierarchical distribution follows as the ratio of (6.27) to (6.29):

$$f(x|g) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2} \sum_{z=1}^{g} f(x|Z=z) \cdot \pi_z}{\sum_x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2} \sum_{z=1}^{g} f(x|Z=z) \cdot \pi_z}. \tag{6.31}$$

For the Poisson mixture case,

$$f(x|g) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x))]^2} \left[\sum_{j=1}^{g} \pi_j \text{Poi}(\lambda_j) \right])}{\sum_x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2} \left[\sum_{j=1}^{g} \pi_j \cdot \text{Poi}(\lambda_j) \right]}. \tag{6.32}$$

Thus, finally, (6.19) and (6.21) on the one hand, and (6.31) and (6.29) on the other, are two different hierarchical specifications, yielding the same marginal model, (6.22). Once more we have two models that are indistinguishable in terms of fit to the observed data, while the ensuing inferences are sensitive to the particular hierarchical formulation chosen. In the first formulation, it is possible to attribute *a posteriori* component membership to a given datum, through classification, based on (6.25). In the second formulation, however, this concept of classification disappears, because the posterior consists of a continuous density, naturally leading to prediction of the value of the (now continuous) latent variable, g, by noting that $E(g|x) = \gamma x$.

6.4.1 Data Analysis

The above developments are illustrated using the Accident Insurance Policies Data, introduced in Section 2.4. Böhning (2000) employed the Non-parametric Maximum Likelihood Estimation method, as implemented in the package C.A.MAn (Computer-Assisted Analysis of Mixtures and Applications), to fit a finite mixture of Poisson distributions to these data, and reached a three-component solution. We re-analyze these data and use the analysis to illustrate our result. The following model for X, the number of claims, is found:

$$f(x) = 0.4184\text{Poi}(0) + 0.5730\text{Poi}(0.3356) + 0.0087\text{Poi}(2.5454). \tag{6.33}$$

With this result, (6.25) can be used to allocate a specific datum, corresponding to x claims, to any of the mixture model components $z = 1, 2, 3$. For instance, for $x = 2$ counts, the set of probabilities $P(Z\ell|x)$, $\ell = 1, 2, 3$, is easily found to be

$(0.0000, 0.9125, 0.0875)$. Based on the maximal posterior allocation criterion, such a datum would be allocated to component 2. On the other hand, for $x = 5$ counts, the set would be $(0.0000, 0.0234, 0.9766)$, in which case the datum would be allocated to the third component. Of course, this is a clear situation; in cases where there is not a clear winner among the three component, presenting all three would be more insightful. No need to add, though, that this does not remove the enrichment aspect of the problem.

We now move to our second hierarchical formulation, where we assume a normal posterior. Fix the parameter γ to 0.5. Given the continuous nature, the action parallel to the above mixture component membership is to compute the predicted value for the latent variable, g. For $x = 2$ and $x = 5$ claims, respectively, we obtain 1 and 2.5, respectively.

In keeping with the theme of this chapter, the choice between these very different routes is not possible in terms of the observed data. Rather, a researcher must carefully consider the substantive knowledge available, together with the scientific goal of the analysis.

6.5 Case III: Factor Analysis

In the introduction to this chapter, we referred to reification as a typical consequence of naively using methods that combine data with external information, through unobservables. A very early context in this respect was the debate regarding general intelligence, based on different but, in terms of the data alone, indistinguishable forms of factor analysis. We now consider the factor-analytic case from a technical perspective.

We consider the following factor-analytic model:

$$Y_j - \mu_j = \sum_{m=1}^{k} \ell_{jm} F_m + \varepsilon_j, \tag{6.34}$$

$(j = 1, \ldots, p)$, where Y_j is a continuous response variable, with mean μ_j. The variable F_j is a latent continuous variable, called factor, and ε_j are errors. The coefficients ℓ_{jm} are called factor loadings. In matrix notation, the model is $\boldsymbol{Y} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon}$. In line with convention, we make the following assumptions: $E(\boldsymbol{F}) = \boldsymbol{0}$, $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, $\text{cov}(\boldsymbol{F}) = \boldsymbol{I}$ (the assumption of uncorrelated factors), and $\text{cov}(\boldsymbol{\varepsilon}, \boldsymbol{F}) = \boldsymbol{0}$. We make the distributional assumptions that $\boldsymbol{F}$ has a multivariate normal distributional with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{I}_k$, and that $\boldsymbol{\varepsilon}$ also has a multivariate normal distribution, with mean vector $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Psi}$. The response vector for an individual

is enriched with a vector of factors. We therefore first set out all the components in (6.1)–(6.2) for model (6.34). The prior distribution is:

$$\boldsymbol{F} \sim N(\boldsymbol{0}, \boldsymbol{I}_k). \tag{6.35}$$

The marginal distribution is readily shown to be:

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{LL'} + \boldsymbol{\Psi}). \tag{6.36}$$

The joint distribution, when joint normality of $\boldsymbol{Y}$ and $\boldsymbol{F}$ is assumed, can also be represented as: $(\boldsymbol{Y'}, \boldsymbol{F'})'$ with mean vector $(\boldsymbol{\mu'}, \boldsymbol{0'})'$ and covariance matrix

$$\begin{pmatrix} \boldsymbol{LL'} + \boldsymbol{\Psi} & \boldsymbol{L} \\ \boldsymbol{L'} & \boldsymbol{I}_k \end{pmatrix}.$$

Finally, by the conditional distribution property of subsets of multivariate normal distributions, the hierarchical and the posterior distributions, respectively, are

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{F} &\sim N[\boldsymbol{\mu} + (\boldsymbol{LL'} + \boldsymbol{\Psi})\boldsymbol{I}_k(\boldsymbol{f} - \boldsymbol{\mu}_f), \boldsymbol{\Psi}], & (6.37)\\ \boldsymbol{F}|\boldsymbol{Y} &\sim N[\boldsymbol{L}(\boldsymbol{LL'} + \boldsymbol{\Psi})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \boldsymbol{I}_k - \boldsymbol{L}(\boldsymbol{LL'} + \boldsymbol{\Psi})^{-1}\boldsymbol{L}]. & (6.38) \end{aligned}$$

The mean of the posterior distribution provides the predictive distribution of the enrichment, given the data. This is ordinarily used in the estimation of factor scores, where the vector of factor scores for the i^{th} individual, $i = 1, \ldots, n$, is given by $\widehat{\boldsymbol{f}_i} = \widehat{\boldsymbol{L}}(\widehat{\boldsymbol{L}}\widehat{\boldsymbol{L'}} + \widehat{\boldsymbol{\Psi}})^{-1}(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_y)$.

To illustrate arbitrariness of the posterior distribution, and the attendant consequences, we retain the marginal model (6.36) whilst replacing the posterior distribution (6.38). First note that $\boldsymbol{L}$, in the posterior is a $p \times k$ matrix. A particular way, therefore, to change the posterior distribution, is to change $\boldsymbol{L}$ to $\boldsymbol{L}_1$, with $\boldsymbol{L}_1$ being a $p \times k'$ matrix, $k \neq k'$. The new posterior, therefore, becomes

$$\boldsymbol{G}|\boldsymbol{Y} \sim N[\boldsymbol{L}_1(\boldsymbol{L}_1\boldsymbol{L}_1' + \boldsymbol{\Psi})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \mathbf{I}_{k_1} - \boldsymbol{L}_1(\boldsymbol{L}_1\boldsymbol{L}_1' + \boldsymbol{\Psi})^{-1}\boldsymbol{L}_1]. \tag{6.39}$$

This corresponds to the posterior of a factor-analytic model with a different number of factors than the initial model (6.34). The new joint model follows as the product of (6.36) and (6.39):

$$\begin{aligned} f(\boldsymbol{Y}, \boldsymbol{G}) &= 2\pi^{-(\frac{p+k_1}{2})}|\boldsymbol{\Sigma}_1|^{\frac{-1}{2}}|\boldsymbol{\Sigma}_2|^{\frac{-1}{2}} \\ &\quad \times e^{\frac{-1}{2}\left[(\boldsymbol{y}-\boldsymbol{\mu}_y)'|\boldsymbol{\Sigma}_1|^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_y) + (\boldsymbol{g}-\boldsymbol{\mu}_{g|y})'|\boldsymbol{\Sigma}_2|^{-1}(\boldsymbol{g}-\boldsymbol{\mu}_{g|y})\right]}, \end{aligned} \tag{6.40}$$

where $\boldsymbol{\Sigma}_1 = \text{cov}(\boldsymbol{Y})$, $\boldsymbol{\Sigma}_2 = \text{cov}(\boldsymbol{g}|\boldsymbol{Y})$, and $\boldsymbol{\mu}_{g|y} = E(\boldsymbol{g}|\boldsymbol{Y})$, components which are all described above. The new prior distribution follows as

$$\int_y 2\pi^{-\left(\frac{p+k_1}{2}\right)} |\boldsymbol{\Sigma}_1|^{\frac{-1}{2}} |\boldsymbol{\Sigma}_2|^{\frac{-1}{2}} \times e^{\frac{-1}{2}\left[(\boldsymbol{y}-\boldsymbol{\mu}_y)'|\boldsymbol{\Sigma}_1|^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_y)+(\boldsymbol{g}-\boldsymbol{\mu}_{g|y})'|\boldsymbol{\Sigma}_2|^{-1}(\boldsymbol{g}-\boldsymbol{\mu}_{g|y})\right]} d\boldsymbol{y}. \tag{6.41}$$

The new hierarchical model is therefore

$$\frac{2\pi^{-\left(\frac{p+k_1}{2}\right)} |\boldsymbol{\Sigma}_1|^{\frac{-1}{2}} |\boldsymbol{\Sigma}_2|^{\frac{-1}{2}}}{2\pi^{-\left(\frac{p+k_1}{2}\right)} |\boldsymbol{\Sigma}_1|^{\frac{-1}{2}} |\boldsymbol{\Sigma}_2|^{\frac{-1}{2}}} \times \frac{e^{\frac{-1}{2}\left[(\boldsymbol{y}-\boldsymbol{\mu}_y)'|\boldsymbol{\Sigma}_1|^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_y)+(\boldsymbol{g}-\boldsymbol{\mu}_{g|y})'|\boldsymbol{\Sigma}_2|^{-1}(\boldsymbol{g}-\boldsymbol{\mu}_{g|y})\right]}}{\int_y e^{\frac{-1}{2}\left[(\boldsymbol{y}-\boldsymbol{\mu}_y)'|\boldsymbol{\Sigma}_1|^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_y)+(\boldsymbol{g}-\mu_{g|y})'|\boldsymbol{\Sigma}_2|^{-1}(\boldsymbol{g}-\boldsymbol{\mu}_{g|y})\right]} d\boldsymbol{y}.} \tag{6.42}$$

We note that (6.41) and (6.42) complete a new hierarchical formulation, which produces the same marginal model, hence with the same marginal fit, as that from the initial formulation composed of (6.35) and (6.37). At the same time, however, important inferences ensuing from the two formulations are totally different. In particular, the estimation of factor scores, which uses the predictive distribution of the enrichment given the data, will be sensitive to the full model formulation. From the $(1 \times k)$ vectors of factor scores for the respondents, which would be the outcome of the initial formulation, we move to very different $(1 \times k')$ sets of factor scores, resulting from the new formulation. In addition, whereas ranking of individual respondents would be with respect to k components in the first formulation, it would be with respect to an arbitrary k', in the second. Vindication of any one formulation can only come through independent substantive information. Our illustration reiterates the fact that there is a completely unidentifiable part of the model. Therefore, we can view Gould's argument expressed through indeterminacy of the axis rotation, in the same way as the arbitrariness of the posterior in our enrichment terms. We emphasize that the enrichment view merely presents the well-known result about indeterminacy of factor rotation in a broader framework, underlining the commonality with other data-enrichment settings. Thus, also here, it follows that a good working knowledge of the difference between what can be learned from the data and what is identifiable through assumptions only, is a necessary part of the appropriate use of factor analysis.

6.5.1 Data Analysis

We analyze the track record data, described in Section 2.6, converting the record times into speed (in metres/second). We fit a 2-factor analysis model, using the

maximum likelihood method, as implemented in SAS Version 9.2. In Table 6.2, the factor-loading pattern for the rotated solution is presented.

Table 6.2: *National Track Records for Women: Factor analysis.*

Distance	Factor 1	Factor 2
100 metres	0.4406	0.8376
200 metres	0.4352	0.8908
400 metres	0.4116	0.8164
800 metres	0.7266	0.5673
1500 metres	0.8592	0.4822
3000 metres	0.9138	0.3859
Marathon	0.7654	0.3888

Factor 1 loads rather highly on the distances from 800 metres to the marathon, while factor 2 loads highly on the distances from 100 metres to 400 metres. We may therefore deem factor 1 to represent the middle and long-distance events, with factor 2 representing the short-distance events. We now turn to the factor scores, and, indeed, consider ranking of the countries involved, based on the respective factors. The U.S.A., Germany, the Czech Republic, France, and Russia complete the list of the top 5 countries with respect to the short-distance factor, while Kenya, Ireland, China, North Korea, and Norway top the middle-and-long-distance factor. Note that, for each country, there is a 1×2 vector of factor scores. As described earlier, by arbitrarily replacing the number of columns in the matrix $\boldsymbol{L}$, in the posterior, from 2 to, say, 4, and, of course, leaving the marginal model unaltered, we would end up with, for each country, a 1×4 vector of factor scores. We also note that in that case, we would also have complete freedom to arbitrarily specify the parameters in the now 7×4 matrix $\boldsymbol{L}_1$, in the new posterior, because the data carry no information about them. In view of these developments, the ranking that we initially conducted, based on the 2 factors, would completely change. Indeed, we would now be considering ranking based on completely different 1×4 vectors for the countries. Neither formulation is self-evidently appropriate and only independent substantive information can allow us to distinguish between them.

Thus, rather than extracting additional insight out of the data, our analysis shows that one has to be very aware of the arbitrary nature of at least some part of the conclusions.

6.6 Case IV: Frailty Models for Repeated Survival Outcomes

Repeated survival data are frequently modeled using so-called frailty models (Duchateau and Janssen, 2008), which are random effects models with, typically, random effects drawn from distributions other than the normal. A common choice is the gamma, combined with a Weibull model for the outcomes.

While such models are now well established, there are non-trivial implications for their use. For example, Molenberghs and Verbeke (2011) showed that the marginal distribution, generated from a Weibull-gamma frailty model, is of log-logistic type and only has a finite number of finite moments. There are examples where not even the second and first moments would be finite. However, this is an issue that takes us beyond the arbitrariness described for the linear mixed model case (Molenberghs *et al.*, 2012), an analogy of which for the Weibull-gamma case will be described next.

The term "frailty", and its use in survival data, has its roots in gerontology. In the latter field, it is used to indicate the increased mortality and morbidity risks of the more frail patients; in line with natural history, it is expected to increase with age. In statistics, it is taken to be constant within a patient in general statistical modeling and rather describe heterogeneity between patients. The introduction of random effects in survival data modelling dates back to Beard (1959), who, in modeling mortality, introduced the random effect in a univariate setting, and called it the "longevity factor." Vaupel *et al.* (1979) on the other hand, in attempting to allow individual differences in mortality hazard rates, introduced the random effect and termed it "frailty." In illustrating our general result on the arbitrariness of the posterior in frailty models, we focus on the parametric proportional hazards Weibull-Gamma frailty model:

$$h_{ij}(t|u) = h_0(t)u_i \exp(\boldsymbol{X}_{ij}^t\boldsymbol{\beta}), \tag{6.43}$$

where $h_{ij}(t|u)$ is the hazard of the j^{th} individual from the i^{th} cluster, $h_0(t) = \lambda\rho t^{\rho-1}, \lambda > 0, \rho > 0$, $\boldsymbol{X}_{ij}^t$ is the covariates' vector, $\boldsymbol{\beta}$ is the fixed effects vector, and u_i is the frailty for cluster i. The frailty distribution is gamma, which, in this context, is normally taken such that its mean equals one and hence the one-parameter gamma distribution is used:

$$f(u) = \frac{\alpha^\alpha}{\Gamma(\alpha)}u^{\alpha-1}e^{-\alpha u}. \tag{6.44}$$

We now spell-out, for (6.43), all components in (6.1) and (6.2). The prior distribution, $f(u)$, is, of course, (6.44). For the hierarchical distribution, using the fact that $f_{ij}(t) =$

$h_{ij}(t)S_{ij}(t)$, where

$$S_{ij}(t) = \exp\left(-\int_0^t \lambda\rho s^{\rho-1}u_i \exp(\boldsymbol{X}_{ij}^t\boldsymbol{\beta})ds\right),$$

it follows that the event times, given the frailty, are Weibull distributed with parameters $\lambda u_i \exp(\boldsymbol{X}_{ij}^t\boldsymbol{\beta})$ and ρ. It follows that the hierarchical distribution is

$$f(t|u) = \lambda\rho t^{\rho-1}u \exp\left[\boldsymbol{X}_{ij}^t\boldsymbol{\beta}) \exp(-\lambda u t^\rho \exp(\boldsymbol{X}_{ij}^t\boldsymbol{\beta})\right]. \tag{6.45}$$

The marginal distribution, $f(t)$, given as $\int_u f(t|u)f(u)du$, is easily shown to be

$$\frac{\lambda \exp(\boldsymbol{X}_{ij}^t\boldsymbol{\beta})\rho t^{\rho-1}\alpha^{\alpha+1}}{[\alpha + \lambda \exp(\boldsymbol{X}_{ij}^t\boldsymbol{\beta})t^\rho]^{\alpha+1}}. \tag{6.46}$$

The posterior distribution, $f(u|t)$, follows as the product of (6.45) and (6.44), divided by (6.46). Evaluating this gives the posterior as

$$\text{Gamma}\left(\alpha + 1, \frac{1}{\zeta_\rho + \alpha}\right), \tag{6.47}$$

where $\zeta_\rho \equiv \lambda \exp(\boldsymbol{X}_{ij}^t\boldsymbol{\beta})t^\rho$. This implies that

$$E(u|t) = \frac{\alpha + 1}{\zeta_\rho + \alpha} = \widehat{u_i}. \tag{6.48}$$

Prior to arbitrarily replacing (6.47), we derive some further quantities, which are of particular relevance in survival data settings. The population survival function, $S_f(t)$, corresponding to (6.43), is evaluated as $\int_0^\infty S_{ij}(t)f_u(u)du$, giving $[(1 + \zeta_\rho\alpha^{-1})^\alpha]^{-1}$. This implies that the population hazard function is

$$h_f(t) = \frac{\zeta_{\rho-1}\rho}{1 + \zeta_\rho\alpha^{-1}}. \tag{6.49}$$

We now return to our arbitrary replacement of the posterior. Specifically, we replace the Gamma posterior (6.47) with a normal posterior with mean μ and variance σ^2, for which we choose $\sigma^2 = 1$ and $\mu = \mu(t) = \varphi\boldsymbol{X}_{ij}^t t$, $\boldsymbol{X}_{ij}$ being as defined earlier. We note here that the normal posterior implies that

$$E(x|t) = \varphi\boldsymbol{X}_{ij}^t t = \widehat{x_i}. \tag{6.50}$$

The new joint model follows as:

$$\frac{\zeta_{\rho-1}\rho\alpha^{\alpha+1}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}{[\alpha + \zeta_\rho]^{\alpha+1}\sqrt{2\pi}}. \tag{6.51}$$

The new prior distribution follows as

$$\int_0^\infty \frac{\zeta_{\rho-1}\rho\alpha^{\alpha+1}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}{[\alpha+\zeta_\rho]^{\alpha+1}\sqrt{2\pi}}dt. \tag{6.52}$$

Hence, the new hierarchical model is

$$\frac{\frac{\zeta_{\rho-1}\rho\alpha^{\alpha+1}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}{[\alpha+\zeta_\rho]^{\alpha+1}\sqrt{2\pi}}}{\int_0^\infty \frac{\zeta_{\rho-1}\rho\alpha^{\alpha+1}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}{[\alpha+\zeta_\rho]^{\alpha+1}\sqrt{2\pi}}dt}. \tag{6.53}$$

We note that (6.53) and (6.52) complete a new formulation, which, as with the initial formulation consisting of (6.45) and (6.44), defines the same marginal distribution for the event times, given by (6.46). Thus our new, admittedly contrived model is indistinguishable from the original model in terms of fit to the data. Like in the other enrichment settings, though, the prediction of h_{ij} (the conditional hazard), through the mean of the hierarchical model, is different from what it was before. This difference will have an impact on inferences, without ability for the data to testify whether this or the original formulation is better or worse.

6.6.1 Data Analysis

6.6.1.1 Data on Recurrent Asthma Attacks in Children

We now illustrate our results using the recurrent asthmatic attacks in children' data, introduced in Section 2.3. Though various time-representations exist to analyze data of this type, we hereby assume that interest is on the event rate in calendar time, leading to the model given below. Furthermore, for purposes of illustration, we restrict ourselves to risk times which culminate in an event (asthmatic attack), *i.e.*, we only consider that subset of data not consisting of censored observations (we only consider data points for which the corresponding "end-of-observation" period corresponds to an attack). Consider the following model

$$h_{ij}(t|u) = \begin{cases} h_0(t)u\exp(X_i^t\beta) & \text{if } \mathrm{y}_{ij1} \leq \mathrm{t} \leq \mathrm{y}_{ij2}; \\ 0 & \text{otherwise,} \end{cases} \tag{6.54}$$

where $u \sim \text{Gamma}(\alpha, 1/\alpha)$ and h_{ij} denotes the hazard for the i^{th} child, time j. We note that in the model specification above, the subscript j has been dropped in denoting the drug covariate, X, since a given child is under either study drug or placebo at all time points. Further, β is the parameter corresponding to the drug effect, (y_{ij1}, y_{ij2}), $j = 1, \ldots, n_i$, denotes the pairs corresponding to the beginning and

end of each risk period for child j, and t is the time since entry into the trial. We optimize the marginal likelihood using the R 2.11.1 software. Following Duchateau and Janssen[20], we convert time from days to months, to avoid convergence issues arising when λ is too small. Parameter estimates obtained are $\widehat{\lambda}$=0.2306 (s.e. 0.0234), $\widehat{\rho}$=1.2576 (s.e. 0.0309), $\widehat{\beta}$=-0.0159 (s.e. 0.0749) and $\widehat{\theta}$= 0.1606 (s.e. 0.0290). We now partially replace the model defined above, by retaining its resultant marginal model, and coupling it with a normal posterior. We set $\varphi = 0.5$, which is required because the data contain no information about this parameter. We then consider predictions for the conditional hazard under the two model formulations. In Figure 6.1, we present, for the study drug and placebo, the population and conditional hazard functions, for the models composed of marginal model (6.46) with each of the two different posterior specifications.

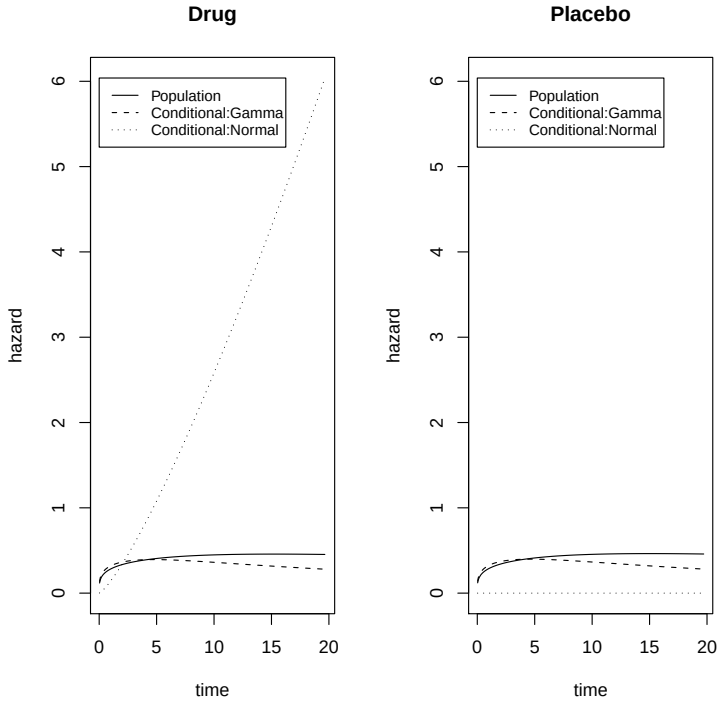


Figure 6.1: *Recurrent Asthma Data: Population and conditional hazard functions.*

We note that the two formulations give totally different predictions for the conditional hazard. The Gamma choice produces a prediction which lies much closer to the population hazard than the normal choice, which, clearly, produces a prediction

which is very different. These disparate inferences occur in disturbing conjunction with an unaltered marginal model, but is in line with all other illustrations in this chapter and the general result spelled out in Section 6.2.

6.6.1.2 Time-to-insemination Data

This data set was introduced in Section 2.5. For our purposes, we restrict attention to event times and only consider that subset of data not consisting of censored observations. For these data, we also consider (6.54), with now h_{ij} the hazard for the j^{th} cow in the i^{th} herd, X_{ij}^t the parity covariate, and β the corresponding parameter. We convert time to months. Optimization of the marginal likelihood is done in R 2.11.1 software. Due to computational challenges, we use, for our model fitting, at most 20 cows in a herd. Parameter estimates obtained are $\widehat{\lambda}$=0.0569 (s.e. 0.0035), $\widehat{\rho}$=2.538 (s.e. 0.082), $\widehat{\beta}$=-0.2210 (s.e. 0.0331) and $\widehat{\theta}$= 0.3248 (s.e. 0.0394). We now partially replace the model defined by (6.54), by retaining its resultant marginal model, and coupling it with a normal density. Again, we set $\varphi = 0.5$ and consider predictions for the conditional hazard under the two model formulations. In Figure 6.2, we present, for each parity category, the population and conditional hazard functions, for the models composed of marginal model (6.46) with each of the two different posterior specifications. Also here, we note that the two formulations give totally different predictions for the conditional hazard. The prediction based on the gamma choice is closer to the population hazard than is the case for the normal choice; this is similar to observations for the linear mixed model case (Molenberghs *et al.*, 2012).

6.7 Discussion

In this chapter, we have considered specific settings within enrichment, and brought out the common feature of unobservables, shared by all. The information required to identify such models is divided in that supplied by the data and that supplied externally, through assumptions and/or scientific knowledge. This implies that entire classes of models exist, coinciding in their description of the observed data, but different in their representation of the unobservables given the observed data.

For the data analyst, this means that every model in an enriched-data setting can be factored into a product of two components: the first one, termed the marginal model, fully identifiable from the observed data; the second one, the predictive distribution, entirely arbitrary. As a consequence, the conventional modeling route, consisting of formulating a model and judging its quality based on goodness-of-fit

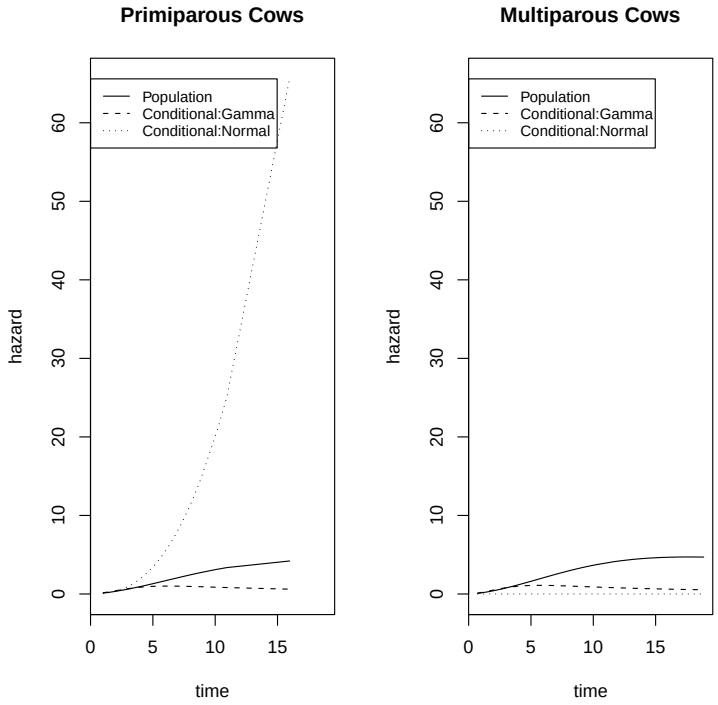


Figure 6.2: *Time-to-insemination Data: Population and conditional hazard functions.*

alone, is inadequate. This is because the inferences drawn and the assumptions made about the unobservables cannot be divorced. As we see it, there are therefore two alternative modes of analysis, that nevertheless fully exploit the information contained in the empirical data. In the first one, non-verifiable assumptions are based on substantive knowledge and/or statistical design. In the second one, sensitivity of the inferences to the non-verifiable assumptions are assessed formally.

This common issue is pertinent in a range of seemingly disparate settings; not only incomplete-data and random-effects models (Verbeke and Molenberghs, 2010), but also frailty models, latent classes, latent variables, factor analysis, and mixture models (Molenberghs *et al.*, 2012). Of course, various of these are interconnected and can be placed under the general umbrella of structural equations modeling (Skrondal and Rabe-Hesketh, 2004). Bringing it out has some value, we believe, as there are instances in the literature where it is missed (Bollen, 1989).

We have not been exhaustive in our coverage of enrichment. Other areas include

censored survival data, which is very similar to the incomplete-data case, grouped data, and situations where more than one type of enrichment occur simultaneously, such as, for example, incomplete data in random-effects models. Another major omission is that of methodology for causal inference. The issues raised here have been widely discussed in the appropriate literature. For example, Pearl (2010) states: "Alternative causal models usually exist that make contradictory claims and, yet, possess identical statistical implications. Statistical test (sic) can be used for rejecting certain kernels, in the rare cases that such cases have testable implications, but the lion's share of supporting causal claims falls on the shoulders of untested causal assumptions." By kernel, Pearl refers broadly to a minimal set of assumptions required to identify the underlying causal model. Within the causal framework, we can include inferences drawn from randomized clinical trials (Rubin, 1974). In the incomplete-data setting sensitivity analysis is particularly well developed, for both the parametric and non-parametric settings (Molenberghs and Kenward, 2007; Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005; Beunckens *et al.*, 2009; Chickering and Pearl, 1997). For example, Creemers *et al.* (2011) showed how the generalized shared-parameter model for incomplete data can be used, not only to demonstrate that one cannot choose based on the data between MAR and MNAR, but also how it can be used as a vehicle for sensitivity analysis. Much work on sensitivity analysis can also be found in the causal literature (Daniels and Hogan, 2008; Greenland, 2005).

Chapter 7

General Conclusions and Future Research

In Chapter 3, we have outlined a broad and extended, hence flexible, joint modelling framework. Apart from the correlation induced by repeated measurements from the same subject and the association between the different outcomes, the often-restrictive mean-variance prescription in the model for the non-Gaussian outcome has explicitly been addressed. This has been done conveniently through the inclusion of conjugate random effects, an aspect that has been exploited to easily estimate the framework in standard software, through partial marginalization.

Through our analyses of data from the area of chronic heart failure, we have shown that our extended framework provides improvement to model fit, while still maintaining parsimony, and also offers better prediction. We have also observed impact on significance tests.

However, there is need for follow-up work, involving general assessments through simulation studies, on the effects of omitting the conjugate random effects. More specifically, the impact on specific model parameters, under various scenarios, will be investigated. These scenarios include the level of overdispersion in each of the non-Gaussian outcomes, the amount of censoring, as well as the length of the longitudinal sequence.

This framework also opens avenues for further research work in related areas. It is possible to derive the marginal joint correlation functions, which may be of interest, for example, in surrogate marker evaluation and in psychometrics.

We have explored how dynamic prediction can assist physicians make intervention decisions in telemonitored CHF patients, in Chapter 4. The dynamically updated con-

ditional survival probabilities, and their confidence intervals, can provide physicians with additional information on which to base such decisions. We have also explored how well each of the available biomarkers discriminate between patients who are and those who are not going to get rehospitalized. This approach does not only provide a sound statistical modelling approach to predicting rehospitalizations in telemonitored chronic heart failure patients, it also provides a practical solution in heart failure management.

However, we have only addressed dynamic prediction in relation to the time to first hospitalization. There is therefore need for methodological extension to cope with dynamic prediction in the context of recurrent events. Another limitation is that we have analyzed each biomarker separately. It would be important to consider them jointly, taking their association structure into account. More software implementation work is needed to allow dynamic prediction and calculation of accuracy measures when multiple longitudinal biomarkers are available.

In Chapter 5, we have used the strong connection between the longitudinal and time-to-event setting, and the missing data one, to build an extended shared random effects joint model, similar in spirit to that of Creemers *et al.* (2011) in the context of longitudinal data subject to missing observations, but now transposed to the current more complex setting. In doing so, we have taken a slightly different perspective on joint models than is prevalent in the literature.

Within the extended framework, we have provided a characterization of MAR, consistent with the one in the missing data setting. While the framework has been built conceptually, we have used the elaborate random-effects structure as an avenue for sensitivity analysis in this context. We have illustrated such an analysis using data from a study on liver cirrhosis.

Needless to say, the longitudinal and time-to-event settings are more complex than the missing data one, given the added levels of coarsening. It has been highlighted that model formulation under the framework studied here then becomes exceedingly complex, given the interrelationships that arise among model components. The effect of such difficulties on the studied framework, especially on the results established here, and the possibility of expanding the framework to take into account such difficulties, requires further investigation.

A sensitivity analysis has been conducted by formulating an extended model, next to a conventional one. It is still possible, within the elaborate random effects structure, to formulate additional alternative models, all making different assumptions. The stability of inferences across such models can then be considered. This provides a rich avenue for sensitivity analyses.

Finally, in Chapter 6, we have considered specific settings within enrichment, and brought out the common feature of unobservables, shared by all. The information required to identify such models is divided in that supplied by the data and that supplied externally, through assumptions and/or scientific knowledge. This implies that entire classes of models exist, coinciding in their description of the observed data, but different in their representation of the unobservables given the observed data.

For the data analyst, this means that every model in an enriched-data setting can be factored into a product of two components: the first one, termed the marginal model, fully identifiable from the observed data; the second one, the predictive distribution, entirely arbitrary. As a consequence, the conventional modeling route, consisting of formulating a model and judging its quality based on goodness-of-fit alone, is inadequate. This is because the inferences drawn and the assumptions made about the unobservables cannot be divorced.

We have not been exhaustive in our coverage of enrichment. Other areas include censored survival data, which is very similar to the incomplete-data case, grouped data, and situations where more than one type of enrichment occur simultaneously, such as, for example, incomplete data in random-effects models.

Bibliography

Agresti A. (2002). *Categorical Data Analysis*. 2nd ed. New York: John Wiley & Sons.

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

Antolini, F., Boracchi, P., Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine*, **24**, 3927–3944.

Beard, R.E. (1959). Note on some mathematical mortality models. In: *The lifespan of animals. Ciba colloquium on Aging,* G.E.W. Wolstenholme and M O'Connor (eds.). Little, Brown, Boston, 302–311.

Beunckens, C., Sotto, C., Molenberghs G., and Verbeke., G. (2009). A multifaceted sensitivity analysis of the Slovenian Public Opinion Survey data. *Applied Statistics* **58**, 171–196.

Bjørnstad, J.F. (1996). On the generalization of the likelihood function and likelihood principle. *Journal of the American Statistical Association*, **91**, 791–806.

Böhning, D. (2000). *Computer-Assisted Analysis of Mixtures and Applications: meta-analysis, disease mapping, and others*. Boca Raton: Chapman & Hall/CRC.

Bollen, K. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Carlin B.P. and Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall.

Chaudhry, S.I., Wang, Y., Concato, J., Gill, T.M., and Krumholz, H.M. (2007). Patterns of weight change preceding hospitalization for heart failure. *Circulation*, **116**, 1549–1554.

Chickering, D. and Pearl, J. (1997). A clinician's tool for analyzing non-compliance. *Computing Science and Statistics*, **29**, 424–431.

Chin, M.H. and Goldman, L. (1997). Correlates of Early Hospital Readmission or Death in Patients With Congestive Heart Failure. *American Journal of Cardiology*, **79**,1640–1644.

Collins, L.M, and Lanza, S.T (2009). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. New Jersey: Wiley.

Creemers, A., Hens, N., Aerts, M., Molenberghs, G., Verbeke, G., and Kenward, M.G. (2010). A sensitivity analysis for shared-parameter models for incomplete longitudinal outcomes. *Biometrical Journal*, **52**, 111—125.

Creemers, A., Hens, N., Aerts, M., Molenberghs, G., Verbeke, G., and Kenward, M.G. (2011). Generalized shared-parameter models and missingness at random. *Statistical modeling*, **11**, 279–311.

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

Dempster, A.P., Laird, N.M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Dendale, P., De Keulenaer, G., Troisfontaines, P., Weytjens, C., Mullens, W., Elegeert, I., Ector, B., Houbrechts, M., Willekens, K., and Hansen, D. (2011). Effect of a telemonitoring-facilitated collaboration between general practitioner and heart failure clinic on mortality and rehospitalization rates in severe heart failure: the TEMA-HF 1 (TElemonitoring in the MAnagement of Heart Failure) study. *European Journal of Heart Failure*, **14**, 333–340.

DeGruttola, V. and Tu, X.M. (1994). Modeling progression of CD4 lymphocyte count and its relationship to survival time. *Biometrics*, **50**, 1003–1014.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. New York: Springer.

Duchateau, L., Opsomer, G., Dewulf, J., and Janssen, P. (2005). The non-linear effect (determined by the penalised partial-likelihood approach) of milk-protein concentration on time to first insemination in Belgian dairy cows. *Preventive Veterinary Medicine*, **68**, 81–90.

Faucett, C.L. and Thomas, D.C. (1996). Simultaneously modeling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, **15**, 1663–1685.

Goodman, L.A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I-A Modified latent structure approach. *American Journal of Sociology*, **79**, 1179–1259.

Gould, J.S. (1981). *The Mismeasure of Man*. New York: W.W. Norton and Company.

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society, Series A*, **168**, 267–306.

Harrell, F., Kerry, L., Mark, D. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361–387.

Heagerty, P., and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, **61**, 92–105.

Heitjan, D.F. and Rubin, D.B. (1991). Ignorability and coarse data. *The Annals of Statistics*, **19**, 2244–2253.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modeling of longitudinal measurements and event time data. *Biostatistics*, **1**, 465–480.

Hogan, J.W. and Laird, N.M. (1998) Increasing efficiency from censored survival data by using random effects to model longitudinal covariates. *Statistical Methods in Medical Research*, **7**, 28–48.

Hogan, J.W. and Laird, N.M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, **16**, 239–258.

Johnson, R.A. and Wichern, D.W. (2007) *Applied Multivariate Statistical Analysis*. New Jersey: Pearson Education, Inc.

Klein, J.P, and Moeschberger, M.L.(1997) *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer.

Kurland, B.F. and Heagerty, P.J. (2005) Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by death. *Biostatistics*, **6**, 241–258.

Kurland, B.F., Johnson, L.L., Egleston, B.L., and Diehr, P.H. (2009) Longitudinal data with follow-up truncated by death: Match the analysis method to research aims. *Statistical Science*, **24**, 211–222.

Lavalley, M.P. and DeGruttola, V. (1996). Models for empirical Bayes estimators of longitudinal CD4 counts. *Statistics in Medicine*, **15**, 2289–2305.

Law, N.J., Taylor, J.M.G., and Sandler, H. (2002). The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics*, **3**, 547–563.

Lewin, J., Ledwidge, M., OLoughlin, C., McNally, C. *et al.* (2005). Clinical deterioration in established heart failure: What is the value of BNP and weight gain in aiding diagnosis? *European Journal of Heart Failure* **7**, 953–957.

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

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M.G. (2008). Every missing not at random model has got a missing at random counterpart with equal fit. *Journal of the Royal Statistical Society, Series B* **70**, 371–388

Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Studies.* Chichester: John Wiley & Sons.

Molenberghs, G., Njagi, E.N., Kenward, M.G., and Verbeke, G. (2012). Enriched-data problems and essential non-identifiability. *International Journal of Statistics in Medical Research*, **1**, 16–44.

Molenberghs, G., Michiels, B., Kenward, M.G., and Diggle, P.J. (1998). Monotone missing data and pattern-mixture models. *Statistica Neerlandica* **52**, 153–161.

Molenberghs, G. and Verbeke, G. (2004) Meaningful statistical model formulations. *Statistica Sinica*, **14**, 989–1020.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

Molenberghs, G. and Verbeke, G. (2011). On the Weibull-Gamma frailty model, its infinite moments, and its connection to generalized log-logistic, logistic, Cauchy, and extreme-value distributions. *Journal of Statistical Planning and Inference*, **141**,861–868.

Molenberghs, G., Verbeke, G., and Demétrio, C.G.B. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis*, **13**, 513–531.

Molenberghs, G., Verbeke, G., Demétrio, C.G.B., and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, **25**, 325–347.

Molenberghs, G., Verbeke, G., and Iddi, S. (2011). Pseudo-likelihood methodology for partitioned large and complex samples. *Statistics & Probability Letters*, **81**, 892–901.

Molenaar, P.C.M. A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, **2**, 201–218.

Molenaar, P.C.M. On the implications of the classical ergodic theorems: Analysis of developmental processes has to focus on intra-individual variation. *Developmental Psychobiology 2008*, **50**, 60–69.

Njagi, E.N., Molenberghs, G., Kenward, M.G., Verbeke, G., and Rizopoulus, D. (2013c). A Characterization of Missingness at Random in a Generalized Shared-parameter Joint modelling Framework for Longitudinal and Time-to-Event Data, and Sensitivity Analysis. *Biometrical Journal*. Under Revision.

Njagi, E.N., Molenberghs, G., Rizopoulus, D., Verbeke, G., Kenward, M.G., Dendale, P., and Willekens, K. (2013b). A flexible joint-modeling framework for longitudinal and time-to-event data with overdispersion. *Statistical Methods in Medical Research.* Published online before print July 18, 2013, doi: 10.1177/0962280213495994.

Njagi, E. N., Rizopoulos, D., Molenberghs, G., Dendale, P., and Willekens, K. (2013a). A joint survival-longitudinal modelling approach for the dynamic prediction of re-hospitalization in telemonitored chronic heart failure patients. Statistical Modelling, **13**, 179–198

Pawitan, Y. and Self, S. (1993). Modeling disease marker processes in AIDS. *Journal of the American Statistical Association*, **88**, 719–726.

Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics*, **6**, article 7, 1–58.

Rizopoulos D. (2011). Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-event Data. *Biometrics*, **67**, 819–829.

Rizopoulos, D. (2012a). *Joint Models for Longitudinal and Time-to-Event Data*. Boca Raton: Chapman and Hall/CRC.

Rizopoulos, D. (2012b). Package "JM": R package version 1.0-0, URL http://cran.r-project.org/web/packages/JM/JM.pdf.

Rizopoulos, D., Verbeke, G. and Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society, Series B*, **71**, 637–654.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, **66**, 688–701.

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

Simar, L. (1976). Maximum likelihood estimation of a compound Poisson process. *The Annals of Statistics*, **4**, 1200–1209.

Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling.* London: Chapman & Hall/CRC.

Tatsuoka, K.K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glazer, A. Lesgold, & M.G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Lawrence Erlbaum Associates.

Taylor, J.M.G., Cumberland, W.G., and Sy, J.P. (1994). A stochastic model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association*, **89**, 727–736.

Thyrion, P (1960). Contribution à létude du bonus pour non sinistre en assurence automobile. *ASTIN Bull*, **1**: 142–162.

Tsiatis, A.A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, **72**, 20–22.

Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, **14**, 809–834.

Tsiatis, A.A., DeGruttola, V., and 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**, 27–37.

Vaupel, J.W., Manton, K.G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**, 439–454.

Verbeke, G. and Molenberghs, G. (2010). Arbitrariness of models for augmented and coarse data, with emphasis on incomplete-data and random-effects models. *Statistical Modelling*, **10**, 391–419.

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

Verbeke, G., Molenberghs, G., and Rizopoulos, D. (2010). Random effects models for longitudinal data. *Longitudinal Research with Latent Variables*. K. van Montfort, H. Oud, and Al Satorra (Eds.). New York: Springer, pp. 37–96.

Xu, H. and Craig, B.A. (2009). A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. *Biometrics*, **65**, 1145–1155.

Xu, J. and Zeger, S.L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics*, **50**, 375–387.

Yu, M., Law, N.J., Taylor, J.M.G., and Sandler, H.M. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*, **14**, 835–862.

Yu, M., Taylor, J.M.G., and Sandler, H.M. (2008). Individual Prediction in Prostate Cancer Studies Using a Joint Longitudinal SurvivalCure Model. *Journal of the American Statistical Association*, **103**, 178–187.

Zhang, J., Goode, K.M., Cuddihy, P.E., and Cleveland, J.G. (2009). Predicting hospitalization due to worsening heart failure using daily weight measurement: analysis of the Trans- European Network-Home-Care Management System (TEN-HMS) study. *European Journal of Heart Failure*, **11**, 420–427.

Zhang, J. and Heitjan, D.F. (2007). Impact of nonignorable coarsening on Bayesian inference. *Biostatistics*, **8**, 722–743.

Zheng, Y., Heagerty, P. (2007). Prospective Accuracy for Longitudinal Markers. *Biometrics*, **63**, 332–341.

Appendix A

A Flexible Joint Modeling Framework for Longitudinal and Time-to-event Data With Overdispersion

In this appendix, we highlight the derivation of the fully marginalized joint model probabilities for Case 1 and Case 3 of Section 3.3. We also provide the analysis program for the case study analyzed in Section 3.5 of the same chapter.

A.1 Derivations for the Joint Marginal Probabilities

A.1.1 Case 1: Repeated Time-to-event and Repeated Continuous Outcomes

Consider the conditional joint distribution given by (3.9). The following expansion

$$e^{-\lambda_k t_{ik}{}^{\rho_k}\psi_{ik}e^{\mu_{ik}+d_{ik}}} = \sum_{\ell_k=0}^{\infty} \frac{(-1)^{\ell_k}}{\ell_k!}\lambda_k{}^{\ell_k}\psi_{ik}^{\ell_k}t_{ik}^{\ell_k\rho_k}e^{\ell_k\mu_{ik}}e^{\ell_k d_{ik}} \tag{A.1}$$

facilitates the integration process and, integrating over the normal random effects, the joint distribution, conditional on the gamma random effects, takes the form:

$$\begin{aligned} & f(\boldsymbol{t}_i, \boldsymbol{y}_i|\boldsymbol{\psi}_i) \\ &= \frac{1}{(2\pi)^{\frac{n_i}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{\frac{-1}{2}(\boldsymbol{y}_i - X_i\boldsymbol{\beta})'\Sigma_i^{-1}(\boldsymbol{y}_i - X_i\boldsymbol{\beta})} \\ &\quad \times \sum_{\boldsymbol{\ell}} \frac{(-1)^{\sum_k \ell_k}}{\prod_k \ell_k!} e^{\sum_k(\ell_k+1)\mu_{ik}} \left(\prod_k {\psi_{ik}}^{\ell_k+1} \lambda_k^{\ell_k+1} \rho_k t_{ik}^{\rho_k\ell_k+\rho_k-1} \right) \times a_{\ell k} \\ &\quad \times \frac{1}{|D|^{\frac{1}{2}}} \times \left|(D^{-1} + Z_i'\Sigma_i Z_i)^{-1}\right|^{\frac{1}{2}}, \end{aligned} \tag{A.2}$$

where

$$a_{\ell k} = e^{\frac{1}{2}\left[\sum_k(\ell_k+1){\boldsymbol{w}_{ik}}' - 2(\boldsymbol{y}_i - X_i\boldsymbol{\beta})'\Sigma_i^{-1}Z_i\right]\Pi\left[\sum_k(\ell_k+1){\boldsymbol{w}_{ik}}' - 2(\boldsymbol{y}_i - X_i\boldsymbol{\beta})'\Sigma_i^{-1}Z_i\right]'},$$

with

$$\Pi = \left[D^{-1} + {Z_i}'\Sigma_i Z_i\right]^{-1}.$$

Integrating-out the gamma random effects, we get

$$\begin{aligned} & f(\boldsymbol{t}_i, \boldsymbol{y}_i) \\ &= \frac{1}{(2\pi)^{\frac{n_i}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{\frac{-1}{2}(\boldsymbol{y}_i - X_i\boldsymbol{\beta})'\Sigma_i^{-1}(\boldsymbol{y}_i - X_i\boldsymbol{\beta})} \\ &\quad \times \sum_{\boldsymbol{\ell}} \frac{(-1)^{\sum_k \ell_k}}{\prod_k \ell_k!} e^{\sum_k(\ell_k+1)\mu_{ik}} \left(\prod_k \lambda_k^{\ell_k+1} \rho_k t_{ik}^{\rho_k\ell_k+\rho_k-1} \right) \times a_{\ell k} \times b_{\ell k} \\ &\quad \times \frac{1}{|D|^{\frac{1}{2}}} \times \left|(D^{-1} + Z_i'\Sigma_i Z_i)^{-1}\right|^{\frac{1}{2}}, \end{aligned} \tag{A.3}$$

where $b_{\ell k} = \left\{[\beta^{\ell_k - 1}(\ell_k + \alpha)(\ell_k + \alpha - 1)\dots(\alpha)]^{p_i}\right\}$, and $a_{\ell k}$ is as defined above. Further, summation over $\boldsymbol{\ell}$ is shorthand for multi-index summation over all components of $(\ell_1, \ell_2, \dots, \ell_{p_i})$.

A.1.2 Case 3: Repeated Time-to-event and Repeated Count Outcomes

Consider the conditional joint distribution given in (3.13). We make use of the expansion

$$e^{-\theta_{ij}e^{\tau_{ij}} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i} = \sum_{m_j=0}^{\infty} \frac{(-1)^{m_j}}{m_j!} {\theta_{ij}}^{m_j} e^{m_j\tau_{ij}} e^{m_j\boldsymbol{z}_{ij}'\boldsymbol{b}_i}$$

and of expansion (A.1), and after integrating-out the normal random effects, the joint distribution, conditional on the two sets of gamma random effects, is

$$\begin{aligned} & P(\boldsymbol{Y}_i = \boldsymbol{y}_i, \boldsymbol{T}_i = \boldsymbol{t}_i | \boldsymbol{\Theta}_i, \boldsymbol{\Psi}_i) \\ = & \left(\prod_{j=1}^{n_i} \frac{1}{y_{ij}!} \right) \sum_{\boldsymbol{m}} \sum_{\boldsymbol{l}} (-1)^{\sum_j m_j + \sum_k \ell_k} \left(\prod_j \theta_{ij}^{m_j + y_{ij}} \right) \left(\prod_k \Psi_{ik}^{\ell_k + 1} \right) \\ & \times \prod_k \left(\lambda_k^{\ell_k + 1} \rho_k t_{ik}^{\rho_k \ell_k + \rho - 1} \right) \times e^{\left[\sum_j (m_j + y_{ij}) \tau_{ij} + \sum_k (\ell_k + 1) \mu_{ik} \right]} \\ & \times g_{m_j \ell_k}, \end{aligned} \tag{A.4}$$

where

$$g_{m_j \ell_k} = e^{\frac{1}{2} \left[\sum_j (m_j + y_{ij}) \boldsymbol{z}'_{ij} + \sum_k (\ell_k + 1) \boldsymbol{w}'_{ik} \right] D \left[\sum_j (m_j + y_{ij}) \boldsymbol{z}_{ij} + \sum_k (\ell_k + 1) \boldsymbol{w}_{ik} \right]}.$$

Upon integrating-out the gamma random effects, we obtain the marginal joint distribution as

$$\begin{aligned} & P(\boldsymbol{Y}_i = \boldsymbol{y}_i, \boldsymbol{T}_i = \boldsymbol{t}_i) \\ = & \prod_{j=1}^{n_i} \frac{1}{y_{ij}!} \sum_{\boldsymbol{m}} \sum_{\boldsymbol{l}} (-1)^{\sum_j m_j + \sum_k \ell_k} \\ & \times \left[\prod_j \frac{\beta_j^{m_j + y_{ij}} \Gamma(\alpha_j + m_j + y_{ij})}{\Gamma(\alpha_j)} \right] \left[\prod_k \frac{\beta_k^{\ell_k + 1} \Gamma(\alpha_k + \ell_k + 1)}{\Gamma(\alpha_k)} \right] \\ & \times \prod_k \left(\lambda_k^{\ell_k + 1} \rho_k t_{ik}^{\rho_k \ell_k + \rho - 1} \right) \times e^{\left[\sum_j (m_j + y_{ij}) \tau_{ij} + \sum_k (\ell_k + 1) \mu_{ik} \right]} \\ & \times g_{m_j \ell_k}, \end{aligned} \tag{A.5}$$

with $g_{m_j \ell_k}$ as defined above.

A.2 Analysis program

Our analysis program is provided below. Notice that the variable "outcome" brings together a patient's measurements on both outcomes, with the variable "type" distinguishing between survival and count outcomes, in order to correspondingly specify the respective likelihood contributions.

```
proc nlmixed data=chf;
bounds rho>0, alpha>0, lambda>0, alpha2>0;
parms zeta0=-3.5165 zeta1=1.0288  xi=0.076599190
lambda=0.003375627 rho=0.809716599;
```

```
offset=1;
if type='outc1' then do;
eta=offset*logtime+zeta0+zeta1*(lvef=1)+b;
loglik=count*eta-lgamma(count+1)+lgamma(count+alpha2)
  -lgamma(alpha2)+alpha2*log(alpha2)
  -(count+alpha2)*log(exp(eta)+alpha2);
end;
else if type='outc2' then do;
loglik=censor*log(lambda*rho*(day**(rho-1))*exp(xi*(lvef=1)+scalef*b))
  +lgamma(alpha+censor)-log(((lambda*(rend**rho-rstart**rho)*
  exp(xi*(lvef=1)+scalef*b)+alpha)**(alpha+censor))*((1/alpha)**alpha))
 -lgamma(alpha);
end;
model outcome ~ general(loglik);
random b ~ normal(0, sigmab**2) subject=ptid;
estimate 'Variance of Normal R.E.s' sigmab**2;
estimate 'Variance of Gamma R.E.s Survival' 1/alpha;
estimate 'Variance of Gamma R.E.s Count' 1/alpha2;
contrast 'Joint effect hypothesis' xi,zeta1;
run;
```

Appendix B

A Joint Survival-Longitudinal Modelling Approach for the Dynamic Prediction of Rehospitalization in Telemonitored Chronic Heart Failure Patients

In this appendix, we present the dynamic discrimination results for systolic blood pressure, and weight, discussed in Section 4.3.3.

B.1 Systolic Blood Pressure

Table B.1: *Systolic Blood Pressure. DDIs.*

	Time window Δt	DDI
First Step Model	2	0.6009
	4	0.6048
	8	0.6760
	16	0.6145
Second Step Model		
NTproBNP	2	0.4917
	4	0.4995
	8	0.5712
	16	0.5223
Heart Rhythm	2	0.5844
	4	0.6038
	8	0.6551
	16	0.6064
NYHA	2	0.5585
	4	0.5585
	8	0.6561
	16	0.5869
Sex	2	0.5977
	4	0.6051
	8	0.6597
	16	0.6304
LVEF	2	0.4823
	4	0.4863
	8	0.5839
	16	0.5749
Age	2	0.6453
	4	0.6453
	8	0.6771
	16	0.6748

B.2 Weight

Table B.2: *Weight. DDIs.*

	Time window Δt	DDI
First Step Model	2	0.3877
	4	0.3877
	8	0.4648
	16	0.5020
Second Step Model		
NTproBNP	2	0.5885
	4	0.5885
	8	0.6392
	16	0.5388
Heart Rhythm	2	0.5433
	4	0.5515
	8	0.5747
	16	0.5199
NYHA	2	0.5227
	4	0.5304
	8	0.5598
	16	0.5044
Sex	2	0.5317
	4	0.5317
	8	0.5434
	16	0.5047
LVEF	2	0.4565
	4	0.4605
	8	0.5085
	16	0.4685
Age	2	0.6634
	4	0.6634
	8	0.6434
	16	0.5978

Appendix C

A Characterization of Missingness at Random in a Generalized Shared-parameter Joint Modelling Framework for Longitudinal and Time-to-Event Data, and Sensitivity Analysis

In this appendix, we present our program for fitting the extended model described in Section 5.5.1. To use the program, the data need to be restructured accordingly.

```
proc nlmixed data=LongSurv;
bounds lambda1>0, rho1>0, lambda2>0, rho2>0;
parms Beta0=0.4349 Beta1=0.1165 Beta2=-0.04118 Beta3=0.03530
xi1=0.07475  lambda1= 4.1621  rho1=0.9308
xi2=-0.1661  lambda2= 3.0619  rho2=1.0224;
if type01='continuous' then do;
loglik=(-0.5*log(2*22/7))-log(sigmaError)-(0.5*(1/(sigmaError**2))*
```

```
(pro-(Beta0+Beta1*time+Beta2*(treat=1)+Beta3*((treat=1)*time)+g))**2);
end;
else if type01='survtime' then do;
loglik=log(lambda1)+log(rho1)+(rho1-1)*log(time)+(xi1*(treat=1)+kappa1*g)-
lambda1*(time**rho1)*exp(xi1*(treat=1)+kappa1*g);
end;
else if type01='censntime' then do;
loglik=log(lambda2)+log(rho2)+(rho2-1)*log(time)+(xi2*(treat=1)+kappa2*g)-
lambda2*(time**rho2)*exp(xi2*(treat=1)+kappa2*g);
end;
model outcome01 ~ general(loglik);
random g ~ normal(0,sigmab**2) subject=id;
estimate 'Error Variance' sigmaError**2;
estimate 'Variance of Normal R.E.s' sigmab**2;
run;
```