Predicting pregnancy outcomes using longitudinal biomarkers: analysis of urinary human chorionic gonadotrophin levels in normal and failing pregnancies

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Nuzhat Banu Ashra, BSc. MSc.

Department of Health Sciences

University of Leicester

April 2022

Predicting pregnancy outcomes using longitudinal biomarkers: analysis of urinary human chorionic gonadotrophin levels in normal and failing pregnancies

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Nuzhat Banu Ashra, BSc. MSc.

Department of Health Sciences

University of Leicester

April 2022

Abstract

Predicting pregnancy outcomes using longitudinal biomarkers: analysis of urinary human chorionic gonadotrophin levels in normal and failing pregnancies

Nuzhat Ashra

Early miscarriage affects approximately 25% of confirmed pregnancies and can adversely impact a woman's body and mind. Human chorionic gonadotrophin (hCG) is used to confirm pregnancy and as a triaging tool in cases of suspected pregnancy loss. Thought its current use may be limited, the potential of hCG in the context of miscarriage is greater than is currently acknowledged. Profiles of hCG for healthy and failing pregnancies have consistently been shown to be distinct, providing motivation for quantifying the association between repeatedly observed hCG and miscarriage.

Naive approaches model this association via techniques typically reserved for modelling each outcome separately. Such methods fail to consider the continuous nature of the biomarker, measurement error and appropriate estimation of uncertainty. The joint longitudinal-survival model provides a framework to simultaneously model a longitudinally observed biomarker and time-to-event outcome. In its most common incarnation, the joint model consists of a linear mixed-effects model and a proportional hazards survival submodel. The dependency between the biomarker response and survival outcome is underpinned by shared random effects, laying the groundwork for subject-specific survival predictions.

The aim was to use advanced statistical models to estimate the association between miscarriage and hCG, and to extend this to predict individual outcomes. These methods were applied to data collected by SPD Development Company Ltd. The study prospectively followed women as they tried to conceive. Volunteers collected daily urine samples from the start of their cycle to up to day 60 if they conceived. The application of cutting-edge methods to this unique dataset allowed the association to be estimated using complete longitudinal profiles of hCG. Analysis extended to diary data, to establish whether timing of intercourse can alter the rate of miscarriage. A key modelling assumption of the joint models fitted in this thesis were also assessed via a simulation study.

COVID Impact Statement

This thesis was affected by the onset of the pandemic mainly due to the lockdowns which were implemented from March 2020. This impacted timelines for completion due to the nursery and school closures, and subsequent home-schooling I had to undertake for my then four and eight-year old children. This was compounded by the University closure as I had no alternative space to work, and did not until July 2021 when my funding was close to ending. The home-schooling lasted for about 9 months, with additional periods where my children had to self-isolate due to cases arising in their bubbles. My four-year especially required a lot of support with the work the school was setting. I was the only person available to provide this support, as my partner was and still is a key worker and was working throughout the pandemic, leaving the bulk of childcare and schooling support to me. The planned projects have been completed, but within a shorter time-frame. I have tried to make an effort to ensure that the quality of this thesis has not suffered as a result.

Due to the work from home mandates, I was not able to complete a three-month placement at SPD Development Company Ltd, which was due to take place during the PhD. Had I been able to undertake the placement, the thesis would have benefited from my gaining a deeper understanding of the data analysed.

Acknowledgements

I would like to extend my gratitude to those who have provided support and encouragement throughout this PhD. To my supervisors past and present; thank you for your input and guidance. Dr Michael Crowther - thanks for hanging in there with me. Many thanks to Professor Paul Lambert for stepping into the breach. Thank you to Professor Keith Abrams; your tangents always meandered round to some great ideas!

Heartfelt thanks go to Dr Sarah Johnson and Lorrae Marriott at SPD Development Company Ltd for their unparalleled knowledge and expertise. Discussions with you both have been some of the most enjoyable and interesting moments of this PhD.

I'd like to acknowledge Dr Mark Rutherford and Professor Lucy Smith's input during each annual review. You have bolstered my confidence when it was flagging and have always been interested and encouraging.

Thank you to my former colleagues and fellow PhD students for talking me down from imaginary ledges more times than I can count, and for all of the laughter and baked goods. I'm especially grateful to Dr Sarah Booth, Dr Ridhi Agarwal and Dr Sarwar Mozumder, for cheering me on to the very end.

And last of all to my family, M, M and M (we didn't plan that). You have collectively seen me through ten years and three degrees, and the pain that goes along with it all. Thanks for making me smile when I needed it and letting me brood when I needed that. I couldn't have done it without you, and I promise we can finally have our weekends back!

Contents

List of Tables

List of Figures

List of Abbreviations

AIC: Akaike's information criterion

AIDS: Acquired immunodeficiency syndrome

ALSPAC: Avon Longitudinal Study of Parents and Children

ANOVA: Analysis of variance

AUC: Area under the curve

BIC: Bayesian information criterion

BLUE: Best linear unbiased estimate

BLUP: Best linear unbiased prediction

BMI: Body mass index

CD4: Cluster of differentiation 4

CI: Confidence interval

CrI: Credible interval

DNA: Deoxyribonucleic acid

EAPU: Early pregnancy assessment unit

FPM: Flexible parametric model

FSH3: Follicle stimulating hormone

GCC: General Cycle Collection

GEE: Generalised estimating equations

GLM: Generalised linear model

GLMM: Generalised linear mixed model

hCG: Human chorionic gonadotrophin

HIV: Human immonodeficiency virus

HR: Hazard ratio

IPCW: Inverse probability of censoring weighting

IUA: Intrauterine adhesion

JLCM: Joint latent class model

LH: Luteinising hormone

LMM: Linear mixed-effects model

LMP: Last menstrual period

LOCF: Last observation carried forward

MAR: Missing at random

MCAR: Missing completely at random

MCMC: Markov Chain Monte Carlo

MNAR: Missing not at random

NFP: Natural family planning

NICE: National Institute for Health and Care Excellence

P3G: Urinary Progesterone

PCOS: Polycystic ovary syndrome

PH: Proportional hazards

PSA: Prostate-specific antigen

QOL: Quality of life

RCS: Restricted cubic spline

REML: Restricted maximum likelihood

ROC: Receiver operating characteristic (curve)

RR: Relative risk

SD: Standard deviation

SE: Standard error

SPD: Swiss Precision Diagnostics

SREM: Shared random effects model

UK: United Kingdom

Chapter 1

Introduction

1.1 Miscarriage

Up to 25% of confirmed pregnancies end in loss, with approximately three-quarters of these losses occurring in the first trimester [5; 6]. The event of a miscarriage is a trying time, and the frequency with which the average woman experiences such an outcome can make it easy for it to be brushed aside without further thought. The trauma of a loss is far-reaching, in many cases requiring not only physical intervention but prolonged psychological support for both prospective mother and partner [7; 8]. Even a single loss can be devastating and though the majority of women go on to experience a healthy pregnancy, a select number of women (1 to 3%) will suffer recurrent losses, defined as three or more consecutive miscarriages [9]. It is for these women, who possibly exist in a greater state of anxiety than a woman who has not experienced a loss, that pregnancy monitoring may prove a boon.

1.2 Human chorionic gonadotrophin

This is where human chorionic gonadotrophin (hCG), colloquially known as the pregnancy hormone, may be useful. As a pregnancy-specific hormone, the behaviour of hCG, particularly in the first trimester, has been extensively studied. The hormone is first produced when the fertilised ovum successfully implants

into the vaginal wall. It then increases reliably and rapidly in early pregnancy, doubling every 24-48 hours until a plateau is reached at around week ten of pregnancy [10]. Though other markers are acknowledged for their association with pregnancy, hCG continues to predominately be used in practice [11; 12]. In the first instance, hCG is primarily used to confirm pregnancy. However due to its reliably consistent nature in early pregnancy, it has also been used as a means of triage in conjunction with ultrasound technology in cases of suspected miscarriage, particularly where a woman presents with signs such as bleeding or pain [13]. Moreover, hCG profiles of miscarried pregnancies have been shown to be distinct from viable pregnancies, with a generally slower rate of increase and a lower peak of hCG [14; 15]. However, apart from in assisted pregnancy settings, where hCG is tracked to monitor progression in the early stages, the information hCG can provide in the context of early loss has yet to be fully exploited [12; 16]. In clinical practice where a miscarriage is suspected the loss is commonly allowed to naturally resolve on its own [17]. Depending on how advanced a pregnancy is, the woman may be advised to take a test to establish pregnancy (hCG $>$ 25 mIU/ml), or alternatively hCG observations may be taken 48 hours apart to establish the level of hCG activity [17]. Both of these give a short-sighted view of the situation compared with the depth of information which could be gained through serial collection of hCG in early pregnancy [12; 18].

1.3 Recurrent loss

Ostensibly, miscarriage cannot be prevented, however it can be managed in a way that gives some autonomy to the woman. Where a loss is suspected, early identification through monitoring could not only prevent complications requiring

surgical intervention, but also avoid protracted emotional distress [19]. Anecdotal evidence suggests that women who have experienced previous losses become 'hyper-vigilant.' One manifestation of this is taking multiple pregnancy tests for reassurance that the pregnancy is progressing [20]. The idea of tracking hCG during early pregnancy builds on this, with a more concrete output of hCG. This is in a similar vein to having more frequent ultrasound scans, yet less time-consuming and expensive. Established and emerging treatments focus on bringing a subsequent pregnancy to term for those women who have experienced recurrent losses [21; 22; 23]. Serial hCG tracking could be implemented in conjunction with such treatments, much like in the *in vitro* fertilisation (IVF) pregnancy setting. Scenarios in which tracking hCG in relation to early pregnancy outcomes could prove beneficial certainly warrant further investigation.

1.4 Urinary observations

Up to now however, there has been a severe lack of suitable data, which would allow the association between serial hCG and early miscarriage to be appropriately modelled. In a natural pregnancy setting conception is unknown so tracking must begin prior to ovulation to capture the point at which implantation occurs and hCG begins to be produced. This, however, assumes that conception is certain to occur, when in reality the probability of conception in the first month is approximately 30% [24]. The cumulative probability of conception only increases to 75% at six months, meaning collection over several cycles would be required for relevant data [24]. Repeatedly observing a biomarker over time is costly, particularly where the norm is a serum observation. A lesser utilised but more cost-effective alternative to serum is the urinary hCG observation, which is the backbone of home pregnancy testing. It has been shown that urinary hCG

reliably follows the same pattern of rise as serum observations [25]. Advances in testing also mean that even home pregnancy tests can detect hCG values as low as 6 mIU/ml [26]. So for a reference 28-day cycle, the test claims to give results just two days post implantation [27]. Urinary hCG observations could then prove an attractive alternative to serum where early pregnancy monitoring is concerned. However, several questions remain. With established evidence of an association between hCG and early miscarriage, how can the association be appropriately quantified? And if pregnancy monitoring is a goal, how can an imminent loss be predicted based on the most current hCG profile? One answer is the joint longitudinal-survival modelling framework, which will be the focus of developments in this thesis.

1.5 Jointly modelling longitudinal and survival data

This thesis aims to make inferences about a time-to-event outcome, early miscarriage, by treating a longitudinally observed biomarker, here hCG, as a surrogate for the survival outcome. Prognostic survival models have traditionally ignored serial observations in favour of a simpler baseline value approach. Where attempt has been made to include time-varying biomarker observations, standard survival techniques have not been able to capture uncertainty or marker measurement error (sections 5.3.2 and 5.3.3) [28]. Hence, analysis of the two types of data requires the amalgamation of techniques for serial observations and time-to-event data, in a way that appropriately links the expected biomarker value to survival. The simultaneous analysis of the two types of data can be achieved by fitting a joint longitudinal-survival model.

The classical shared parameter joint model is made up of longitudinal and survival component submodels, with a dependency structure underpinned by shared

random effects [29]. This requires the specification of a trajectory function for the biomarker, which can then be included within the equation for the survival model complete with its own association parameter [30]. Joint models first evolved in the context of acquired immunodeficiency syndrome (AIDS) where the association between longitudinal cluster of differentiation 4 (CD4) counts and progression to AIDS was modelled [31; 32]. Since then the framework has been extended to different settings and different types and numbers of outcomes [33; 34; 35]. The work in this thesis, however, harks back to those early foundations, focussing for the most part on a single, continuous longitudinally observed biomarker.

1.5.1 Modelling a repeatedly observed bioamrker

A key characteristic of a biomarker such as hCG is that it is intrinsically linked with the time-to-event outcome. For hCG to be produced the pregnancy must continue to progress and if the pregnancy is interrupted, the level of hCG will also be impacted. The thesis will be dealing with these so-called endogenous biomarkers [36; 37]. Several accommodations must be made when modelling continuous repeated measures data. For one a continuous biomarker is often subject to the phenomenon of measurement error, where an observation is rounded or observed incorrectly due to human or instrumental error [38]. Moreover, longitudinal studies are often plagued with missing data, where observations for individuals are observed intermittently, not necessarily observed the same number of times and/or at the same follow-up times. The linear mixed-effects model (LMM) is perfectly placed to deal with these issues and so forms the basis of the longitudinal component model. Most important of all, the LMM framework allows for intra- and inter-subject variation, ultimately supporting the prediction of individual biomarker trajectories, and within the joint model setting, prediction of subject-specific conditional survival probabilities.

1.5.2 Linking the biomarker to survival

Survival analysis describes the analysis of data where individuals are followed up over time until they experience an event (very often death), or are censored at the last point they were seen and known to be event-free. Unique to the survival context is the incorporation of the time taken for the event to occur, so that risk is estimated as a hazard rate [39]. The proportional hazards model, which can be estimated parametrically or semi-parametrically, forms the foundation of the survival component of the joint model. The former relies on a distributional assumption for the baseline hazard, most commonly the Weibull distribution or explicit modelling of the hazard using restricted cubic splines [40; 41]. The semi-parametric approach avoids modelling the baseline hazard completely giving the Cox proportional hazards model [42]. For the joint model context it is important for prediction purposes that the baseline hazard is modelled, hence parametric estimation will be favoured. The main driver for applying these models to early pregnancy outcomes is to output real-time subject-specific conditional survival probabilities for individual pregnancies [43]. This supports dynamic monitoring, where predictions of risk can be updated based on the hCG profile to date and maternal characteristics. From a clinical perspective this feeds into the move towards personalised medicine, where predictions can prompt intervention tailored to the individual. In addition, an individualised schedule for observations can be built, depending on the hCG observations and risk predictions to date [28]. The joint model framework and its extensions have great clinical potential.

1.6 Application to data

The data analysed in this thesis offers a unique insight into early miscarriage, which usually goes unobserved in clinical practice. Collected by SPD Development Company Ltd - the industrial collaborator for this thesis - the focus for each of the two datasets was to prospectively follow up women aged 18-45 years who were intending to conceive. For each dataset, described in detail in Chapters 6 and 7, daily urinary hCG was observed from first detection, allowing the analysis of complete hCG profiles. Women who conceived collected up to day 60 of the cycle, or for up to a week after the event if they miscarried. The initial dataset is an abbreviated (n=129 vs n=367) version of the larger general cycle collection (GCC) dataset and has previously been analysed using a simple two-stage joint modelling approach [44]. The GCC study represents an expanded collection to additional biomarkers, more detailed maternal history and self-reported diary data. Enhanced prospectively collected biomarker data such as this is virtually unheard of for natural pregnancies, particularly where extensive laboratory testing has been used to pinpoint key cyclical milestones such as ovulation and the hormone surges leading up to it. To this author's knowledge there are only a handful of research studies which utilise serial hCG observations as a means for predicting pregnancy loss, though there have been numerous studies which attempt to characterise hCG curves for viable, miscarried or ectopic pregnancies [14; 45; 46; 47]. Where serial observations have been utilised, data for each individual is sparse, limited to two or three observations [45; 46; 48]. As a result, analysis in these cases follow simpler established methods, such as direct estimation of sensitivity and specificity based on whether a hCG ratio over a certain threshold can predict early loss [48]. Where more complex methods have

been proposed, notably (Bayesian) semi-parametric mixed-effects models, the aim has been to classify miscarried/healthy pregnancies [45; 46]. Access to this data presents a rare opportunity to more accurately estimate the association between longitudinally observed hCG and early pregnancy loss.

1.7 Evolving methodology

Extensions to the joint modelling framework are continually being developed, with multiple markers, competing risks and non-continuous markers all covered in the literature [34; 49; 50; 51]. However, basic modelling considerations, such as non-linear covariate effects in standard survival models, has been ignored in the joint model setting. A cornerstone of prognostic survival model building is to evaluate how the hazard changes with different values of a given covariate. This should also be the case for a biomarker which is modelled via a longitudinal trajectory function, as is the case for the joint model. If the true association between the survival outcome and longitudinal submodel is non-linear then misspecification could lead to model predictions which are biased. The standard joint model parameterisations - current value or first derivative association structures - are the most frequently discussed and implemented. Yet each association is based on linear changes in current value or slope. It is unlikely that assuming linearity will reflect the true nature of the association, hence it is necessary to evaluate the effect of such misspecification on model output.

Simulation studies are an important tool used to evaluate and compare new and existing methods or hypotheses, with data generated under a pre-specified truth. Misspecification of the submodels of the joint longitudinal-survival model have been addressed in the literature [52; 53; 54]. However, to date no study has

addressed misspecification of a non-linear association structure. A novel simulation study will be presented in this thesis exploring the effect of assuming a linear association on survival predictions when a quadratic association is simulated as the truth.

1.8 Aims of the thesis

The overarching aim of this thesis is to apply both established and cutting-edge joint modelling techniques to novel pregnancy datasets, in order to answer several hypotheses related to the early pregnancy setting. I also aim to investigate the consequences of a key modelling assumption made when fitting the joint longitudinal-survival model via a simulation study, namely that the association between the biomarker trajectory function and survival is linear, when it may not truly be the case.

The analyses undertaken in this thesis will add to the evidence base by first further examining the relationship between hCG and early miscarriage, with expansion to serial hCG observations. This aim is many-faceted, for one the focus of this thesis will be on repeated urinary observations of hCG, as opposed to the serum observations that are commonly encountered in the literature.

Secondly the association between longitudinal hCG and time-to-miscarriage will be quantified utilising the joint longitudinal-survival model. The development of a prediction model, will in turn lead to the output of subject-specific conditional survival probabilities. Producing up-to-date estimates of risk based on individual characteristics and hCG profiles will ultimately allow for dynamic monitoring. These predictions present a unique aspect of the joint modelling framework, which have yet to be applied to the early pregnancy context.

Miscarriage is much studied, yet continues to pose a mystery to clinicians. As such, many questions surrounding the whys and wherefores of a loss remain suppositions rather than evidence-based claims. One such conjecture concerns how the timing of intercourse can affect (yet to be) established pregnancy [55]. The third aim is then to analyse self-reported intercourse diary data using Cox proportional hazards models in order to establish evidence for or against several hypotheses on timing of intercourse and the rate of miscarriage. Briefly these are,

(1) Does intercourse in the luteal phase or peri-implantation window increase the rate of miscarriage?
(2) Are pregnancies conceived of acts of intercourse in advance of ovulation more likely to end in miscarriage due to ageing sperm?
(3) Do more acts in the fertile window reduce the rate of miscarriage by increasing sperm quality?

Finally, a simulation study will be carried out, aiming to evaluate an aspect of the association structure of the joint model which has been ignored until now. It is common for an association between a covariate and outcome to be non-linear in nature. Age and BMI, are examples of such variables which are commonly modelled quadratically, to describe how the risk at lower and higher values is greater than for moderate ages. However, this is a little considered scenario in the joint model setting, where the commonly described association structures between the biomarker and time-to-event outcome assume linearity. This simulation study will evaluate the effect of misspecifying a quadratic association structure on predicted survival probabilities across two scenarios and six models.

1.9 Layout of the thesis

The thesis will adhere to the following format. Chapter 2 will give some background on early miscarriage and hCG. This includes the biological foundations of pregnancy, the role of hCG and its behaviour during early pregnancy. I also touch upon the current diagnosis and care pathway of early loss and where hCG tracking, particularly of the urinary variety could fit into current clinical practice.

Chapters 3 and 4 are precursors to Chapter 5 and detail the building blocks of repeated measures and time-to-event analysis, forming the basis of the joint longitudinal-survival model. The characteristics of longitudinally observed data will be discussed in Chapter 3. Several techniques for modelling such data will be presented though the emphasis will be on the linear mixed-effects model. Particular attention will be given to the estimation of subject-specific predictions from the fitted LMM, via specification of random effects. The unique aspects of survival data will be described in Chapter 4. Non-, semi- and fully parametric approaches to analysing time-to-event data will be presented with key associated formulae. The information from these chapters will be consolidated in Chapter 5, in which the joint modelling framework will be introduced. Naive methods to simultaneously model survival and longitudinal data will be presented initially, before expanding further to how the joint longitudinal-survival model improves upon these techniques. Methods for the estimation of subject-specific conditional survival predictions, utilising the best linear unbiased predictions (BLUPs) for the random effects as discussed in Chapter 3 will be presented. Further model assessment techniques will be covered with particular focus on discrimination and calibration measures specific to the joint model setting.

Methods presented in the previous chapters will be applied to two novel datasets in Chapters 6, 7 and 8. Chapter 6 presents a re-analysis of a dataset of 129 women using a joint longitudinal-survival model. It has previously been analysed using a simpler two-stage model approach [44]. The focus here will be to establish the association between hCG and early pregnancy loss within the joint model framework. Further work on how the longitudinal trajectory for hCG, and the baseline hazard for the survival submodel should be modelled will be detailed. The knowledge gained from this analysis will be taken forward into the analysis of the larger and more detailed GCC dataset in Chapter 7. The collection of this data is a scaled up version of the smaller dataset, with more detailed collection of maternal history, additional biomarker information as well as self-reported intercourse, bleeding and morning sickness diary information. The analysis of this dataset follows a prognostic model building approach, utilising a combination of model selection techniques and expert opinion. The corresponding conditional survival probabilities dependent on longitudinal observations to date will be presented graphically. Consideration will also be given to how well the fitted model predicts event probabilities and whether false positives and false negatives are minimised. The final applied project will make use of the diary data collected as part of the GCC study and will be presented in Chapter 8. Various hypotheses around the timing of intercourse and subsequent pregnancy outcome will be explored utilising Cox models.

The research projects will be rounded off with a simulation study which addresses an aspect of the joint model specification that has been ignored up to now. The focus will be on the consequences of misspecifying a non-linear association structure for a joint model. More specifically data will be simulated from a joint model with a quadratic association structure. Two scenarios will be

described, where an increase in biomarker conveys an increase in risk and vice versa. For each scenario, three models will be fitted, assuming an increasingly greater quadratic effect. The data generating mechanisms for simulating data will be described. The true parameter estimates will be based on an example joint modelling dataset. For each non-linear association type the effect of assuming a linear association between the longitudinal trajectory function and the survival outcome will be evaluated by predicting survival probabilities for discrete time-points for each treatment group. These probabilities will be compared to the predictions from the true model via bias, empirical standard error and average model standard error estimates.

Finally, in Chapter 10 the discussion will consider where the work conducted in this thesis fits in the current evidence base. The strengths and limitations of the analyses presented in the thesis will be discussed, as well as future directions of research.

Chapter 2

Pregnancy loss and human chorionic gonadotrophin

2.1 Chapter overview

Miscarriage is a common consequence of pregnancy, particularly in the first trimester. In this chapter the biological processes which lead to pregnancy will be presented, with reference to key points in the menstrual cycle. The factors known to be associated with sporadic and recurrent miscarriage, as well as the current diagnostic pathway will be discussed. The role of the pregnancy hormone, human chorionic gonadotrophin, will be highlighted, as well as its potential to be used serially as a diagnostic marker for early pregnancy loss.

2.2 Menstrual cycle and conception

A number of biological processes must take place for a woman to conceive successfully, and these rely heavily on specific aspects of the menstrual cycle. The average menstrual cycle spans 28 days, however this can vary widely between women. Studies looking at cycle lengths have reported durations as short as 15 days to as long as 45 days; however a cycle of 21 to 40 days is thought to be within normal range [56; 57; 58]. Variation in cycle length is thought to be connected to maternal age, with greater variation reported amongst women under 25 years of

age, and over the age of 40 [59]. The cycle is split into four phases, comprising of menstruation, the follicular phase, ovulation and the luteal phase. Day one of the cycle coincides with the first day of menstruation, colloquially known as the period. It is usual for a period to last anywhere from two to seven days, and is the process by which the uterine wall sheds its lining after fertilisation does not occur. The days after menstruation are known as the follicular phase, which is characterised by an elevation in key hormone levels in preparation for ovulation. Follicle stimulating hormone (FSH3) is responsible for stimulating the growth of the immature egg cell or oocyte into a mature secondary follicle [60]. Post menstruation, oestrogen levels begin to rise from their constant low levels, peaking the day prior to the luteinising hormone (LH) surge [61]. Oestrogen enables the maturation of the most dominant follicle, namely the one that will release an ovum for fertilisation [60]. Luteinising hormone regulates the function of the ovaries and a surge in LH triggers the release of the ovum, which is known as ovulation [62]. This usually occurs about halfway through a cycle, so day 14 or 15 for an average 28 day cycle. If sexual intercourse takes place close to ovulation, ideally in the fertile window (the five days preceding and day of ovulation), the ovum may be fertilised [63]. The luteal phase follows ovulation and spans the second half of the cycle. It is at this stage that progesterone (P3G), secreted by the corpeus luteum, rises to prepare the womb for implantation of the fertilised ovum. The corpeus luteum is comprised of the remnants of the collapsed dominant follicle that housed the released egg [64]. P3G is essential for implantation and the maintenance of pregnancy. If conception is successful the fertilised ovum embeds into the uterine wall. This implantation is the initial trigger for the production of hCG, levels of which are used to determine pregnancy by home pregnancy tests [65]. If fertilisation is not successful then the lining of the uterine wall, which

has been readied for implantation will break down resulting in a period and the start of the next cycle. Figure 2.1 illustrates the key hormone fluctuations for a woman with a 28 day cycle.

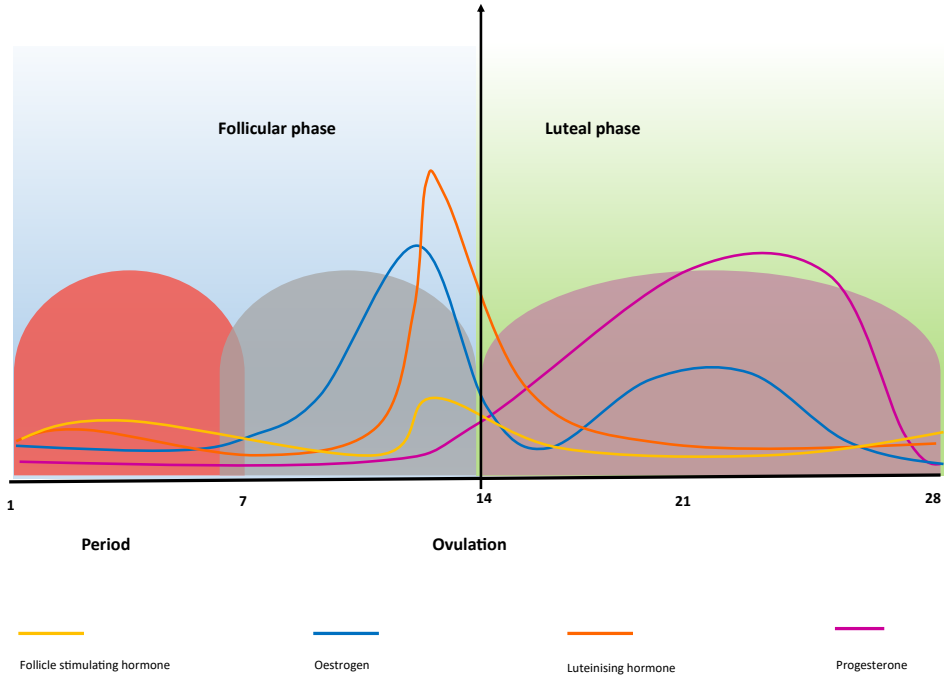


FIGURE 2.1. Hormone fluctuations for a typical 28 day cycle, adapted from 'The menstrual cycle: more than just your period' [1]

2.3 Miscarriage

Miscarriage is a frequent complication of pregnancy and in the United Kingdom (UK) is defined as the spontaneous termination of a pregnancy before week 24, with early pregnancy loss defined as loss up to and including week 12 [66]. The incidence of early loss has been reported variably from 10 to 24% of all clinically confirmed pregnancies [5; 6; 67]. The uncertainty around these figures is compounded by the fact that the majority of losses tend to resolve themselves

without medical intervention, and so go undetected and/or unreported. A Norwegian study looking at all registered pregnancies over a four year period found 11.9% of pregnancies ended in loss before 12 weeks [68]. However, as even these estimates are based on confirmed pregnancies, the true figure is suspected to be higher when including losses which occur prior to the day of the missed period. An estimated 70 to 80% of all losses occur in the first trimester, with half of all early losses attributed to foetal chromosomal abnormalities [6; 69]. Increasing maternal age is a strong prognostic factor for miscarriage from the age of 35 years onwards, with a 75% increase in odds of miscarriage for women aged between 35-39, and a five times increase in odds for women aged 40 and over, when compared to the 25-29 year age group [70]. Furthermore, women with a BMI greater than 30kg/m^2 have a 20% higher odds of miscarriage and 3.5 times odds of recurrent early loss when compared to healthy weight controls [71]. Smoking during pregnancy is another common factor known to adversely affect outcomes, with a 1% increase in pooled relative risk for every cigarette smoked per day [72]. Based on landmark studies utilising hCG to identify preclinical pregnancies, it has been postulated that 22 to 30% of ovum are lost after implantation but before the day of the missed period (biochemical or occult pregnancy) [73; 74; 75; 76]. However, it has also been suggested that 15 to 30% of ovum never successfully implant and so are lost even before hCG can be detected to identify them [75; 76]. The discrepancies in figures makes it difficult to pinpoint the true incidence of early pregnancy loss, yet it is unfortunately clear that it is a common outcome of conception. What is more, experiencing one miscarriage increases the probability of another loss [67; 77].

2.3.1 Recurrent miscarriage

Recurrent miscarriage, which in the UK refers to three consecutive losses, affects 1 to 3% of women [9]. Once a woman experiences this arbitrary number of miscarriages, further investigations can take place to identify the reason for these continual losses. A probable cause, however, can only be identified in 50% of cases [9; 69]. In those instances where a contra-indicator for bringing a pregnancy to term can be identified, treatment offered aims to prevent a further loss. Chromosomal abnormalities are usually the most common unpreventable culprit, and the probability of subsequent losses increases with the age of the mother (over 35) and father (over 40) [69; 78].

A treatable cause of recurrent loss is antiphospholipid syndrome (APS), which is an illness which makes blood more likely to clot. In the first trimester this can interfere with implantation and consequently the initial establishment of the pregnancy. The clotting also affects the adequacy of blood flow through the placenta [78]. Treatment for APS constitutes a low dose of aspirin and heparin at pregnancy onset [79]. Thrombophilias, the catch-all term for a set of blood clotting problems, have also been associated with recurrent loss, although later in pregnancy [78]. Conditions such as polycystic ovary syndrome (PCOS), uncontrolled diabetes and thyroidism; infections such as toxoplasmosis, rubella, listeria and genital infection; as well as elevated levels of uterine NK (uNK) cells have all been linked to an increased risk of recurrent miscarriage, although the evidence base remains sparse [78].

Hormone levels during pregnancy and the link to recurrent loss has received much attention in the literature, specifically levels of progesterone. P3G is essential for the maintenance of pregnancy, therefore depleted levels could be a cause of some recurrent losses. Both the PROMISE and PRISM trials evaluated the effect

of progesterone therapy for women who experienced recurrent loss and those who presented with bleeding in early pregnancy [21; 80]. No statistically significant difference in live birth rates was found between the recurrent loss and placebo groups in the former trial [80]. Treatment with progesterone corresponded to a 3% greater birth rate for women who experienced bleeding in early pregnancy. Though this was not a statistically significant finding, there were no adverse effects observed related to the therapy. In addition a greater proportion of live births were seen amongst the subgroup of women who had previously suffered three or more losses when compared to the placebo group [22]. Research suggests there may be clinical value in progesterone therapy for some groups of women.

A relatively recent discovery has pointed towards a scarcity of stem-like progenitor cells in the endometrium lining as the cause of recurrent loss for some women [81]. These cells are more specific than stem cells in that they already have a target cell for differentiation [82]. A small feasibility study found that sitagliptin, a drug usually used to treat insulin resistance, successfully increased the stem-like cell counts in women who experienced three or more previous losses [23]. Although promising, a full-scale randomised controlled trial is required to fully evaluate the efficacy of this treatment regimen.

In some cases recurrent loss can be attributed to the quality of the sperm, with poorer sperm motility identified amongst those couples who experience recurrent miscarriages [83]. Sperm DNA fragmentation is also increasingly being linked to recurrent loss, particularly where no other explanation has been found [84]. The sperm DNA damage is thought to affect embryo development and the success of implantation [85]. It is clear that the reasons behind loss are complex with interplaying factors involved, some of which can be addressed whilst others

cannot. In either case the effects of loss on couples can be far-reaching both mentally and physically.

2.3.2 Consequences of miscarriage

Due to the relatively high occurrence of miscarriage the associated healthcare costs are also substantial [5]. Treatment strategies vary depending on whether products of conception are retained in the womb or passed naturally. Ranging from the least to most invasive, care can follow an expectant, medical or surgical management approach. The former allows the tissue to pass with no intervention, whereas the latter two utilise medication or surgical intervention to remove any retention. In addition to medical intervention, those losses which are not self-resolving may also require diagnostic testing to decide on the correct treatment pathway and follow-up care [5]. The aftermath of a miscarriage can leave behind signs of trauma in the womb, with one in five women left with intrauterine adhesions (IUA) post-loss. A meta-analysis of ten studies found women who experienced three miscarriages had two times greater pooled odds of IUAs than women who suffered one loss [86]. Where surgical intervention is required the severity of such adhesions may be exacerbated and can impact the time taken to conceive again [7].

In the majority of cases, no physical intervention is required, however the abrupt halt to impending motherhood inevitably has ramifications for mental health. Pregnancy loss is understandably associated with emotional distress [87]. Women who suffer from a miscarriage are likely to report symptoms associated with depression and anxiety, with the number of affected women ranging from 20% to 55% [88]. It is common for women to blame themselves, and recurrent losses can leave women feeling disconnected from a subsequent pregnancy as they

anticipate another loss [89; 90]. Women who are pregnant for longer before miscarrying, experience a prolonged period of grief, having established a more secure bond [91]. The residual effects of a loss persist years later, and the birth of a healthy baby does not necessarily alleviate grief [92]. Evidence suggests that the number of previous miscarriages is associated with signs of depression in a subsequent pregnancy [93]. Feelings of grief are not confined to women, but also affect prospective fathers, which has the potential to affect relationships [89]. The psychosocial effects of a loss are long standing, which suggests that women will require access to adequate support during and long after the loss occurs. Early identification of a miscarriage could potentially mitigate the sense of loss somewhat. It may also reduce further physical and psychological deterioration, precluding the need for already scarce counselling or therapy resources over a protracted period of time. Miscarriage however can only be pre-emptively identified if diagnosis can be adjusted to allow so.

2.3.3 Current diagnosis of miscarriage

Women who are suspected to be miscarrying usually present with vaginal bleeding and pain in the lower abdomen. Heavy bleeding rather than spotting is usually more indicative of an impending loss [94]. Additional symptoms can include a discharge of fluid or tissue and the absence of pregnancy symptoms [95].

Currently diagnosis of very early miscarriage (< six weeks) falls under an 'expectant management' framework. Where miscarriage is suspected because of vaginal bleeding, but there is no pain, it is advised that the woman should take a pregnancy test 7-10 days later [17]. If pregnancy is confirmed or symptoms worsen then further action can be taken, but otherwise it is more common for a miscarriage to naturally resolve itself at this early stage. As a result women who

experience loss at this point are left, in a sense, to fend for themselves, as no level of medical intervention can induce a positive pregnancy outcome.

For pregnancies beyond the sixth week, if a woman presents with serious complications, such as abdominal or pelvic pain, then hospital admission is deemed necessary. If there is no pain, and the pregnancy has progressed beyond the sixth week, then referral is advised to an early pregnancy assessment unit (EPAU) for further investigation [17]. This can involve blood tests 48 hours apart to observe changes in hCG levels. However, it is more usual for transvaginal or external ultrasounds to be performed in conjunction with the blood tests [13]. This is in spite of the fact that before the fifth week of pregnancy ultrasounds cannot reliably detect the gestational sac to confirm viability, whereas hCG is readily present at this point [96].

Women who experience pregnancy loss prior to the six-week 'cut-off' point are currently only directed to a pregnancy test to track possible miscarriage. This could feel inadequate from the patient's perspective, and exacerbate feelings of helplessness. The current management framework, advocating a 'wait and see' approach can result in greater distress for some women even after six weeks. Some women prefer intervention in the case of loss to avoid waiting for a natural resolution, whilst others prefer to undergo a natural miscarriage. What is most apparent from the literature is that women would like greater choice in their care when it comes to pregnancy loss [97].

Some women experience no recognisable signs of loss. These cases are referred to as missed or delayed miscarriages and are only identified at routine appointments where a heartbeat can no longer be detected [95]. As the body fails to expel the pregnancy tissue of its own accord, this can be a particularly traumatic loss. Not only is the woman blind-sided, but she faces intervention to pass

the deceased foetus. This is another scenario in which early detection would be beneficial, to prevent prolonged suffering.

Miscarriages vary in their presentation, which makes misdiagnosis a concern. To avoid this, final diagnosis is usually based on ultrasound scans. Current guidelines define pregnancy loss based on size of gestational sac and foetal measurements (crown rump length) and a detectable heartbeat [98]. Once miscarriage is strongly suspected, it is essential that healthcare professionals correctly distinguish between incomplete and complete miscarriages [5]. This means sometimes several ultrasounds are necessary to confirm that all foetal tissue has been expelled [99]. In cases where a complete miscarriage is incorrectly diagnosed, women can experience abnormal bleeding, abdominal pain and/or fever due to retention of foetal tissue. If left untreated this can result in internal bleeding, infection and in some cases even infertility [7]. Confirming miscarriage must then tread the line between haste to prevent further deterioration of the patient and thoroughness to avoid a miscalculated diagnosis.

2.4 Human Chorionic Gonadotrophin

Human chorionic gonadotrophin is a hormone typically produced in early pregnancy. It is produced first by the embryo and then the placenta takes over thereafter. The role of hCG is to encourage the body to continue producing P3G, preventing the onset of the next cycle via menstruation. This ensures the corpeus luteum and lining of the womb continues to support the pregnancy [100]. Most importantly of all hCG stimulates the maternal thyroid gland to encourage successful implantation [101].

The hormone can be detected in maternal blood (serum) once implantation has occurred, as early as six days after conception, and in maternal urine a few

days later [102]. Currently urinary hCG is almost exclusively used for detection of pregnancy in home pregnancy tests. A level of 25 mIU/ml indicates a confirmed pregnancy. In healthy pregnancies serum hCG concentrations are reliably known to rise at a rate of approximately 50% per day, before decreasing and plateauing to a stable concentration after week 10. However, the range of 'normal' values can vary greatly across pregnancies (See Figure 2.2) [14]. This pattern is similarly replicated in maternal urine [25].

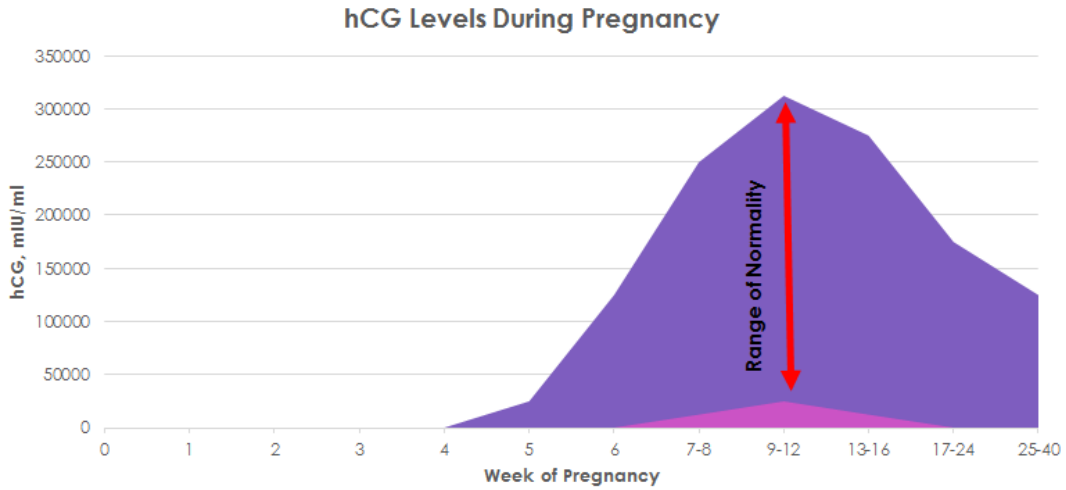


FIGURE 2.2. Range of hCG levels during pregnancy

Although ultrasounds are the main mode of diagnosis in pregnancy loss, hCG can be measured as a first port of call. As isolated measurements their only use is to confirm pregnancy, and cannot be used to confirm viability. Yet in clinical practice, hCG measurements are currently used to gain a snapshot understanding of the rate of increase in cases of suspected pregnancy loss [103]. In contrast to current practice, monitoring hCG as a series of collected measurements has been advocated. This would allow the trajectory of measurements to be compared to expected growth patterns for a viable pregnancy, and hence allow a failing pregnancy to be distinguished from a healthy pregnancy [18]. The evidence base for

hCG suggests that trajectories of hCG in failing early pregnancies are markedly different from viable pregnancies. This emphasises the potential to extend its use in a diagnostic capacity and suggests an association between hCG and miscarriage [104].

The Prior *et al.* [105] publication of the priorities for research within miscarriage ranked the identification of effective interventions to prevent (threatened) miscarriage highest. This research area encompasses the plausibility of using biomarkers, such as hCG, to track pregnancy progression through to viability or miscarriage. A meta-analysis looking at the predictive capabilities of various hormones found, across eight studies which measured intact or β - hCG in serum, that overall hCG had good sensitivity for detecting miscarriage at the eighth week of pregnancy, when compared with another biomarker. Yet, once a foetal heartbeat was detectable ($\sim$ six weeks), diagnostic accuracy for hCG was lower than other biomarkers [11]. In the case of biochemical pregnancies there is agreement of the benefits of hCG monitoring. A study conducted by Wilcox *et al.* [73] in the 1980s followed 221 women who were attempting to conceive. Of these women 32% experienced pregnancy loss and two thirds of these losses, which occurred around the expected time of implantation, would have been unidentifiable without measuring hCG.

In recent years, monitoring of hCG measurements in pregnancy has become common in women who undergo IVF [106]. Where hCG has been tracked for spontaneous conceptions the evidence is largely based on serum hCG observations. Yet a more cost-effective approach could be found in tracking maternal urine hCG measurements. Similar patterns of hCG have been noted across serum and urinary measurements, and current technology allows for detection of extremely low hCG values even in urine [44]. This could then be an opportunity

for a greater number of measurements to be taken at a lower cost to form a larger picture of how hCG trajectories differ between viable and failing pregnancies.

2.4.1 Modelling serial hCG

Ample evidence of the distinguishing features of viable and miscarried hCG trajectories is emerging. What is more uncertain is how these trajectories can be utilised in practice and which statistical techniques can best link hCG with the outcome of the pregnancy. Measurements of hCG are only sparsely observed in practice, and so statistical techniques classifying these unbalanced trajectories into abnormal and normal pregnancy groups have been investigated. De la Cruz *et al.* [46] proposed fitting semi-parametric non-linear mixed effects models, with penalized splines modelling the non-parametric component, to each pregnancy class. The superiority of modelling individual deviation via the parametric component using random effects versus directly estimating the correlation structure of the errors was assessed. The latter was found to more accurately classify trajectories [46]. An alternative approach advocated using a single Bayesian non-parametric classification model for the joint outcome of pregnancy outcome and longitudinal hCG, avoiding the requirement of separate discriminant models for each pregnancy outcome [45]. A further analysis extended classification with respect to multiple hormones, utilising a multivariate non-linear mixed effects model with random effects [107]. All of these methods were applied to a dataset of 173 pregnant Chilean women contributing serum hCG measurements over urinary measurements. Ninety-eight percent of women had three or fewer measurements, with 28% of women contributing only one measurement. This is common amongst serial hCG studies, where serial is usually a misnomer for just the two measurements where a ratio of observations is analysed [48]. In practice this is likely probably all that would be observed.

A study, which prospectively followed women who contributed daily urinary hCG samples through conception and early pregnancy outcome, provides a unique insight into the benefits of urinary hCG and continuous follow-up and provides the foundation for this thesis [44]. Two-stage models, discussed in section 5.3.3, were implemented to directly estimate the risk of miscarriage using survival models (see section 4.6), based on subject-specific predictions from an LMM (see section 3.5) used to model hCG profiles. This analysis found a delay in the time from LH surge to hCG reaching 25 mIU/mL increased the risk of miscarriage, though prediction of pregnancy outcomes was not considered. The data used for the analysis performed by Marriott *et al.* [44] will be described further and re-analysed using the more sophisticated joint longitudinal-survival model, described in Chapter 6. This analysis will be further extended utilising a larger collection of longitudinal hCG and early pregnancy outcome data in Chapter 7 with a view to predict early pregnancy outcomes. This type of monitoring is currently exclusive to IVF settings. Transferability of such tracking to a natural pregnancy setting has the potential to help women who have suffered recurrent losses [15]. The combination of cheaper to collect urinary samples and sampling more often may be the key to identifying losses as early as possible, and serve to reassure prospective mothers.

2.5 Discussion

Miscarriage is relatively common and with it comes substantial healthcare costs and emotional trauma. Currently diagnosis of miscarriage can vary depending on presentation, but almost always in cases after the sixth week of pregnancy involve diagnostic tests such as ultrasounds. Serial hCG monitoring as a potential predictor of miscarriage has not been embraced in clinical practice. With such varying aetiological mechanisms behind early loss, the common thread amongst

all pregnancies is that hCG profiles may reliably reflect viability. With urinary measurements simpler and cheaper to collect, the way forward may be to observe greater numbers of cost-effective measurements to track a pregnancy in high risk women. The IVF model has shown tracking to be beneficial, which means women who have experienced loss and are actively attempting to conceive may benefit from prospectively testing their hCG levels. This can reassure women and in the event of loss provide impending warning of a miscarriage, even when clear symptoms do not manifest themselves.

Chapter 3

Longitudinal models

3.1 Chapter Overview

This chapter will introduce conventional methods used to analyse a continuous repeatedly measured biomarker. Simple methods will first be introduced, typically revolving around summary variables and related analysis methods, including linear regression. The multilevel model framework will be discussed in detailed, with particular focus on the estimation of subject-specific effects. Alternative modelling frameworks will be presented briefly, including generalised estimating equations.

3.2 Introduction

Longitudinally observed data - also known as panel and serial data - is frequently encountered in clinical trial and observational study contexts. Collecting repeated observations for individuals can have many advantages, particularly when the objective is to study change in an outcome over time [108]. When an individual is known to suffer from a condition, it is commonplace to observe serial measurements to monitor deterioration or improvement. Medical scenarios where serial measurements are encountered, include blood pressure measurements in relation to cardiovascular disease and glucose measurements in diabetic patients [33; 109; 110]. Growth studies, particularly in child development, often look at

markers at varying ages. An example is the Avon Longitudinal Study of Parents and Children (ALSPAC), first established in 1991, which investigated the interplay of social and environmental influences and genetics on child development [111]. Though collecting repeatedly has higher cost implications, there are many reasons why these observations might aid in answering varied and more nuanced research questions. For example, interest may lie in making inferences about the overall mean response, such as for a given therapy. Alternatively, with multiple observations available it would be feasible to more completely estimate the variation between individuals. Or if individual differences are of importance, the dependency between measurements for a given person could be estimated, ultimately allowing prediction of individual response profiles [108]. Serial observations are clearly characterised by different levels of variation. Observations for a given individual are correlated and measurements from the same individual vary (within-subject variation). Longitudinal trajectories also vary between individuals, leading to between-subject variation [108]. The subsequent method of analysis must appropriately model the different levels of variation which are intrinsic to longitudinally observed data and ultimately decide whether population level or subject-specific inferences are to be made.

3.3 Measuring longitudinal continuous biomarkers

Though methods discussed in this chapter can be adapted to binary or categorical data, the focus of this thesis is on continuous biomarkers, as this type of data will be modelled in later chapters. Measuring a continuous biomarker is challenging. It is rarely possible for a measurement on a continuous scale to be observed 100% accurately. Inherently, instruments used to observe biomarker measurements can only do so to a pre-specified degree of accuracy. Furthermore, a biomarker may be

measured and then rounded further for manageability. Where a measuring device is incorrectly calibrated, or incorrectly used further disparities may arise. This means when modelling the possibility of measurement error must be accounted for [52]. Typically, biomarkers will be observed at discrete time-points, so to infer how the biomarker changes between measurements it is necessary to build a trajectory of biomarker measurements accounting for systemic measurement error. To do this one must consider that observations for a given individual will be correlated and so methods that assume independence, such as simple regression models, are invalid [30].

3.3.1 Missing data

Modelling data repeatedly over a length of time may result in incomplete or intermittently observed data. For instance, an individual may contribute a biomarker observation, before missing the next follow-up, but then attend and contribute measurements for the remainder of the study. Missing observations arising in this way is commonly encountered in longitudinal studies and results in unbalanced data [108]. The series of observations can also frequently be cut short due to an individual dropping out before the end of follow-up [112]. In each case, if the underlying missing mechanism is related to the longitudinal response then bias can be introduced during estimation. Missing data is classified into three types, missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [113]. These are defined as follows,

- MCAR: the probability of the observation being missing is independent of the longitudinal response
- MAR: the probability of an observation being missing is related to the *observed* longitudinal data only

- MNAR: the probability of an observation being missing is dependent on unobserved values of the longitudinal response

Examples of MCAR could be missing a follow-up visit due to work commitments or leaving the study due to a move. Neither of these events are linked to the biomarker, so standard analysis methods could be undertaken, effectively ignoring the missing data. The MAR assumption does presume an association between the longitudinal observations and the state of being missing. This scenario may arise if a healthcare professional advises leaving a study due to measurements observed up to that point. Typically longitudinal data which are MAR can be modelled using random effect models, which are designed to accommodate unbalanced data. The inclusion of the random effects provide an implicit prediction of unobserved data [108]. The linear mixed-effects model (LMM) which will be discussed in detail in section 3.5 is one such modelling framework. The final definition, MNAR, is the most problematic in terms of the analysis as the underlying missing mechanism depends on both observed and unobserved values of the longitudinal response. This definition of missing data could occur as a consequence of an adverse event or death. These examples are also instances of informative drop-out where a patient leaves the study due to worsening outcomes, which links the reason for the 'missingness' to both the observed measurements and what would have been observed had the patient not dropped out. This necessitates the modelling of the longitudinal and missing mechanisms via a joint distribution [112]. The joint longitudinal-survival model, which will be introduced in Chapter 5, can address the issue of informative drop-out described in this final scenario by jointly modelling the missing mechanism and time-to-event outcome [30][114].

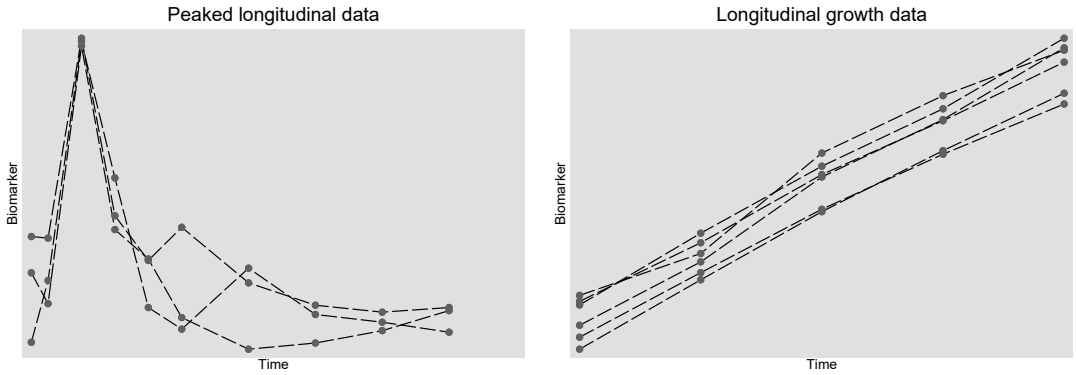
3.4 Approaches to longitudinal modelling

Where the number of measurements exceeds two, which will be the case for data analysed in this thesis, a variety of analyses can be undertaken. These broadly fall into three categories. Summary measure approaches are the most simplistic and rely on paring the data down to one or two summary statistics. These approaches will be discussed further in sections 3.4.1 and 3.4.2. More commonly random effect models, specifically linear mixed effect models are utilised, which allow inclusion of all of the longitudinal data whilst also accounting for confounding variables. The variation within and between individuals is explicitly estimated through inclusion of random effects. These models will be implemented in the joint modelling context in later chapters and detailed in section 3.5. Generalised estimating equations (GEEs) are an alternative modelling approach for longitudinal data which produce only population-average effects. These will be discussed briefly in section 3.6.

3.4.1 Summary measures

A naive approach to analysing longitudinal data is to compute summary measures, which reduce individual responses down to a single or pair of observations [115]. These are used to assess differences between groups, eliminating within-subject observations from consideration [116]. The choice of method is usually driven by the type of data being collected. These are either peaked (characterised by a rise before reaching a maximum and then decreasing), or growth data which increases over time [117]. Illustrations of both types of data are shown in Figure 3.1.

A common method of analysis for longitudinal data, where group differences are of interest, is to estimate the mean response for individuals in a given category



(A) Longitudinal peaked data (B) Longitudinal growth data

FIGURE 3.1. Examples of types of longitudinal data

at each time-point, with corresponding precision estimates, usually one standard deviation. A t-test is then used to compare groups at each time-point [117]. There are a number of issues with this approach. Carrying out several t-tests is akin to making multiple comparisons and if the significance level is not adjusted for multiple testing then spurious significant results are likely. Though individual variation is eliminated, even mean responses at successive time-points are likely to be correlated, meaning the t-tests cannot be considered independent [117]. Furthermore, comparing mean responses at two time-points does not consider the possibility that the same (numbers of) individuals may not be being compared between time-points [117]. Most concerning of all is that summarising in this way eliminates the original information and prevents the study of individual variation.

Other simple analysis measures have been proposed [117]. When looking at peaked pre and post-prandial glucose response curves it is usual to estimate the area under the curve (AUC) between groups [118]. If the maximum or minimum response is important then the mean maximum (minimum) could be compared between groups or even mean time to maximum (minimum) response. With growth data, hypotheses related to the rate of change can utilise standard linear

regression models, which can incorporate a grouping variable for comparisons between groups at a single time-point [119]. Simple analyses have their place, particularly as they can be implemented with ease. However, cross-sectional analyses provide only a snapshot of the true scenario, preventing any inferences of changes over time [120].

3.4.2 Analysis of variance

A way of incorporating all measurements into an analysis is to use the analysis of variance (ANOVA) method. This technique is usually applicable to what is called balanced data, where a measurement is observed at every fixed time-point for every individual, i.e. no missing data, which as discussed in section 3.3.1 is rarely true of longitudinally measured observations. The repeated measures ANOVA considers time to be an independent factor with n levels or related groups. At each level the same subjects provide observations of the response variable [121]. A one-way repeated measures ANOVA extends the standard one-way ANOVA by contrasting within-level and between-level variation. It is assumed that residuals of the response are normally distributed and that the variance of the difference between any possible combination of levels (time) are equal. The two-way repeated measures ANOVA extends this to also include a factor variable. This can allow inferences about a grouping variable of interest and its interaction with time, with the null hypothesis assuming there is no difference in mean response between groups over time. These are simple analysis methods which can give an indication of whether there is a difference in response over time or not, however several provisos need to be satisfied to avoid inappropriate conclusions. ANOVA methods treat time as a discrete variable, requires complete, balanced data and do not allow for several levels of clustering or investigation of covariates which change over time [108]. Extensions to non-balanced data have been proposed

however, if sample sizes and variances are both unequal then statistical power is reduced greatly [122]. With ANOVA more focussed on hypothesis testing than estimation, when interest lies in quantifying an association a linear mixed-effects model would be preferable to the limited output of an ANOVA.

3.5 Linear mixed effects models

Linear mixed effects models are able to model unbalanced data. In the case of longitudinal studies, measurements can be sporadically observed, measured at unequal time-points and with varying frequency for a given individual, with missing data common [123]. Each set of repeated measurements are nested within an individual, creating a hierarchical structure [124]. An illustration of unbalanced longitudinal blood pressure data is shown in Figure 3.2. The shaded boxes indicate missing observations.

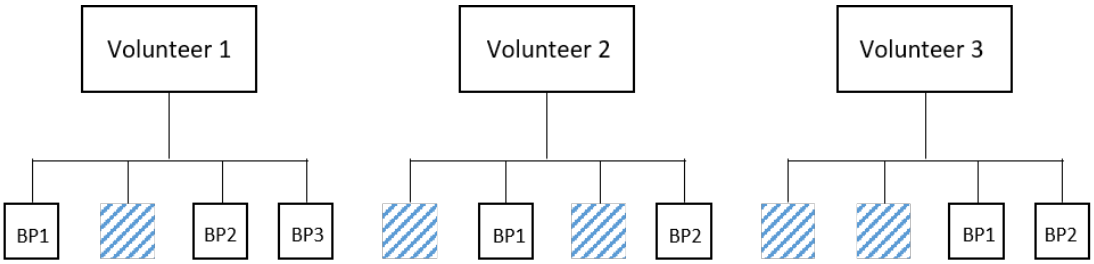


FIGURE 3.2. Example of unbalanced longitudinal data

Typically different observations from the same individual are correlated. The within-subject variance can be appropriately modelled, through the formulation of the variance components within the multilevel model framework, the first incarnation of which was proposed by Laird and Ware [125]. As discussed in section 3.3, measurement error is an enduring issue when observing a continuous biomarker, which can also be accommodated in the mixed model context. Variance components account for both variation between measurements for a given

individual as well as between individuals, through the inclusion of random effects [126].

The general formulation of an LMM for a biomarker response variable Y_{ij} observed for individual i at time j is given by Equation 3.1 [108].

$$y_{ij}(t) = \mathbf{X}_i(t_{ij})\boldsymbol{\beta} + \mathbf{Z}_i(t_{ij})\mathbf{u}_i + e_{ij}$$
$$\mathbf{u}_i \sim MVN(0, \mathbf{G}) \qquad e_{ij} \sim N(0, \sigma_e^2) \tag{3.1}$$

Here $\mathbf{X}_i$ represents the design matrix of fixed effects for patient i. This consists of all covariates that need to be be incorporated into the analysis, but don't need to necessarily vary over time, with $\boldsymbol{\beta}$ the associated fixed effects parameters. The fixed effects represent the population-average effects. Conversely $\mathbf{Z}_i$ is the design matrix for the random effects and $\mathbf{u}_i$ the associated random effects parameters. The random effects parameters are assumed to be multivariate normally distributed with mean zero and matrix of variance components $\mathbf{G}$. For a two-level model $\mathbf{G}$ is comprised of the between-subject variances, σ_u^2. The overall between-subject variance is partitioned appropriately depending on the number and structure of the specified random effects (see section 3.5.1) [124]. The residual error term e_{ij} takes into account the measurement error associated with continuous biomarkers and is normally distributed with mean zero and level-one or within-subject variance σ_e^2. All elements of the variance-covariance matrix $\mathbf{G}$ and residual error e_{ij} are assumed to be independent. Residuals at each level are assumed to be independent. The covariance between biomarker responses for the same individual is assumed to be positive. More formally these assumptions are written as,

$$cov(u_i, e_i) = 0 \qquad cov(u_i, u_j') = 0 \qquad cov(y_{i_1}, y_{i_2} \mid x_i) \geq 0 \quad (3.2)$$

The LMM is a variance components model and the within- and between-subject variances can be partitioned to describe the variation for different components of the model. The intra-correlation coefficient describes the correlation between observations on the same subject and is estimated as $ICC = \sigma^2_{u_0}/(\sigma^2_{u_0} + \sigma^2_e)$. Figure 3.3, adapted from Merlo *et al.* [2], illustrates the levels of variation for two-level longitudinal data. The fixed effects give the population-average mean of the biomarker. The subject-level deviations or residuals are represented by the solid black lines which indicate higher or lower mean biomarker values for individuals. The between-subject variance describes how individuals vary from the overall mean. This can be described using random effects. Within-subject variance is an estimate of how biomarker measurements for a given individual vary from observation to observation.

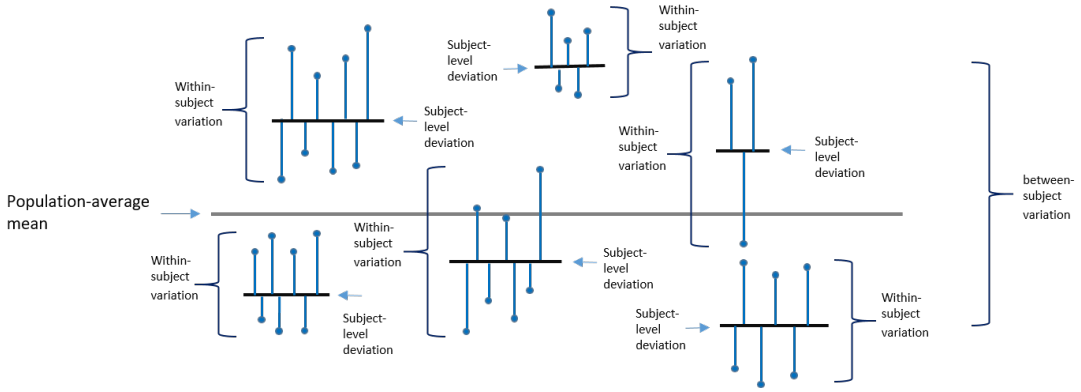


FIGURE 3.3. Illustration of the levels of variation for a two-level longitudinal data adapted from Merlo *et al.* [2]

The between-subject variance can be decomposed by including a random intercept. Figure 3.4 presents a visualisation of a random-intercept LMM.

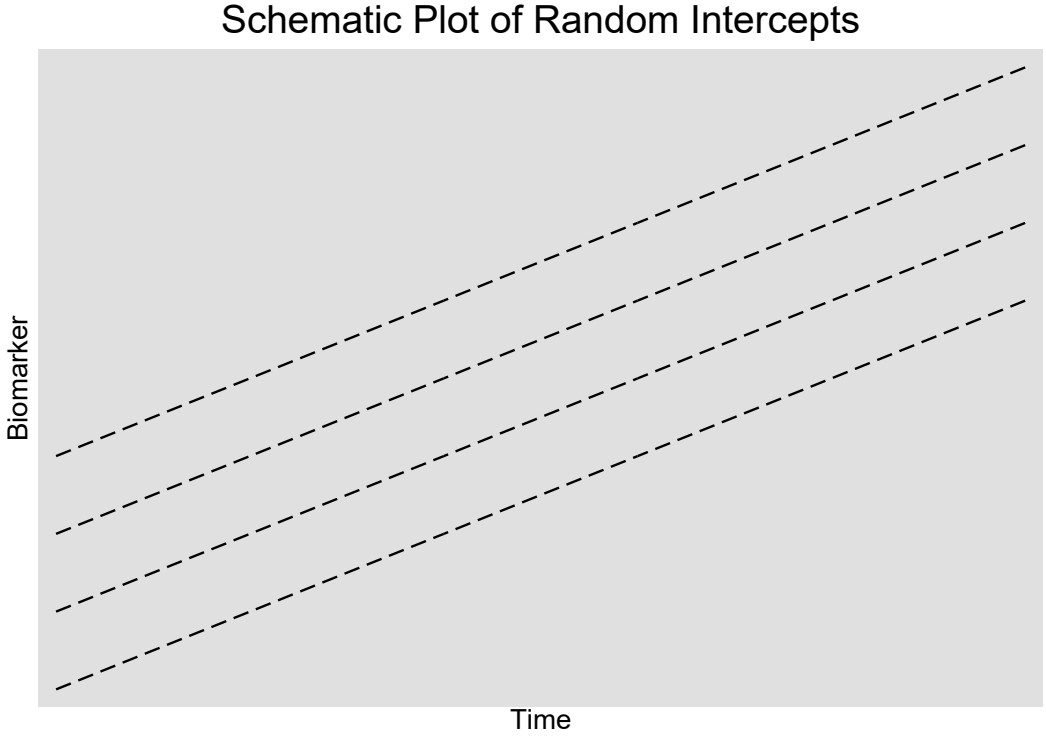


FIGURE 3.4. Schematic plot of random intercept model

An example of a two-level random intercept model is given in Equation 3.3.

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + u_{0i} + e_{ij}$$

$$u_i \sim N(0, \sigma^2_{u_0}) \qquad\qquad e_{ij} \sim N(0, \sigma^2_e) \tag{3.3}$$

In this scenario t_{ij} represents the time at which the biomarker response was observed, j, for individual i. Inclusion of a random intercept, u_{0i} allows individual biomarker values at $t = 0$ to vary from the mean given by the fixed intercept β_0. This means an individual's baseline biomarker value is allowed to be higher or lower than another's. The random intercept u_{0i} term represents the subject-specific baseline biomarker deviation from the average. The deviations

are assumed to follow a normal distribution with mean zero and between-subject variance $\sigma^2_{u_0}$.

This model can be extended with the addition of a random slope, which allows the gradient of individual biomarker trajectories to vary. A visual representation of this is shown in Figure 3.5.

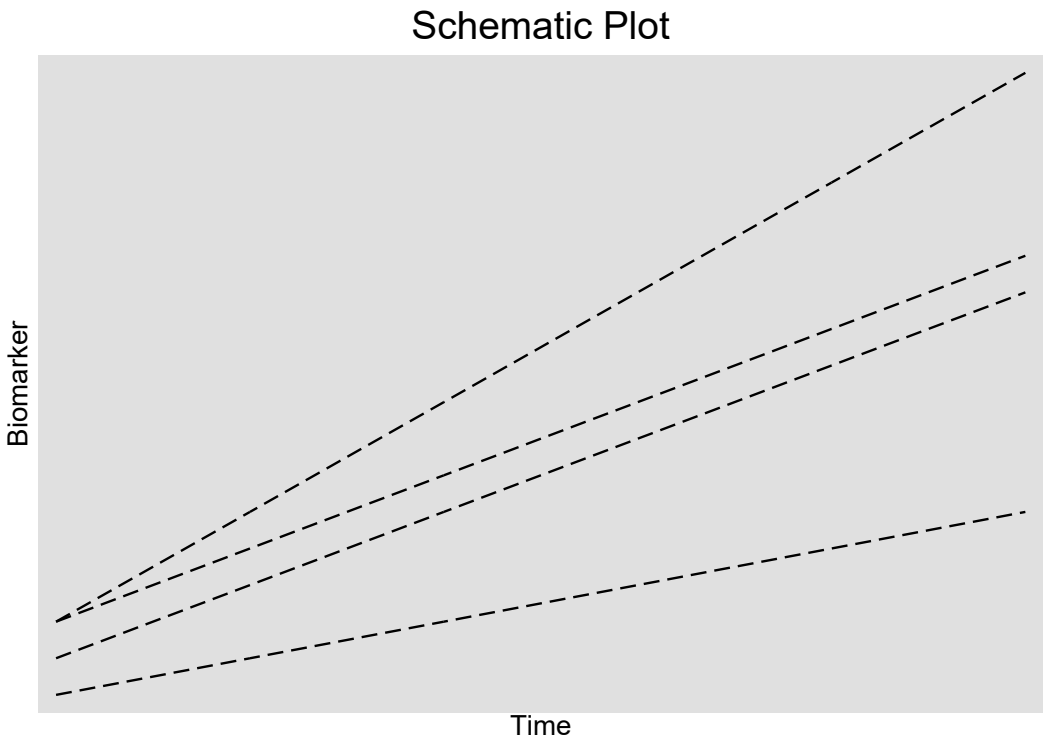


FIGURE 3.5. Schematic plot of random intercept and slope model

The formulation of a two-level random intercept and slope model is given in Equation 3.4. This assumes a linear slope for t_{ij}, with unstructured variance-covariance matrix, which will be discussed further in section 3.5.1.

$$y_{ij} = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i})t_{ij} + e_{ij}$$

$$u_i \sim N(0, \mathbf{G}) \qquad \mathbf{G} = \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_0,u_1} \\ \sigma_{u_0,u_1} & \sigma^2_{u_1} \end{bmatrix} \qquad e_{ij} \sim N(0, \sigma^2_e) \qquad (3.4)$$

The addition of the random slope u_{1i} represents the subject-specific deviation from the population-average estimate of the gradient, represented by β_1. In addition to the between-subject intercept variance term, $\sigma^2_{u_0}$, a second variance term is estimated, $\sigma^2_{u_1}$, which represents the between-subject slope variance. An additional covariance term, σ_{u_0,u_1}, describing the correlation between the intercept and slope, is estimated when assuming an unstructured matrix structure for $\mathbf{G}$.

When fitting either the random intercept or combined random intercept and slope model the individual deviations from the mean are not estimated, rather the between-subject variances are. These models are only the simplest examples of linear mixed effects models. If it is sensible to do so, further random effects can be included, perhaps to describe a non-linear slope, and additional clustering would imply more levels are appropriate. If the slope is non-linear additional (fractional) polynomial or spline terms could be included to achieve a better fit. Polynomial terms and restricted cubic splines (see section 4.6.8) will be implemented in Chapters 6 and 7 respectively.

3.5.1 Variance-covariance structure

As noted in section 3.5 the random effects are not directly estimated, but rather they are assumed to be normally distributed with mean zero and variance-covariance matrix $\mathbf{G}$. The covariance structure of matrix $\mathbf{G}$ must be pre-specified for estimation of the LMM. The diagonal of the matrix contains the variances for the random parameters, whilst the covariances are the off-diagonal entries. Those most commonly implemented are the independent, exchangeable, autoregressive

and the unstructured covariance structures [127]. Each of these specifications will be presented for a two-level random intercept and slope model.

The independent structure, presented in Equation 3.5, sets the covariance to 0, estimating one common variance term σ^2.

$$\mathbf{G}_{\text{indep}} = Var(u_i) = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \tag{3.5}$$

The exchangeable structure assumes a single variance for all random effects, and consequently a single covariance parameter. This implies a constant variance and constant covariance, with only two parameters estimated, σ^2 and σ_1. This is shown in Equation 3.6.

$$\mathbf{G}_{\text{exch}} = Var(u_i) = \begin{bmatrix} \sigma^2 & \sigma_1 \\ \sigma_1 & \sigma^2 \end{bmatrix} \tag{3.6}$$

The unstructured covariance structure, shown in Equation 3.7, allows for unique variances to be estimated for each specified random effect. Covariance is estimated for each pair of random effects. For a two-level random intercept and slope model this results in three estimable covariance parameters, $\sigma^2_{u_0}, \sigma_{u_0,u_1}$, and $\sigma^2_{u_1}$.

$$\mathbf{G}_{\text{unst}} = Var(u_i) = \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_0,u_1} \\ \sigma_{u_0,u_1} & \sigma^2_{u_1} \end{bmatrix} \tag{3.7}$$

Finally the autoregressive structure is shown in Equation 3.8. This assumes homogeneous variances for all random effects, however the covariance between each additional pair of random effects tends becomes smaller still than the previous pair. Two parameters are estimated, the variance σ^2 and correlation ρ. If

an additional random effect was included the covariance would be $\rho^2\sigma^2$ and so on for each addition.

$$\mathbf{G}_{\text{ar}} = Var(u_i) = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \tag{3.8}$$

For applications in this thesis, the unstructured covariance structure will be assumed, as this allows unique variances to be estimated for all random effects. For the level-1 errors e_{ij}, an independent structure is usually assumed, with a common variance σ_e^2 and covariance set to 0.

3.5.2 Model assumptions

The linear mixed effects model framework is based on several assumptions which must be satisfied for the resulting estimates to be valid. These assumptions build on those for the linear regression model, but take into account the addition of the random effects [128]. The most important requirement when fitting an LMM or indeed any class of model is to correctly specify the covariates in the model and to have an adequate sample size for inferences to be made from the fitted model [128]. Further assumptions of the LMM include [129],

- a linear relationship between the response and explanatory variables
- residuals at each level are independent
- residuals at each level are normally distributed
- the residuals have constant variance or are homoscedastic

Though a linear relationship between the response and covariates is assumed, it is possible to relax this requirement by including higher order terms to capture non-linear associations between the response and predictor variables [130]. Further assumptions are based around the residuals of the fitted model, rather than the response variable. Visual examination of residuals at each level of the fitted

LMM can usually confirm whether these assumptions have been met or not. If patterns are noted then there may be indication of a lack of independence or existence of heteroscedastic errors [131]. Assumptions of normality can be assessed through normal probability plots by plotting standardised residuals for each level against the expected normal probability score [132].

Several studies have looked at the effect of the violation of these assumptions on model estimates. McCulloch and Neuhaus [133] and Jacqmin-Gadda *et al.* [132] investigated the effect of misspecifying the shape of the random-effects error distribution, i.e. violating the normality assumption. Fixed effect parameter estimates were found to be unbiased for both random intercept and random intercept-slope models, even when the error was highly skewed [132]. Furthermore, estimates of random effect variances were robust to deviations from normality [133]. According to Maas and Hox [134] if level-two residuals are not normally distributed this is most likely to impact only the standard error estimates, which can be remedied by calculating robust standard errors. The estimate of the intercept, however, can be biased if the normality assumption for the level-one residuals or random intercept is not satisfied [133; 135]. Fixed effect estimates have been shown to be robust in the presence of heteroscedastic errors, except when there is a dependency between the error variance and a covariate-time interaction term in the model [132]. In general, provided that the sample size is large enough, the LMM is robust to deviations from the assumptions made [136]. However, if normality and homoscedasticity assumptions are violated, then an alternative specification for the residuals can be assumed [129].

3.5.3 Estimation

Estimation of these models is usually by maximum likelihood (ML) or restricted maximum likelihood (REML). The former includes both the fixed effects and

variance components in the likelihood estimation. REML is a two step procedure, which omits the dependency on the fixed effects including only the variance components in the initial likelihood function before estimating the fixed effects in a second step [131]. The expectation and variance of the biomarker response $\mathbf{Y}_i$ are given by,

$$E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta} \tag{3.9}$$

$$\mathbf{V}_i = \text{var}(\mathbf{Y}_i) = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T + \sigma_e^2\mathbf{I}_N \tag{3.10}$$

$\mathbf{X}_i$ is the design matrix of fixed effects with associated parameter estimates $\boldsymbol{\beta}$. The design matrix of random effects is denoted by $\mathbf{Z}_i$ and $\mathbf{Z}_i^T$ is the transpose of $\mathbf{Z}_i$. $\mathbf{G}$ is the variance-covariance matrix for the random effects, $\sigma_e^2\mathbf{I}_N$ the identity matrix for the within-subject variance.

Let $\boldsymbol{\theta}$ collectively denote the covariance parameters for $\mathbf{G}$ and $\sigma_e^2\mathbf{I}_N$. The full likelihood is written as the product,

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{i=1}^{p}(2\pi)^{-\frac{n}{2}}|\mathbf{V}_i|^{-\frac{1}{2}}\exp\left\{\frac{1}{2}(y_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(y_i - \mathbf{X}_i\boldsymbol{\beta})\right\} \tag{3.11}$$

The resulting log likelihood is expressed in Equation 3.12

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{n}{2}\ln 2\pi - \frac{1}{2}\sum_{i=1}^{p}\ln|\mathbf{V}_i| - \frac{1}{2}\sum_{i=1}^{p}(y_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(y_i - \mathbf{X}_i\boldsymbol{\beta}) \tag{3.12}$$

Parameter estimates for $\boldsymbol{\beta}$ can be obtained by assuming the covariance parameters $\boldsymbol{\theta}$ are known. Using a generalised least squares approach the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ is shown in Equation 3.13 [137][138]. This estimator minimises the sampling variance (best), and is unbiased such that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i)^{-1} \mathbf{X}_i^T \mathbf{V}_i^{-1} y_i \tag{3.13}$$

In order to obtain estimates for the covariance parameters $\boldsymbol{\theta}$ the BLUE of $\boldsymbol{\beta}$ ($\hat{\boldsymbol{\beta}}$) are substituted into the log likelihood in place of $\boldsymbol{\beta}$, to form a profile log likelihood.

$$\ln L_{\mathrm{ML}}(\boldsymbol{\theta}) = -\frac{n}{2} \ln 2\pi - \frac{1}{2} \sum_{i=1}^{p} \ln |V_i| - \frac{1}{2} \sum_{i=1}^{p} (r_i^T \mathbf{V}_i^{-1} r_i) \tag{3.14}$$

where $r_i = y_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}$. There is no closed form solution to this profile likelihood and so numerical approximation methods are utilised. One approach is to use a combination of the Expectation-Maximisation (EM) algorithm and the Newton-Raphson method[125; 139]. The former is an iterative process which assumes that the random effects are unobserved data. The EM algorithm first calculates values $\boldsymbol{\theta}^w$ for the unknown vector $\boldsymbol{\theta}$ and uses this to obtain the expectation of the log likelihood for the subsequent value of $\boldsymbol{\theta}$ conditional on the distribution $\boldsymbol{\beta} \mid y$. The expectation is then maximised with respect to $\boldsymbol{\theta}$ to obtain the next value of the iteration $\boldsymbol{\theta}^{w+1}$. The Newton-Raphson algorithm, discussed in section 4.6.4, is an optimisation procedure which minimises/maximises the score function, or first derivative of the profile likelihood at the current estimate of $\boldsymbol{\theta}^w$ to obtain the next estimate. The matrix of second derivatives or Hessian in conjunction with the score function provide are used to calculate a direction vector. A new estimate is derived from this information. This is repeated until convergence criterion are

met, which means that the root has been found to a pre-specified acceptable level of tolerance [140].

3.5.4 Restricted maximum likelihood estimation

A criticism of utilising ML estimation for linear mixed effects models is that when estimating the variance components via ML, the degrees of freedom lost from estimating the fixed effect parameters are not taken into account [139]. This can lead to biased parameter estimates. An alternative proposition is estimation by restricted maximum likelihood (REML) [141]. This method was first introduced in 1962 and is based on forming linear combinations for Y which do not depend on the fixed effects [142]. These 'linear contrasts' are equivalent to the residuals which can be estimated from fitting only the fixed effects model. The corresponding log likelihood function is written as,

$$\ln L_{\text{REML}}(\boldsymbol{\theta}) = -\frac{n-m}{2}\ln 2\pi - \frac{1}{2}\sum_{i=1}^{p}\ln|\mathbf{V}_i| - \frac{1}{2}\sum_{i=1}^{p}(r_i^T\mathbf{V}_i^{-1}r_i) - \frac{1}{2}\sum_{i=1}^{p}\ln|\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i| \tag{3.15}$$

where $r_i = y - \mathbf{X}\hat{\boldsymbol{\beta}}$. Analyses in Chapter 6 and 7 will rely on maximum likelihood estimation for obtaining parameter estimates, on the assumption that the sample size is large enough that ML and REML estimates will be similar. When comparing models it is preferred that ML estimation be used to allow comparison using likelihood-based methods such as the AIC, BIC or likelihood ratio tests for nested models [143].

3.5.5 Subject-specific predictions

The greatest motivation for utilising mixed effects models is to estimate subject-specific effects. These models allow inferences on the population level but also on

the individual level with some further estimation. Individualised predictions can be obtained by 'estimating' individual deviations, via the empirical Bayes estimation of the best linear unbiased predictor (BLUP) of the random effects $\mathbf{u}_i$ [144]. Predictions are obtained by melding Bayesian and frequentist principles. The method assumes the model parameters, the fixed effect estimates and variance-components are known and equivalent to the ML estimates which are denoted by $\hat{\boldsymbol{\theta}}$ [145] [146]. Due to this assumption the posterior distribution is known as 'empirical.' The prior distribution, $\phi(\mathbf{u}_i; \mathbf{G})$, for the $\mathbf{u}_i$ is taken to be the multivariate normal distribution with covariance matrix $\mathbf{G}$ previously expressed in Equation 3.1. The likelihood is obtained by fitting the relevant LMM. The posterior density of $\mathbf{u}_i$ conditional on observing $Y_i = y_i$, is obtained via Bayes' Theorem [145]. The posterior distribution of $\mathbf{u}_i$ conditional on the data is given by,

$$w(\mathbf{u}_i \mid y_i, \mathbf{X}_i, \mathbf{Z}_i; \hat{\boldsymbol{\theta}}) = \frac{\phi(\mathbf{u}_i; \mathbf{G}) f(y_i \mid \mathbf{u}_i, \mathbf{X}_i, \mathbf{Z}_i; \hat{\theta}^f)}{g(y_i \mid \mathbf{X}_i, \mathbf{Z}_i; \hat{\boldsymbol{\theta}})} \tag{3.16}$$

The vector of parameters appearing in the conditional response distribution is denoted by $\hat{\boldsymbol{\theta}}^f$. The denominator is the likelihood contribution for the i^{th} individual. The optimal empirical Bayes estimate of the BLUPs is given by the mean of the empirical posterior distribution shown in Equation 3.17. Alternatively, the empirical posterior distribution mode can be used instead of the mean. This can be obtained by finding the mode which minimises the posterior expectation of the loss function in Equation 3.18. For linear models the integral in Equation 3.17 can be solved analytically, otherwise estimation is achieved via numerical integration such as adaptive quadrature or simulation methods (e.g. Monte Carlo Markov chain (MCMC)) [145] [147].

$$\hat{\mathbf{u}}_\mathbf{i} = E(\mathbf{u}_i \mid y_i, \mathbf{X}_i, \mathbf{Z}_i; \hat{\boldsymbol{\theta}}) = \int \mathbf{u}_i w\left(\mathbf{u}_i \mid y_i, \mathbf{X}_i, \mathbf{Z}_i; \hat{\boldsymbol{\theta}}\right) du_i \tag{3.17}$$

$$L(u_i, \hat{u}_i) = \begin{cases} 0 & \text{if} \quad |u_i - \hat{u}_i| \leq \epsilon \\ 1 & \text{if} \quad |u_i - \hat{u}_i| > \epsilon \end{cases} \tag{3.18}$$

Estimates of subject-specific residuals are also known as shrinkage estimates [131]. The BLUPs are shrunk towards the overall mean, depending on how reliably they are estimated [131]. If there are fewer observations for a given individual they will be shrunk more [145]. Similarly a BLUP which is far away from the population mean will be deemed less reliable and will shrink more again [131]. This shrinkage introduces bias, which is a criticism levelled at the empirical Bayes method of estimating the BLUPs [131]. The effect of substituting the estimates of parameters in the empirical Bayes predictor to obtain the BLUPs however is small when the sample size is large [145].

An advantage of using this framework is the ease with which standard errors for the realized values of $\mathbf{u}_i$ can be obtained [145]. The corresponding posterior covariance matrix of random effects is written as,

$$cov_y(\hat{u}_i^{EB} - u_i \mid \mathbf{X}_i, \mathbf{Z}_i; \hat{\boldsymbol{\theta}}) = \int (\hat{u}_i^{EB} - u_i)(\hat{u}_i^{EB} - u_i)' w(u_i \mid y_i, \mathbf{X}_i, \mathbf{Z}_i; \hat{\boldsymbol{\theta}}) du_i \tag{3.19}$$

When $\boldsymbol{\theta}$ are assumed known, the variances ($V(\mathbf{u}_i \mid, y_i, \mathbf{X}_i; \boldsymbol{\theta})$) are akin to the conditional mean-squared error of prediction. Assuming normality of the posterior distribution, corresponding credible intervals can be computed using the posterior mean and standard deviation [145]. The integral is calculated numerically using, for example, adaptive quadrature.

3.5.6 Modelling

Several considerations must be made when modelling continuous data. Although the response does not necessarily have to be normally distributed (see section 3.5.2), to allow for models to converge, transforming skewed biomarker data can be beneficial. Standardising or centering (around the mean) explanatory variables to allow for meaningful interpretation is also recommended [131]. Selecting the most appropriate variables for inclusion and a suitable functional form can be difficult. Stepwise model selection procedures are the most commonly used to select covariates for inclusion when modelling. For longitudinal data, Snijders and Bosker [129] advise fitting a random effects ANOVA, ignoring explanatory variables, to establish within- and between-group variances. From there it is sensible to build the model from level one, first explaining within-group variation before tackling between-group variation [124]. For models fitted by maximum likelihood estimation, likelihood ratio tests can be used to compare nested models including either or both fixed and random effects. Differences in only the random effects can be compared for models estimated using the REML, but only if they contain the same fixed effects [129] [131]. Establishing the correct fixed and random effects is essential. In fact, failure to include a necessary random slope can lead unrealistically small standard error estimates [135].

3.5.7 Interpreting model estimates

Consider a random intercept and slope model for the biomarker log billirubin. The Stata output for the fitted LMM is given below. The mean response is estimated at time t. At $t = 0$, the average log billirubin for the population is 0.496, given by the fixed intercept term. The overall fixed effect at time t is estimated via $0.496 + 0.177 \times t$, so that at $t = 5$ the mean response is

$0.496 + 0.177 \times 5 = 1.381$. The variation is captured by the random-effect parameters. For the unstructured variance-covariance matrix a variance parameter is estimated for each random effect included in the model, the intercept (`var(_cons)`) and slope (`var(time)`), as well as a covariance parameter (`cov(time, _cons)`). This describes the between-subject variation. The confidence intervals for the random intercept and slope variances indicate there is evidence of variation in log billirubin intercepts and slopes between individuals. This justifies the inclusion of each of the random effects. The measurement error or within-subject variation is estimated by `var(Residual)`.

```
. *Fit LMM
. mixed logb time || id: time, cov(unstr)
```

logb	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
time	.17742	.0123806	14.33	0.000	.1531544	.2016855
_cons	.4957678	.0579801	8.55	0.000	.3821288	.6094067

Random-effects parameters	Estimate	Std. err.	[95% conf.	interval]
id: Unstructured				
var(time)	.029277	.0040436	.0223338	.0383786
var(_cons)	.9946514	.0847627	.8416524	1.175463
cov(time,_cons)	.0715485	.0151061	.041941	.101156
var(Residual)	.1218071	.004719	.1129004	.1314164

LR test vs. linear model: chi2(3) = 2870.97 Prob > chi2 = 0.0000

The mean response can be estimated at different timepoints by summing over the fixed effects, as shown. By obtaining BLUPs for the random effects, subject-specific intercepts and slopes can be calculated. Combining the fixed effects with the predictions of the random effects, subject-specific log billirubin predictions for time t.

```
. *Calculate fixed effects at t=0 and t=5
. dis _b[_cons] + _b[time]*0
.49576775

.
. dis  _b[_cons] + _b[time]*5
1.3828676

.
. * Predict subject-specific deviations for the slope and intercept
. predict b1 b0, reffects

. gen fitted = (_b[_cons] + b0) + (_b[time] + b1)*time

.
. list id time logb fixed fitted if id ==2
```

	id	time	logb	fixed	fitted
3.	2	0	.0953102	.4957677	.0073679
4.	2	.498302	-.2231435	.5841765	.0991779
5.	2	.999343	0	.6730711	.1914925
6.	2	2.10273	.6418539	.8688335	.3947862
7.	2	4.90089	.9555114	1.365283	.9103352
8.	2	5.88928	1.280934	1.540643	1.092442
9.	2	6.88588	1.435084	1.717461	1.276062
10.	2	7.8907	1.280934	1.895736	1.461196
11.	2	8.83255	1.526056	2.062838	1.634727

```
.
```

3.6 Generalised estimating equations

Though LMMs for continuous responses will form the basis of longitudinal modelling in this thesis, the following method for clustered or longitudinal data is presented here for completeness.

Generalised estimating equations (GEEs) are also commonly used to model longitudinally observed data, particularly when interest lies solely in estimating the marginal effects, i.e. a mean response for groups of individuals who share a covariate pattern[148]. GEEs are an extension of the generalised linear model (GLM), but make no distributional assumptions [149]. Firstly GEEs build on the

GLM by allowing the within-subject correlation to be estimated, and secondly they allow cluster robust standard errors to be estimated for the regression coefficients. The set of GEEs for a vector of responses $\mathbf{y_i}$ for the i^{th} individual is shown in Equation 3.20 [148]. $\mu_\mathbf{i}$ represents the expectation or mean of $\mathbf{y_i}$.

$$\sum_{i=1}^{N} \left(\frac{d\mu_\mathbf{i}}{d\beta} \right) \mathbf{V}_i^{-1} (\mathbf{y_i} - \mu_\mathbf{i}) = 0 \tag{3.20}$$

The correlation is incorporated by specifying a working correlation matrix $\mathbf{R}_i(\alpha)$ to describe the general correlation structure. These can have similar specifications to the variance-covariance matrix for LMMs discussed in section 3.5. Popular definitions include the exchangeable, independent, autoregressive, or unstructured.[150] Using this correlation a covariance matrix, $\mathbf{V}_i$, can be estimated for y_i.

$$\mathbf{V}_i = \frac{\mathbf{A}_i^{1/2} \mathbf{R}_i(\alpha) \mathbf{A}_i^{1/2}}{\phi} \tag{3.21}$$

Here $\mathbf{A}_i$ is a n_i x n_i diagonal matrix with some function of the mean, $g(\mu_i)$, as the i^{th} diagonal element. A quasi-likelihood approach is used to solve the set of GEEs. The iterative procedure first fits a standard GLM to estimate an initial value for β assuming independence. The β estimates are then used to obtain estimates for α and ϕ, given the correlation structure assumed for the matrix $\mathbf{R}$ standardised residuals. A subsequent estimate for the variance $\mathbf{V}_i$ is computed and the estimate for β is updated. These steps are repeated until convergence is reached [148].

GEEs are relatively simple to fit compared to models relying on ML estimation, particularly as they make no underlying distributional assumptions. As a result they can also model non-continuous response variables. Furthermore, even

if the working correlation matrix is misspecified estimates are consistent and robust standard errors can be estimated [150]. However, as noted previously these are population-average models and so for the purposes of predicting individual effects they are unsuitable. To compound matters model selection cannot be carried out using usual likelihood methods as GEEs are estimated using a quasi-likelihood approach. As subject-specific longitudinal trajectories of hCG are an integral aspect of the application to pregnancy outcomes, these models will not be considered further in this thesis.

3.7 Discussion

Foundational methods used to analyse longitudinally measured data have been presented in this chapter. The linear mixed effects model will feature as the now standard longitudinal submodel of the joint model which will be introduced in Chapter 5. For each analysis conducted in Chapters 6 and 7 a longitudinal submodel will be built, giving consideration to how the biomarker should be modelled over time and how it varies between individuals. Ultimately random effects and corresponding subject-specific predictions, obtainable from the mixed model framework, allow for individualised predictions of survival probabilities when linked with a survival outcome in a joint model framework.

Chapter 4

Survival analysis

4.1 Chapter overview

In this chapter the key concepts of survival analysis will be introduced. The hazard and survival functions will be defined and the non-parametric Kaplan-Meier estimation procedure will be presented. The focus will be on the Cox proportional hazards (PH) model and the Weibull and flexible parametric models, which are all implemented in this thesis, both in isolation and as the survival component model of the joint longitudinal-survival model.

4.2 Introduction

Survival analysis is known by various aliases including time-to-event or failure time analysis. All of these names refer to analysis of data where individuals are observed from a pre-defined origin, up to the time at which the event of interest occurs, or the individual is lost to follow-up [151]. Often the event of interest is mortality however other common endpoints are tumour or disease progression, relapse and recurrence of symptoms or adverse events [152; 153]. When modelling a time to an event, interest lies in the rate at which an event occurs as a function of time, as opposed to a cross-sectional probability of event occurrence, as with logistic regression. Survival data is usually skewed, so methods relying on the normality assumption can lead to inappropriate inferences [154]. Typically when

observing specific disease populations for a pre-determined period of time, not all individuals will experience an event by the end of follow-up [155]. This key characteristic of survival data is known as censoring and must be accounted for in the analysis.

4.3 Censoring

During the course of follow-up not everyone will experience the event of interest. In a clinical study, a person may (i) not experience an event before the end of follow-up, (ii) may be lost to follow-up or (iii) may withdraw from the study [156]. Where this occurs, and the reasons behind each are unrelated to survival, the only information available is the time at which the individual was last event-free. As the true survival time is unknown, the individual is said to be censored at their last observation time and is withdrawn from the risk set [155]. Typically this is referred to as right censoring as the event, if it does occur, happens to the right of the last point at which the individual was known to be event-free [151; 154]. Hypothetical survival data is shown in Figure 4.1. The survival time for those who experienced an event is denoted by a filled red circle, whilst the corresponding survival times for those patients who were censored are denoted by an open green circle.

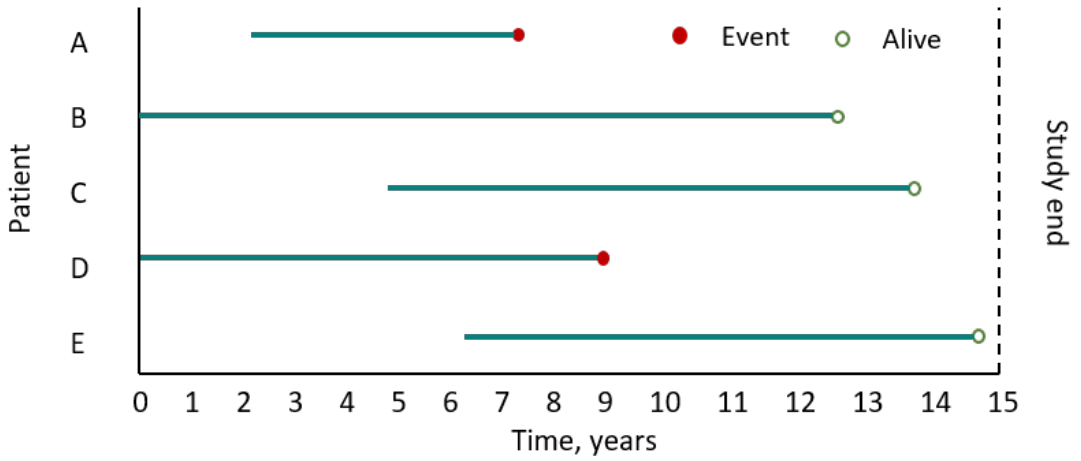


FIGURE 4.1. Hypothetical dataset of survival times

4.3.1 Assumptions

A key assumption made when censoring individuals is that the true survival time t is independent of the underlying censoring process [156]. A censored individual, subject to the same covariate pattern, is assumed to be representative (have the same survival probability) of all other individuals who remain in the risk set at the time of censoring. This concept is called non-informative censoring. If the reason for the censoring is related to the status of the individual then the censoring is instead informative and the assumption is no longer valid [155].

To illustrate this, let us consider a 50 year old female patient who has joined a study which follows up individuals until failure (death). Under the assumption of non-informative censoring, if she withdrew at time t because she moved away, then we assume that her risk of experiencing an event at the time of censoring is the same as a 50 year old female who continues in the study. If, on the other hand, she withdrew due to poor health which in turn makes her more likely to experience an event than another 50 year old women censored at time t, then this

indicates informative censoring [156]. As informative censoring indicates an alternative failure/risk profile for certain individuals this has added implications for analysis, leading to biased estimates if not addressed [157; 158]. For the analyses in Chapters 6, 7 and 8 it will be assumed that censoring is non-informative.

4.3.2 Left and interval censoring

Other types of censoring, which will not be considered further here, include left and interval censoring [151; 155; 159]. Left censoring is the antithesis to right censoring, where the event is known to occur before or to the left of the survival time [159]. For example consider the scenario of pregnancy loss. If at the first contact with a healthcare professional there is no detectable heartbeat, it is clear that the loss has occurred at a time prior to this initial appointment. Interval censoring is as described, when an event is known to have occurred between two time points [155]. In the pregnancy loss scenario this could mean that the pregnancy was progressing at the initial scan appointment, yet at the second detailed scan appointment there are no signs of viability. The event is known to have occurred in the interval between the two scan appointments, and so is interval censored.

4.3.3 Delayed entry

Typically a patient enters a study if they meet the inclusion criteria, which in the case of a time-to-event outcome, means they must be at risk of the event of interest. There may be occasions where a patient does not become at risk of the event until after the specified time origin, so entry is at a time $t > 0$. This is referred to as delayed entry and results in left-truncated data. A common example of this is when age is used as the timescale, as an alternative way for adjusting for age [160]. Consider an epidemiological study following an elderly population of 65

years or over. Utilising age as a timescale, as opposed to time under observation gives a meaningful counterpoint to study ageing [161]. As individuals enter if they are over the specified age, data pre-65 years is left-truncated. For accurate survival times to be calculated this must be incorporated into the analysis [162]. Left truncation will feature prominently in analyses of pregnancy data detailed in Chapters 6, 7 and 8. This is due to the selection of a timeline for which women do not meet the conditions to be at risk at $t = 0$. Namely pregnancy cannot be confirmed until a time after conception, and so women cannot be at risk of miscarriage until this point.

4.4 Hazard and survival functions

Survival data is usually described in terms of the hazard and survival functions. These quantify a patient's potential of experiencing an event and the probability of surviving an event, respectively. Let T be a continuous non-negative random variable signifying the survival times and t a specific time of interest. Then the cumulative probability density function, denoted by $F(t)$, is the probability that the survival time for an individual is less than t. $F(t)$ is defined in Equation 4.1, where $f(t)$ denotes the probability density function of T [156]. Substituting in $t = 5$ (years) into Equation 4.1 would give the probability that an individual experiences an event within 5 years, or in other words that their survival time T is less than 5 years.

$$F(t) = P(T < t) = \int_0^t f(u)du \tag{4.1}$$

The survival function $S(t)$ indicates the probability that a patient does not experience the event up to time t, or the probability that their survival time T is greater than or equal to t [151]. It is written in terms of the cumulative density

function $F(t)$, as shown in Equation 4.2. Here, for $t = 5$, $S(t)$ would be the probability that the patient survives for 5 or more years.

$$S(t) = P(T \geq t) = 1 - F(t) \tag{4.2}$$

The hazard function is a 'conditional failure rate' and is given by Equation 4.3 [156]. The probability of an individual experiencing an event within a time interval t and $t + \delta t$, given they have already survived up to time t is first calculated. This quantity is then divided by the length of the time interval, denoted by δt. As this interval tends to zero the hazard function $h(t)$ gives the instantaneous failure rate at time t, given that an individual has survived up to time t [151].

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\} \tag{4.3}$$

Integrating the hazard function up to time t gives the cumulative hazard function $H(t)$ shown in Equation 4.4. This describes the cumulative event rate up to time t [156].

$$H(t) = \int_0^t h(u)du = -\log\left(S(t)\right) \tag{4.4}$$

The hazard function is intrinsically linked to the survival function and each can be written solely in terms of the other. Equation 4.5 represents the hazard function written in terms of the survival function and Equation 4.6 denotes the survival functions parameterised in terms of the hazard function.

$$h(t) = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}\log S(t) \tag{4.5}$$

$$S(t) = \exp\left[-\int_0^t h(u)du\right] \tag{4.6}$$

The survival and hazard functions are fundamental to survival analysis, as these are predominately used to communicate risk and the event rate. These functions will be discussed further in the context of survival models in later sections.

4.5 Kaplan-Meier Estimator

Statistical methods for survival analysis broadly fall into three categories - non-parametric, semi-parametric and parametric. The Kaplan-Meier estimator is the classic method for non-parametric estimation of the survival function, namely without assuming an underlying distributional form [163]. This is a product-limit estimator of the survival function $S(t)$ and is shown in Equation 4.7 [164]. The number of individuals (n_j) still in the risk set just before each event time t_j is calculated, along with the number of deaths (d_j) at each t_j. The proportion who are still event-free at each time t_j is computed giving a survival probability estimate. Survival probabilities are then updated at each event time t_j by taking the product of the current survival probability and all preceding probabilities [151]. The survival function is only updated when an event occurs and remains constant otherwise, resulting in a step function [151; 165].

$$\hat{S}(t_j) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j=1}^{k} \left(1 - \frac{d_j}{n_j} \right) \tag{4.7}$$

Several assumptions are made when estimating the Kaplan-Meier estimate. Censored observations are assumed to be independent of the event times and non-informative, resulting in censored observations which only contribute to the denominator presented in Equation 4.7 and never the numerator [165; 166]. At times where an event and censored observation are recorded it is assumed that the

event occurs just before the censored observation [151]. Furthermore, individuals who experience events are assumed to be independent of each other [151].

The corresponding standard error for $\hat{S}(t)$ is approximated using the delta method via Greenwood's formula [151]. This is given in Equation 4.8.

$$se\left\{\hat{S}(t)\right\} \approx \hat{S}(t)\left\{\sum_{j=1}^{k}\frac{d_j}{n_j(n_j-d_j)}\right\}^{\frac{1}{2}} \tag{4.8}$$

Obtaining confidence intervals for $\hat{S}(t)$ at time t requires a normal approximation, with mean $\hat{S}(t)$ and variance $\left[se\left\{\hat{S}(t)\right\}\right]^2$ [151; 166]. However, this poses a problem when $\hat{S}(t)$ is close to 0 or 1, as the upper and lower limits of the confidence interval may exceed the allowed range of $[0, 1]$ [151; 166]. To overcome this, the normal distribution is applied to a transformation of $\hat{S}(t)$. The suggested transformation is $\hat{v}(t) = ln\left[-ln\hat{S}(t)\right]$ [166]. The corresponding standard error for $\hat{v}(t)$ is given in Equation 4.9.

$$se\left[\hat{v}(t)\right] \approx \frac{1}{\ln\hat{S}(t)}\left\{\sum_{j=1}^{k}\frac{d_j}{n_j(n_j-d_j)}\right\}^{\frac{1}{2}} \tag{4.9}$$

The 95% confidence intervals for $\hat{S}(t)$ are obtained by back-transforming to Equation 4.10 [151].

$$\hat{S}(t)^{\exp[\pm 1.96 se\{\hat{v}(t)\}]} \tag{4.10}$$

To demonstrate the calculation of the Kaplan-Meier estimate of the survival function consider the following set of survival times representing 20 patients shown in Table 4.1. The symbol $^+$ represents an individual who was right- censored at the year indicated. Conversely those without the symbol experienced an event. Corresponding Kaplan-Meier survival estimates are shown in Table 4.2.

TABLE 4.1. Hypothetical survival times for 20 patients

	1	2	3	3^+	4	5	5^+	5	6	6^+
	7	8^+	8	9	9^+	10^+	11^+	11	13^+	15^+

$^+$ individual was censored

TABLE 4.2. Kaplan-Meier estimates for hypothetical survival data for 20 patients

t	n_j	d_j	$\frac{n_j - d_j}{n_j}$	$\hat{S}(t)$	Confidence Interval
0	20	0	1.0000	1.0000	1.0000, 1.0000
1	20	1	0.9500	0.9500	0.6947, 0.9928
2	19	1	0.9474	0.9000	0.6560, 0.9740
3	18	1	0.9444	0.8500	0.6038, 0.9490
4	16	1	0.9375	0.7969	0.5448, 0.9186
5	15	2	0.8667	0.6906	0.4361, 0.8478
6	12	1	0.9167	0.6331	0.3789, 0.8062
7	10	1	0.9000	0.5698	0.3172, 0.7591
8	9	1	0.8889	0.5065	0.2609, 0.7086
9	7	1	0.8571	0.4341	0.1981, 0.6505
10	5	0	1.0000	0.4341	0.1981, 0.6505
11	4	1	0.7500	0.3256	0.1032, 0.5744
13	2	0	1.0000	0.3256	0.1032, 0.5744
15	1	0	1.0000	0.3256	0.1032, 0.5744

Figure 4.2 shows the plotted step function for the Kaplan-Meier survival function for the hypothetical data, and associated 95% confidence interval.

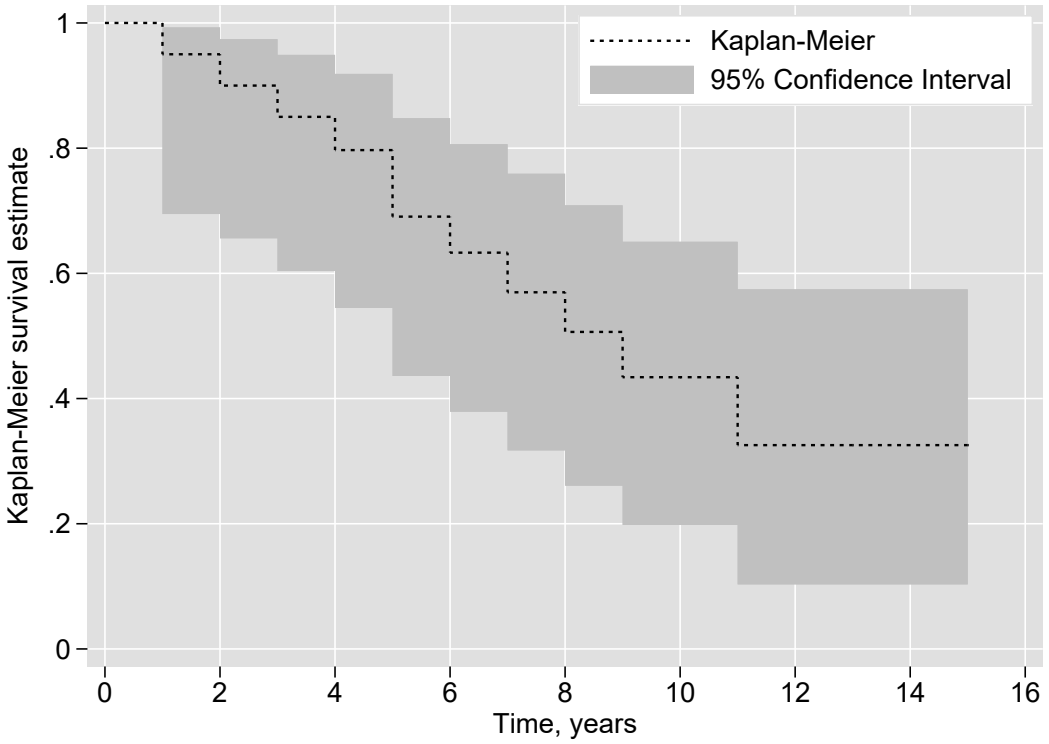


FIGURE 4.2. Plot of Kaplan-Meier survival estimates computed from hypothetical survival data from 20 individuals

4.6 Modelling survival data

The Kaplan-Meier estimate is useful for giving an overall estimate of survival in a specified group, or to compare survival between groups. However, in clinical studies it is usual to collect data on patient characteristics such as age and sex and also assess their impact on survival [151]. To accomplish this a statistical model, including patient characteristic variables, must be fitted to the survival data. Survival models fit into two broad categories; models which are fitted parametrically or semi-parametrically [167]. Semi-parametric models do not make

underlying distributional assumptions for the survival times T, whilst parametric models do [167]. In this section several models will be introduced, including the Cox proportional hazards model (see section 4.6.1), which will be applied in Chapter 8; the Weibull (section 4.6.3) and flexible parametric models (section 4.6.8), which will feature in Chapters 6 and 7.

4.6.1 Cox model

The Cox model is most frequently implemented with the aim to model survival data whilst accounting for and establishing the effect of predictive factors on the outcome of interest [39; 42]. The Cox model specification of the hazard function, $h_i(t \mid \mathbf{x})$, for the i^{th} individual is presented in Equation 4.11 [168]. This is made up of the product of the baseline hazard function $h_0(t)$ and the linear predictor consisting of the vector of explanatory variables $\mathbf{x}_i$ and the corresponding vector of coefficients $\boldsymbol{\beta}$. The baseline hazard represents the underlying common hazard rate when all covariates $\mathbf{x}_i$ are set to 0, and remains unspecified [39].

$$h_i(t \mid \mathbf{x}) = h_0(t) \exp(\boldsymbol{\beta}\mathbf{x}_i) \tag{4.11}$$

The coefficients, $\boldsymbol{\beta}$, for the explanatory variables are estimated on the log scale and exponentiated to produce hazard (rate) ratios. These exponentiated estimates of $\boldsymbol{\beta}$ represent the ratio of the hazard rate for a chosen value of the covariate and the corresponding hazard rate for the baseline value of the covariate [39]. For example, for a categorical variable sex, if male is coded as the reference value 0 then the hazard ratio for sex expresses how much larger (or smaller) the event rate is for females when compared to males. This is shown in Equation 4.12.

$$HR_{ij} = \frac{h_i(t)}{h_j(t)} = \frac{h_0(t)e^{\beta_1 Female}}{h_0(t)e^{\beta_1 Male}} = e^{\beta_1} \tag{4.12}$$

4.6.2 Proportional hazards

A key assumption of the Cox model is one of proportional hazards. This explicitly requires that the ratio of hazard rates for two individuals, with fixed covariate patterns, to be constant over time [165; 168]. The assumption can be formally written as in Equation 4.13 [168]. The ratio of the hazard function for individual i and individual j can be reduced to a ratio of the linear predictor of covariates for each individual, resulting in a constant which is independent of time [165].

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)e^{\beta x_i}}{h_0(t)e^{\beta x_j}} = \frac{e^{\beta x_i}}{e^{\beta x_j}} \tag{4.13}$$

The Cox model falls into the semi-parametric class of model. The parametric portion of the model is given by the explanatory variables and the associated β coefficients, whilst no functional form is assumed for the underlying baseline hazard [167]. Estimation of log hazard ratios $\boldsymbol{\beta}$ is independent of the baseline hazard function, via maximising the partial likelihood, which is introduced in Equation 4.14 [165; 169]. Let i represent those participants at risk of event j, with $t_{(j)}$ denoting failure times and $\mathbf{x}_{(j)}$ the covariate vector for the individual experiencing an event at $t_{(j)}$. $R(t_{(j)})$ represents those still at risk up to the event time $t_{(j)}$.

$$PL(\beta) = \prod_{j=1}^{d} \frac{\exp(\boldsymbol{\beta}\mathbf{x}_{(j)})}{\sum_{i \in R(t_{(j)})} \exp(\boldsymbol{\beta}\mathbf{x}_i)} \tag{4.14}$$

The estimation of model coefficients assumes there are no ties, in that no two individuals experience an event at the same time. In the event of ties, Breslow's estimator is commonly utilised to approximate the partial likelihood [170].

4.6.3 Weibull Model

Unlike the semi-parametric Cox model, parametric survival models assume a functional form for the baseline hazard by allowing the survival times to follow a particular distribution. The advantage of modelling wholly parametrically lies in the relative ease with which predictions can be obtained through direct estimation of the baseline hazard function and without resorting to an approximation, such as the Breslow estimator [171]. This feature of parametric survival models will inform the choice of model in Chapters 6 and 7.

The most commonly assumed distribution for the survival times is the Weibull distribution as shown in Equation 4.15. This consists of scale and shape parameters as denoted by λ and γ respectively. Figure 4.3 illustrates the baseline hazard function for different values of γ. More formally the Weibull baseline hazard function can be written as in Equation 4.16 [151]. Note that when $\gamma = 1$, the baseline hazard is constant and reduces to the exponential functional form [165]. Otherwise the function is monotonic, which increases over time when $\gamma > 1$ and decreases over time when $\gamma < 1$. This in itself is a disadvantage of assuming a Weibull distribution for the survival times, as it is more likely that the hazard will change more fluidly over time [41]. Flexible parametric models, which will be introduced in 4.6.8, allow for turning points in the baseline hazard modelled using restricted cubic splines (RCS).

$$S(t) = \exp\{-\lambda t^{\gamma}\} \tag{4.15}$$

$$h_0(t) = \lambda\gamma t^{\gamma-1} \tag{4.16}$$

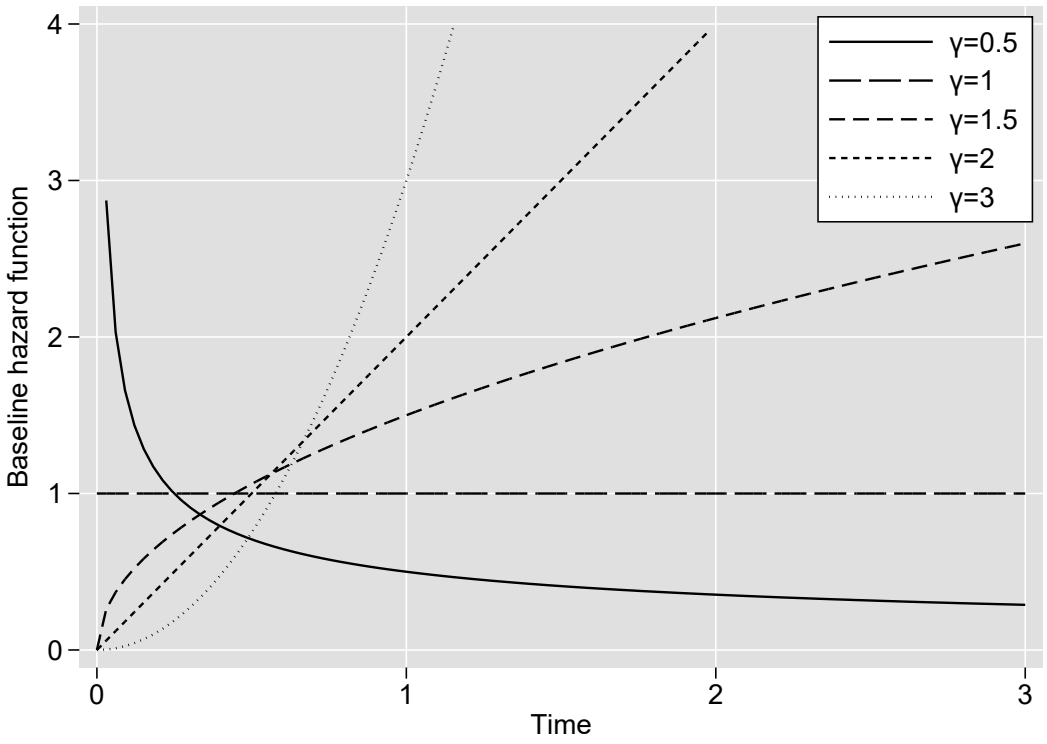


FIGURE 4.3. Examples of a Weibull baseline hazard function for $0.5 \leq \gamma \leq 3$

The Weibull hazard function for the i^{th} individual is expressed in Equation 4.17. The λ term is incorporated into the constant term of the linear predictor, so that $\beta_0 = ln(\lambda)$.

$$\begin{aligned} h_i(t) &= \lambda\gamma t^{\gamma-1} \exp\left(\beta_1 x_{1i} + \cdots + \beta_p x_{pi}\right) \\ &= \gamma t^{\gamma-1} \exp\left(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\right) \end{aligned} \tag{4.17}$$

4.6.4 Parametric model estimation

The estimation of a general parametric PH model is via maximum likelihood with respect to the unknown scale, shape and regression parameters; λ, γ and $\boldsymbol{\beta}$ [151]. Estimation in this case must account for those who are censored. As it is

not known when these individuals experience an event, the probability that they have an event after their censoring time is used to evaluate the likelihood. This is equivalent to the survival function calculated at the censoring time [151].

Equation 4.18 shows the likelihood contribution for the i^{th} individual, which is expressed in terms of the hazard and survival functions [151]. The event indicator is denoted by the term d_i, taking 1 in the case of an event at time t_i and 0 otherwise.

$$L(\boldsymbol{\beta}, \gamma, \lambda) = \prod_{i=1}^{n} [h_i(t_i)]^{d_i} \, S(t_i) \tag{4.18}$$

The corresponding log likelihood is presented in Equation 4.19 [151].

$$\begin{aligned} \ln L &= \sum_{i=1}^{n} [d_i \ln h(t_i) + \ln S(t_i)] \\ &= \sum_{i=1}^{n} [d_i \{\boldsymbol{\beta}\mathbf{x}_i + \ln \lambda\gamma + \gamma \ln t_i\} - \lambda \exp(\boldsymbol{\beta}\mathbf{x}_i) t_i^{\gamma}] \end{aligned} \tag{4.19}$$

The log-likelihood is maximised by implementing the Newton-Raphson algorithm. This method approximates the score vector (vector of first derivatives) using a linear function of each parameter[172]. The linear function consists of a first-order Taylor series approximation which is set equal to zero and solved for an estimate of the parameter. This is updated iteratively until the change in the log likelihood between iterations becomes small according to an acceptable tolerance level [172].

4.6.5 Delayed entry

Equation 4.18 assumes that individuals are at risk of the event from $t = 0$. Analyses in Chapters 6, 7 and 8, however, will assume delayed entry. The likelihood function is then adjusted as in Equation 4.20 to incorporate $S(t_{0i})$, the survival

function at the entry time t_{0i} for the i^{th} participant [173]. The corresponding log likelihood contribution is expressed in Equation 4.21.

$$L(\beta, \gamma, \lambda) = \prod_{i=1}^{n} \frac{[h_i(t_i)]^{d_i} S(t_i)}{S(t_{0i})} \tag{4.20}$$

$$\begin{aligned} \ln L &= \sum_{i=1}^{n} [d_i \ln h(t_i) + \ln S(t_i) - \ln S(t_{0i})] \\ &= \sum_{i=1}^{n} [d_i \{\boldsymbol{\beta}\mathbf{x}_i + \ln \lambda\gamma + \gamma \ln t_i\} - \lambda \exp(\boldsymbol{\beta}\mathbf{x}_i) t_i^{\gamma} + \lambda \exp(\boldsymbol{\beta}\mathbf{x}_i) t_{0i}^{\gamma}] \end{aligned} \tag{4.21}$$

4.6.6 Model comparison

Models can be compared using Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC). Modelling decisions in later chapters, particularly when considering the functional form of the baseline hazard function, will be based on the AIC. The AIC and BIC estimate how far the model deviates from the data, whilst including a penalty term for increasing model complexity [174; 175]. The BIC implements a more severe penalty when compared to the AIC. The estimation of each criterion is given in Equations 4.22 and 4.23 [176]. In each case, L refers to the likelihood, N to the number of observations or the number of events in a survival context and k the number of parameters estimated in the model.

$$AIC = -2\ln(L) + 2k \tag{4.22}$$

$$BIC = -2\ln(L) + 2\ln(N)k \tag{4.23}$$

4.6.7 Time-dependent effects

The PH assumption (see section 4.6.2) may not always be satisfied for a given variable. In these cases the assumption can be relaxed through the inclusion of a covariate interaction with time [177]. This then allows the corresponding hazard ratio for the time-dependent covariate to vary over time [39]. An example of a model with a time-dependent effect for covariate x is given by Equation 4.24. The linear predictor of the hazard function is now a function of time $g(t)$. The most commonly chosen interaction with time is $g(t) = \ln t$, particularly where very large survival times may influence the parameter estimate for the interaction [177].

$$h_i(t \mid x_i) = h_0(t) \exp\left([\beta_1 + \eta g(t)]\, x_i\right) \tag{4.24}$$

Resulting (log) hazard ratios for time-dependent effects can be plotted against time as shown in Figure 4.4. Inclusion of time-dependent effects increases computational burden as it requires the estimation of the linear predictor for all individuals at risk, at all event times [151]. Furthermore, inclusion of time-dependent effects necessitates appropriate modelling of the baseline hazard function [178].

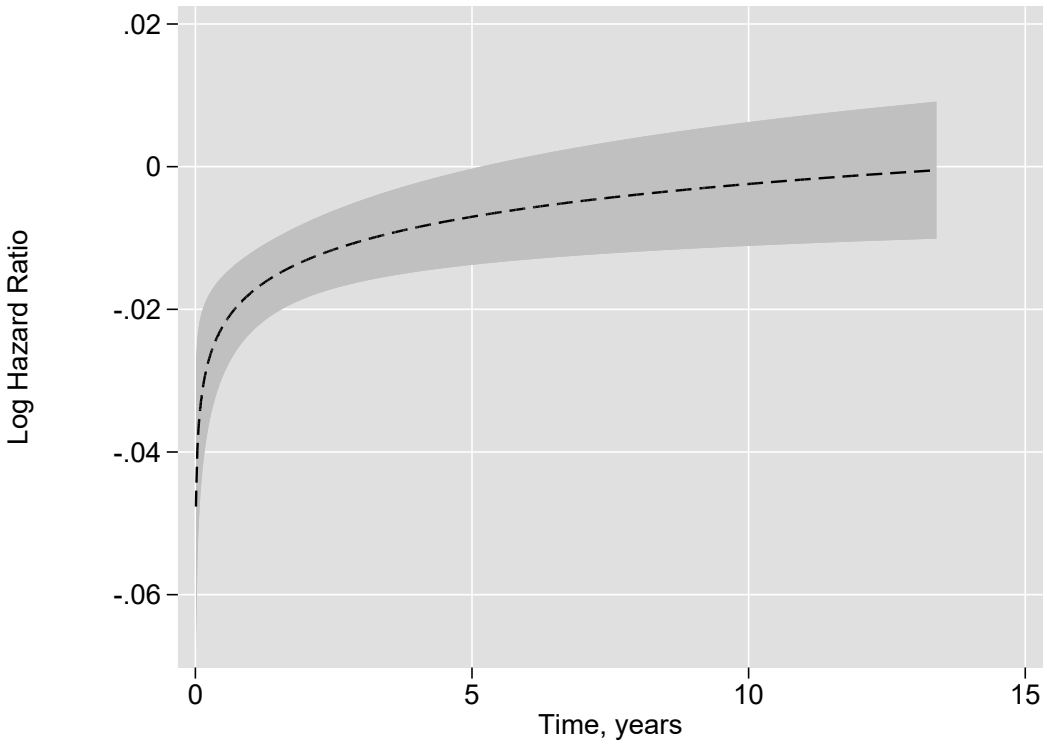


FIGURE 4.4. Plot of the log-hazard ratio for a time-dependent covariate

4.6.8 Flexible parametric models

Flexible parametric models (FPMs), first introduced by Royston and Parmar [179], are fitted parametrically with a fully specified baseline hazard [179]. Restricted cubic splines are used to model the baseline log cumulative hazard function, $\ln H_0(t)$, which allows complex non-linear hazards to be modelled [179; 180].

RCS functions are made up of piecewise polynomials, which are connected at points called knots. To ensure they are smooth, the function and its first and second derivatives are constrained to be continuous at the join points [181]. In addition they are forced to be linear before the first and after the last knot, to provide stability and sensible extrapolations [182]. Formally an RCS function, $s(x)$, for knots $k_1, \cdots, k_K$ is presented in Equation 4.25 [178].

$$s(x) = \gamma_0 + \sum_{i=1}^{K-1} \gamma_i B_i(x) \tag{4.25}$$

γ_i represents the parameter values. For K number of knots there are $K-1$ corresponding basis functions $B_i(x)$ estimates, which are defined as in Equation 4.26, where $(x-k_1)_+^3$ is 0 if the value is not positive. Furthermore, $\lambda_i = \frac{k_K - k_i}{k_K - k_1}$ [178].

$$B_i(x) = \begin{cases} x, & \text{if } i = 1 \\ (x-k_i)_+^3 - \lambda_i(x-k_1)_+^3 - (1-\lambda_i)(x-k_K)_+^3. & \text{if } i = 2, \cdots, K-1 \end{cases} \tag{4.26}$$

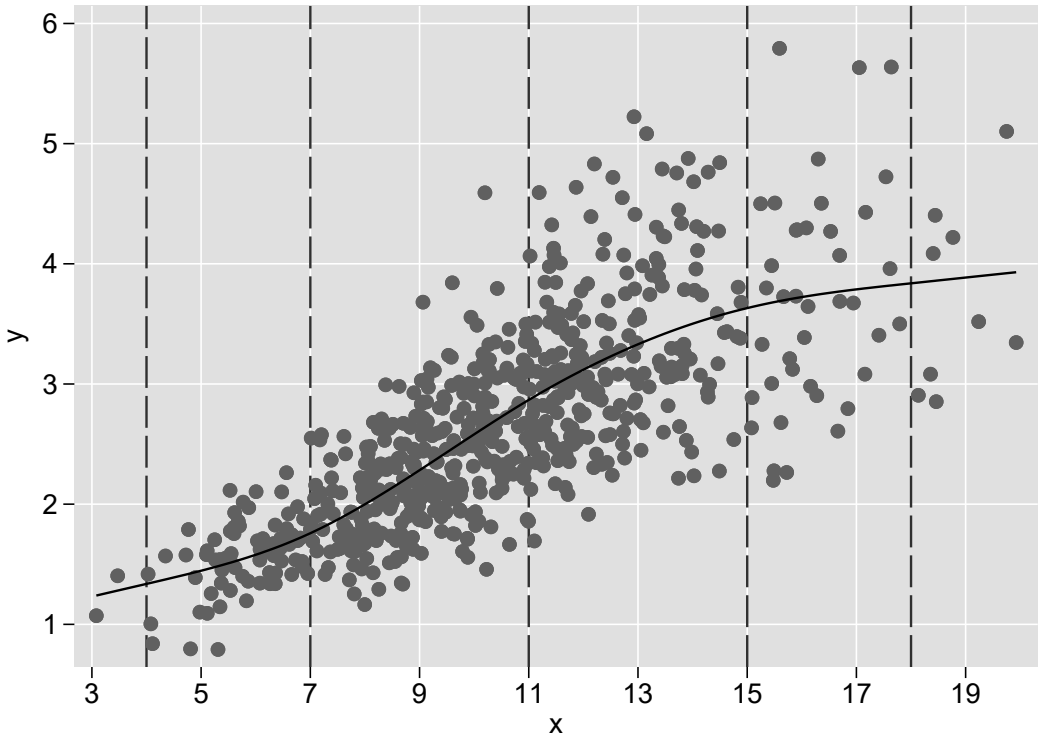


FIGURE 4.5. Example of a restricted cubic spline function with 4 degrees of freedom

An example of an RCS function with three internal knots at $x = 7, 11, 15$ and 2 boundary knots at $x = 4, 18$ is shown in Figure 4.5. The Weibull model is used as the basis for the FPM. The $\ln(t)$ term of the Weibull log cumulative hazard function, however, is instead expanded out into a cubic spline basis function [182]. Assuming proportional cumulative hazards, which follows on from assuming PH (see section 4.6.2), the FPM can be represented by Equation 4.27. This is the sum of the RCS function $s\left\{\ln\left(t\right) \mid \gamma, \mathbf{k}_0\right\}$, with knots $\mathbf{k}_0$ and the explanatory variables $\mathbf{x}_i$ and associated parameter estimates $\boldsymbol{\beta}$ [182].

$$\ln\left\{H(t \mid \mathbf{x}_i)\right\} = \ln\left[H_0(t)\right] + \mathbf{x}_i\boldsymbol{\beta} = s\left\{\ln\left(t\right) \mid \boldsymbol{\gamma}, \mathbf{k}_0\right\} + \mathbf{x}_i\boldsymbol{\beta} \tag{4.27}$$

A sufficient number of knots should be specified to capture the log baseline cumulative hazard, though consideration should be given to the location. Sensitivity analyses have shown that the number of knots chosen is of more importance than the location. In particular the hazard and survival functions are insensitive to the location of the knots [178]. By default software packages (e.g. `stpm2`) place knots at the boundaries of the uncensored log survival times, whilst internal knots are spaced evenly at the relevant number of centiles of the uncensored log survival times [41]. So for a model with three degrees of freedom two internal knots will be placed at the 33^{rd} and 66^{th} centiles. A comparison of baseline hazard and survival functions for the Weibull model and various FPMs of differing degrees of freedom are shown in Figure 4.6. As these are parametric models, survival predictions can be obtained in a straightforward manner by transforming to the survival function using Equation 4.28.

$$S(t \mid \mathbf{x}_i) = \exp\left[-\left\{H(t \mid \mathbf{x}_i)\right\}\right] \tag{4.28}$$

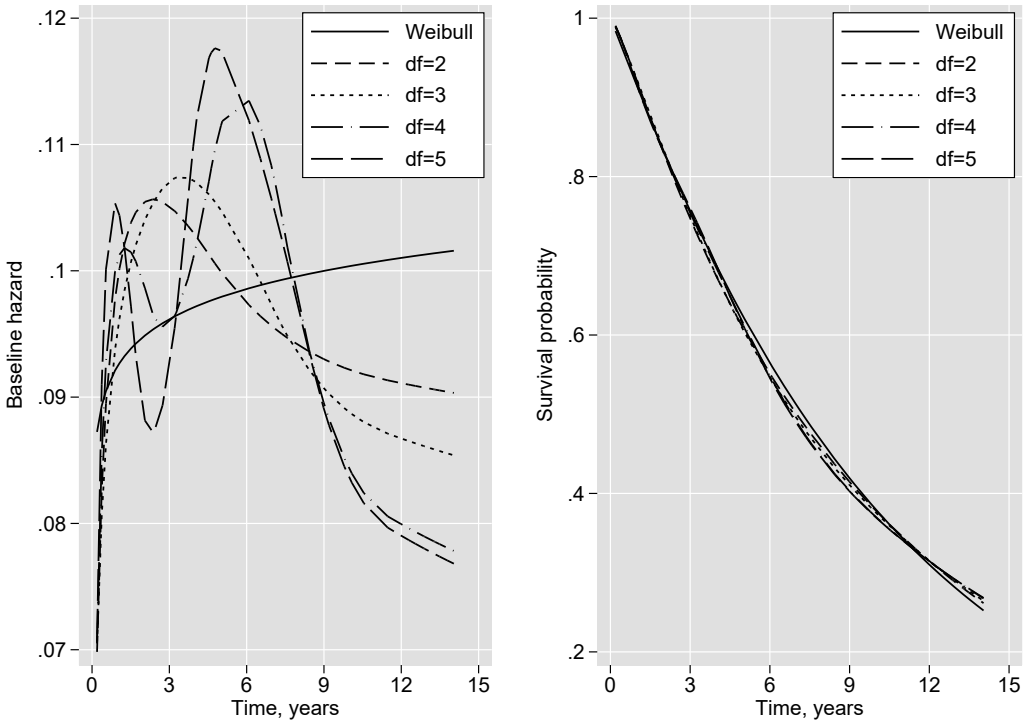


FIGURE 4.6. Baseline hazard and survival functions for the Weibull model and flexible parametric models of differing degrees of freedom

Though proportional (cumulative) hazards are assumed this can be relaxed in the FPM framework. Equation 4.29 shows the model formulation with the inclusion of time-dependent effects. These take the form of covariate (x_{ij}) interactions with spline modelled time functions $s(\ln t \mid \delta_j, \mathbf{k}_j)$.

$$\ln\left\{H_i(t \mid \mathbf{x}_i)\right\} = s\left\{\ln(t) \mid \boldsymbol{\gamma}, \mathbf{k}_0\right\} + \sum_{j=1}^{D} s\left\{\ln(t) \mid \boldsymbol{\delta}_j, \mathbf{k}_j\right\} x_{ij} + \mathbf{x}_i\boldsymbol{\beta} \qquad (4.29)$$

The FPM framework presented will be used to allow for non-linear specification of the log cumulative baseline hazard within the survival submodel of the joint longitudinal-survival model in Chapters 6 and 7.

4.7 Discussion

Methods to analyse survival data have been introduced in detail, in order to provide a grounding for applied analysis in future chapters. The Kaplan-Meier estimator will feature across Chapters 6 and 7. Parametric models, both the Weibull and the extension to the FPM will be used as the basis of the survival submodel of the joint-longitudinal survival models for time to miscarriage and longitudinal hCG. The prediction of survival probabilities is of utmost importance for these models and hence the specification of the baseline hazard plays a crucial role. The interplay between longitudinal and survival submodels and estimation of individualised predictions will be elaborated on further in Chapter 5, which will introduce joint longitudinal-survival models. Where the analysis is interested in only the effect of a specific covariate, the Cox model will be favoured over parametric models. This will be the case in Chapter 8, in which the timing of intercourse and its effect on the hazard of miscarriage will be investigated.

Chapter 5

Joint longitudinal-survival models

5.1 Chapter overview

In this chapter methods detailed in Chapters 3 and 4 will be consolidated to introduce techniques which allow the joint modelling of time-to-event and longitudinally observed data. Naive methods which have been used as a prelude to the development of the joint longitudinal-survival model will be discussed. The traditional frequentist joint longitudinal-survival model and its estimation will be explained in detail, as well as the important extensions these models contribute to the landscape of individualised predictions. Methods for joint model assessment, which extend current methods to incorporate dynamically changing observations and censoring, will be presented. Finally, alternatives to the shared random effects joint model framework for simultaneously modelling longitudinal and survival data will be introduced.

5.2 Introduction

The repeated collection of biomarker measurements alongside a survival outcome is common place in clinical practice, particularly in the case of long term conditions which require careful monitoring. Such scenarios may include glucose measurements and a diabetic event such as hypo- or hyperglycaemia or perhaps blood pressure and a cardiovascular event [33; 110]. This repeated collection has

led to the compilation of information-rich electronic databases which can provide detailed pictures of disease progression [183]. Even so many prognostic modelling techniques continue to use snapshot baseline information to assign risk. Where longitudinal information is used to its full extent, the important statistical task revolves around appropriately modelling the association between the longitudinal biomarker and the time-to-event outcome, hence the development of the joint longitudinal-survival model.

5.2.1 Personalised medicine

Disease characteristics and presentation vary by patient, so much so, that a person-centred approach towards treatment has long been advocated. Tailoring treatments and interventions to an individual, based on their particular set of risk factors and predicted prognosis, represents an about turn on traditional 'one-size fits all' approaches to patient care, which require corresponding statistical techniques [184; 185]. Biomarker profiles can provide a personalised medium through which to establish disease progression, by acting as surrogate markers for the outcome of interest [186]. Surrogate markers prove advantageous in a clinical trial setting, resulting in comparatively shorter trials with an easier and cheaper to measure endpoint [187]. For outcomes such as cancer survival it is already standard practice to look at alternative measures closely associated to disease progression to tailor treatment to the patient. For example, in prostate cancer patients it is usual to monitor prostate-specific antigen (PSA)[188; 189; 190]. In order to make inferences about the survival endpoint by way of the biomarker observations, it is necessary for the two outcomes to be linked statistically. Furthermore, individual monitoring can only be achieved if the fitted model allows for subject-level inferences to be made. The joint-longitudinal-survival model links longitudinal and survival models through shared random effects, allowing

for variation to be modelled between individuals [29]. This ultimately allows the prediction of subject-specific effects, and monitoring at the patient level.

5.2.2 Informative drop-out

Studies in which biomarkers are observed repeatedly over time frequently experience a level of patient drop-out. This can affect the timing and number of observations collected per individuals. Where it is suspected that the drop-out is associated with the underlying biomarker response then it is classified as missing not at random, as discussed in Chapter 3 section 3.3.1. This can affect inferences if the analysis is conducted on a complete case basis. The joint modelling framework has been proposed to model the association between the biomarker and drop-out mechanisms to investigate the underlying dependency [191]. In a similar fashion where a biomarker is strongly associated with a a time-to-event outcome, the joint model can be said to be modelling an informative drop-out process. It is likely sicker patients will have more extreme biomarker observations, and in turn will probably drop-out sooner than a healthier patient. An example of this is quality of life (QOL) and survival in cancer studies, where a poorer QOL score may indicate a greater risk of death and therefore drop-out [29; 114]. Information presented in this chapter will be based on the biomarker as a surrogate for a time-to-event outcome, with focus on the subject-specific estimates which are integral to dynamic monitoring.

5.3 Simple approaches to joint modelling

The joint longitudinal-survival model first emerged in the late 1990s and early 2000s, motivated by the desire to simultaneously model an intermittently observed longitudinal biomarker, subject to measurement error, and a time-to-event outcome in the context of human immunodeficiency virus (HIV) progression

[32; 192]. Prior to this simplistic approaches were used, such as the survival model with inclusion of a time-varying biomarker covariate, which will be discussed briefly in section 5.3.2. Improving upon this, the two-stage model has implemented both the LMM and survival model (see Chapters 3 and 4) in a two step procedure and will be presented in section 5.3.3. First it is important to note that biomarkers broadly fall into two categories, endogenous and exogenous, which will be defined in the next section.

5.3.1 Endogenous and exogenous covariates

Time-dependent covariates come in different forms with distinguishing features and modes of treatment as a result. Endogenous or internal covariates and exogenous or external covariates will be defined here [36; 37]. Let T_i be the observed survival time for individual i. Following the exposition by Rizopoulos [37], let $y_i(t)$ represent the covariate vector at time t for individual i, with the complete covariate history denoted by $Y_i(t) = \{y_i(s), 0 \leq s < t\}$. Then formally an exogenous covariate satisfies Equation 5.1 for all s, t such that $0 < s \leq t$ and $ds \to 0$.

$$Pr\{s \leq T_i < s + ds \mid T_i \geq s, Y_i(s)\} = Pr\{s \leq T_i < s + ds \mid T_i \geq s, Y_i(t)\} \quad (5.1)$$

This suggests that $y_i(.)$ is associated with the event over time. Yet, an event at time s remains independent of the future profile of $y_i(.)$ at time $t > s$. Examples of such external covariates are time of day or seasons of the year, which would not directly impact the occurrence of an event.

Longitudinally observed biomarker observations, such as hCG, fall under the endogenous covariate category. These are internal because their existence relies

on the survival of the patient. This means that the biomarker history informs the timing of the event if it occurs, as the biomarker can only be observed if the patient remains event-free. For example hCG can only be observed if the pregnancy progresses and is not miscarried. This relationship is defined in Equation 5.2, so that a person must be alive to contribute for the biomarker to be measurable.

$$S_i(t \mid Y_i(t)) = Pr(T_i > t \mid Y_i(t)) = 1 \tag{5.2}$$

Endogenous covariates, often biomarkers repeatedly observed for an individual, are intrinsically linked to the survival outcome, particularly when that event is death. Furthermore, the nature of an endogenous biomarker means it is measured with error and often observed intermittently. The dependency between the two outcomes requires appropriate modelling, so that the association is established whilst accounting for the features of endogenous biomarkers. This has led to the development of statistical methods which can simultaneously model both outcomes.

5.3.2 Survival model with time-varying covariate

It is standard practice to include baseline-measured covariates in a survival model. In the case of biomarker observations it is common, in the context of prognostic models, to include only the initially observed biomarker measurement in the model. However, this approach ignores valuable time sensitive information which could be gleaned from the repeated observations. From an economic point of view however, collecting panel data when there is no intention of utilising it wastes valuable resources [193]. The most basic way of estimating the dependency between a longitudinal observed biomarker and survival outcome is to fit a standard

proportional hazards model (see Equation 4.11) and include the biomarker as a time-varying covariate [30]. The model then becomes,

$$h_i(t, \mathbf{x}_i, Y_i(t)) = h_0(t) \exp\left[\mathbf{x}_i\boldsymbol{\beta} + \alpha y_i(t)\right] \tag{5.3}$$

A reminder that $h_0(t)$ represents the baseline hazard which can be modelled via a chosen distribution or be left unmodelled as discussed in sections 4.6.3 and 4.6.1. Baseline covariates, i.e. not time-varying, are represented by the vector $\mathbf{x}_i$ with corresponding vector of log hazard ratios $\boldsymbol{\beta}$. The repeated biomarker observations $y_i(t)$ have associated parameter estimate *alpha* which is the log hazard ratio for a unit increase in the biomarker.

For a continuous biomarker, however, this does not provide the most appropriate method of analysis, namely due to issues surrounding the intermittently measured biomarker. First of all a continuous measurement, due to reasons discussed in section 3.3, is often measured with error [38]. However the disparity arises, this measurement error cannot be accounted for using a survival model and so estimates can be unrealistically precise with a small standard error [28]. Secondly, inclusion of the time-varying covariate assumes that the sporadically measured biomarker measurements remain constant between observed values. Figure 5.1 illustrates this phenomena. The resulting model will then likely underestimate the size of the association [194]. A standard survival model with a time-varying covariate uses a last observation carried forward (LOCF) approach, where the hazard at time t is based on the last observed biomarker value [33]. As more time elapses between measurements, a greater level of measurement error will be introduced and the strength of the association will be further diluted. Additionally observations taken for the same individual are likely to be correlated and a time-varying covariate does not account for this within-person variation. It is not

expected that a continuous biomarker will remain constant between measurement times, or that a series of measurements for a given individual are independent. As a result, a survival model including a time-varying continuous biomarker cannot appropriately model the relationship between repeated measurements and time to event.

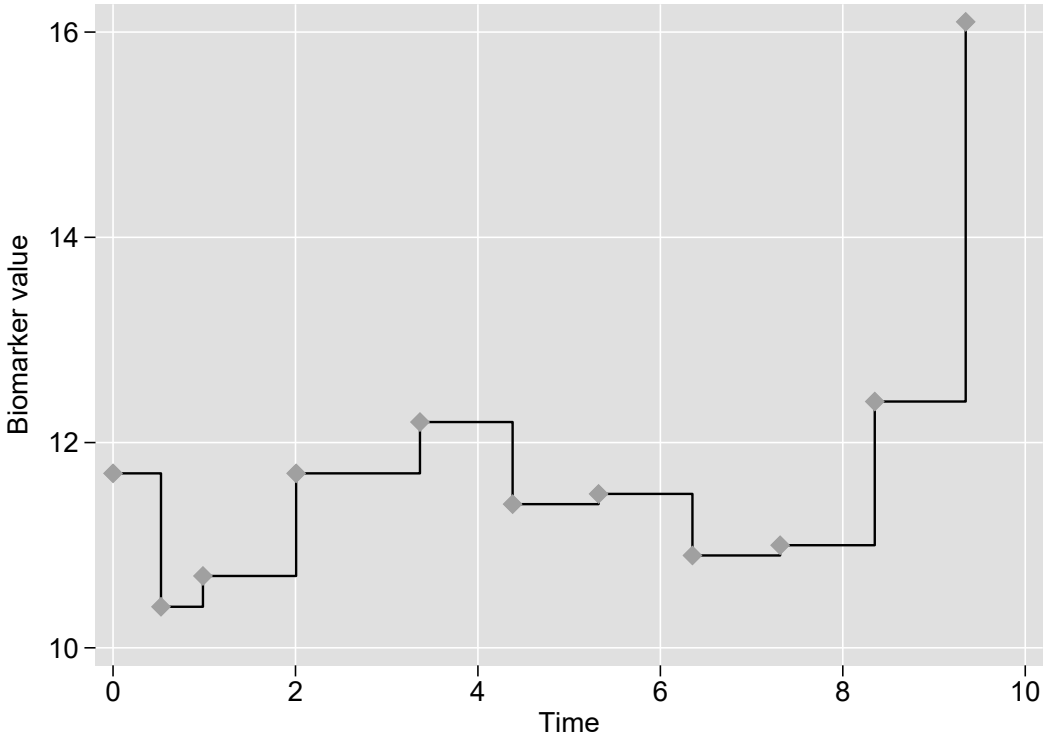


FIGURE 5.1. Observed longitudinal biomarker values as modelled in a survival model with time-varying covariate

5.3.3 Two-stage model

Two-stage models, are by definition models fitted in two stages. They aim to go some way in addressing concerns surrounding measurement error and intermittently observed biomarkers, by first modelling the longitudinal profile using a linear mixed effects model. The subject-specific values from this model are then

included in a survival model as a time-varying covariate, so that it is possible to estimate the association between the repeatedly measured biomarker and time to event. This was developed to model AIDS progression and CD4 counts [195]. This is an improvement on the standard survival model with time-varying covariate as the longitudinal biomarker information is being appropriately modelled.

As discussed in section 3.5, the advantage of modelling panel data using linear mixed effects models are that they can give us subject-specific estimates in addition to the standard population-average estimates. The general formulation of the LMM for a biomarker response variable $Y(t)$ for the ith individual is presented once more in Equation 5.4.

$$y_{ij}(t) = \mathbf{X_i}(t_{ij})\boldsymbol{\beta} + \mathbf{Z_i}(t_{ij})\mathbf{u}_i + e_{ij}$$

$$\mathbf{u}_i \sim MVN(0, \mathbf{G}) \qquad e_{ij} \sim N(0, \sigma_e^2) \tag{5.4}$$

$\mathbf{X}_i$ denotes the design matrix of fixed effects for patient i, with associated fixed effects parameters $\boldsymbol{\beta}$. $\mathbf{Z}_i$ is the design matrix for the random effects and $\mathbf{u_i}$ the associated random effect parameters. The random effects parameters are assumed to be multivariate normally distributed with mean 0 and matrix of variance components $\mathbf{G}$. The residual error term e_{ij} takes into account the measurement error associated with continuous biomarkers and is normally distributed with mean 0 and level one or within-subject variance σ_e^2.

Modelling longitudinally measured biomarkers in this way allows measurement error to be accounted for and also builds a complete profile by implicitly estimating the true unobserved biomarker values, essentially filling in the gaps between the intermittent observations. This model allows inferences at the individual level. The LMM can be crudely linked to the time-to-event outcome by defining a survival model which also includes the output of the longitudinal model,

effectively combining the two. Firstly, the fitted values for the linear predictor are obtained as a combination of the fixed effects and subject-specific deviation from the mean response. The subject-specific deviations are the BLUPs for the random effects based on the specification of the fitted longitudinal model. The corresponding survival model includes the individualised predictions, $\hat{m}_i(t)$, as a time-varying covariate and is shown in Equation 5.5. This allows estimation of the association, α, between the change in biomarker and time-to-event outcome. The log hazard ratio for a unit increase in the fitted subject-specific predictions of the biomarker is represented by α. The model is similar to the time-varying covariate case (Equation 5.3), with the biomarker observations now appropriately modelled.

$$h_i(t, \mathbf{x}_i, m_i) = h_0(t) \exp\left[\mathbf{x}_i\boldsymbol{\beta} + \alpha\hat{m}_i(t)\right] \tag{5.5}$$

This method is computationally attractive as it is simple to fit in statistical software. Although the method reduces bias when compared to the time varying covariate approach, the uncertainty in estimates from the longitudinal stage are not carried through to the survival stage. This means that the estimates are too precise, with unrealistically small standard errors [28; 196]. In addition as there is no dependency structure between the two models informative drop-out is not considered [28]. Though true unobserved biomarker values have been estimated, predictions are still based on the discrete times at which the biomarker was observed. As a result the survival model estimates are again based on the assumption that values do not change between measurements. Despite being an improvement on previous methods, the two-stage approach continues to have drawbacks. However, the two stage model is a reasonable option when computational efficiency

is a concern, as can be the case when fitting joint longitudinal-survival models particularly in small sample settings.

5.4 Joint longitudinal-survival models

Up to now, conventional methods for simultaneously modelling longitudinal and time-to-event data have been discussed. The shortcomings of standard techniques suggest a model developed specifically for the joint modelling of the two types of outcome is required.

5.4.1 Development of the joint model

Joint longitudinal-survival models, briefly joint models, first arose in the field of AIDS in the late 1990s, and purported to address two aspects of the analysis of longitudinal and survival data. Through use of a Markov chain Monte Carlo (MCMC) technique, the authors modelled repeated measurements of CD4 counts as a function of time whilst also relating the effect of the biomarker on time to diagnosis of AIDS [197]. Since this initial analysis, further inroads have been made in joint modelling, both within a frequentist and Bayesian framework [30; 198]. Wulfsohn and Tsiatis [192] first proposed joint maximization of the likelihood from both the longitudinal process and survival data, also in the area of HIV/AIDS. In particular the necessity of incorporating the longitudinal data into the time-to-event process was emphasized, to minimize previously identified biases due to measurement error and intermittently measured biomarkers [192; 194]. A combination of a mixed effects model to build a longitudinal trajectory for the biomarker and a survival model to describe the associated risk of event have been advocated for their efficiency [32]. Further developments have been seen in cancer, particularly in modelling the association between repeated PSA and prostate cancer recurrence [188]. Since their inception, joint longitudinal-survival

models have been extended to settings with alternative time-to-event submodels, competing risks and incorporation of multiple longitudinal biomarkers [34; 50; 199]. Advances have also been made in the availability of statistical software, which can now fit these types of models with relative speed and efficiency [38; 200; 201; 202].

5.4.2 Joint model formulation

The classical (frequentist) joint model, which will be the focus here, can be thought of in terms of two component models. Joint models were first developed in the context of continuous biomarkers, so the longitudinal part has traditionally consisted of a linear mixed effects model, described in Equation 3.1. The survival submodel is a proportional hazards model. As with the two-stage approach measurement error is accounted for via the mixed effects model. In the context of the joint model, it is now reparameterised in terms of the trajectory function $m_i(t)$ (see Equation 5.6). This estimates the true unobserved, subject-specific values of the biomarker $m_i(t)$ for the i^{th} patient at time t, effectively removing measurement error and addressing concerns around the intermittent observations [203].

$$y_i(t) = m_i(t) + e_i(t) \quad e_i(t) \sim N(0, \sigma_e^2)$$

$$m_i(t) = \mathbf{X}_i(t)\boldsymbol{\beta} + \mathbf{Z}_i(t)\mathbf{u}_i \quad \mathbf{u}_i \sim MVN(0, \mathbf{G}) \tag{5.6}$$

The trajectory function is made up of a combination of fixed and random effect design matrices, $\mathbf{X}_i$ and $\mathbf{Z}_i$ respectively with corresponding parameter estimates $\boldsymbol{\beta}$ and $\mathbf{u}_i$. The error terms $e_i(t)$ are normally distributed with variance σ_e^2, and is

independent of the random effects and that residual errors are independent (see Equation 3.2).

A proportional hazards survival submodel is assumed (see section 4.6.2). Borrowing notation from Rizopoulos [30], let us define $M_i(t) = \{m_i(s), 0 \leq s \leq t\}$ to be the true unobserved longitudinal profile up to time t. The survival submodel is then given by,

$$h(t|M_i(t), \mathbf{x}_i) = h_0(t) \exp\left[\mathbf{x}_i\boldsymbol{\beta} + \alpha m_i(t)\right] \tag{5.7}$$

Here, $h_0(t)$ again represents the baseline hazard function and $\mathbf{x}_i$ gives a set of baseline time-independent covariates with associated vector of log hazard ratios $\boldsymbol{\beta}$. The term $\alpha m_i(t)$ is the current value parameterisation. The association parameter α is the log hazard ratio for a unit change in the absolute current value of the biomarker trajectory. Alternative association structures will be discussed in section 5.5. This model is underpinned by shared random effects $\mathbf{u}_i$ as the entire trajectory function is incorporated into the survival submodel. This shared random effects (SREM) joint model will be the default model discussed in this chapter, unless otherwise indicated. Incorporating the longitudinal submodel into the survival submodel effectively links the expected value of the longitudinal response to the survival time, where typically a response would not have been observed [52]. Figure 5.2 illustrates how the biomarker observations are modelled by a joint longitudinal-survival model.

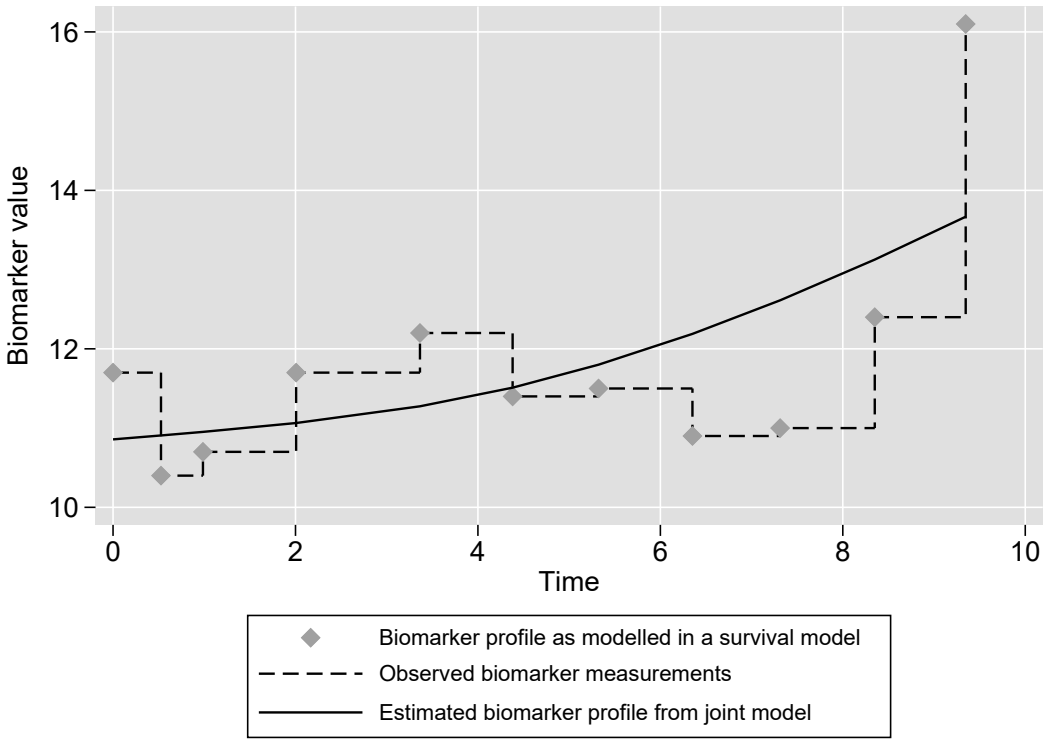


FIGURE 5.2. Observed longitudinal biomarker values as modelled in a joint longitudinal-survival model

The corresponding survival function is described in Equation 5.8.

$$S(t \mid M_i(t), \mathbf{x}_i) = \exp\left(-\int_0^t h_0(u) \exp\left\{\mathbf{x}_i\boldsymbol{\beta} + \alpha m_i(u)\right\} du\right) \tag{5.8}$$

The survival function is directly dependent on the trajectory function $m_i(t)$. With the inclusion of random effects and consequently subject-specific biomarker profiles, the framework naturally lends itself to individualised prediction. This will be discussed further in section 5.11.2. The integral in the survival function is analytically intractable and requires numerical integration to evaluate. This is due to the time-dependency of the longitudinal profile, and as a result makes estimation of the joint model complex. Numerical integration methods will be presented in section 5.6.

5.4.3 Baseline hazard

Several choices of baseline hazard have been proposed. The original development of the joint model utilised the semi-parametric Cox model [31]. As discussed in section 4.6.1, the Cox model does not assume a functional form for the baseline hazard and therefore avoids possibly incorrect distributional assumptions. Conversely, avoiding estimation of the baseline hazard means absolute risk predictions, a great driver for using joint models, cannot be obtained in a straightforward manner [42; 171]. Furthermore, leaving the baseline hazard unspecified results in the underestimation of standard errors [204]. Bootstrapping as a means of obtaining the standard errors adds an additional computational component to an already complex estimation procedure (see section 5.6)[54].

A parametric framework is preferred for the prediction of conditional survival, with the Weibull distribution a popular choice. Splines provide a flexible alternative, as they can be tailored in their number and placement over the follow-up time. Restricted cubic splines will be explored to model the baseline hazard of early pregnancy loss in Chapters 6 and 7 [182]. Alternative spline functions which have been proposed include (penalized) b- and p-splines [205]. The former will be implemented in the Bayesian joint model application in Chapter 7.

5.5 Association structures

The current value association structure is the standard parameterisation for joint models, where α is the log hazard ratio for a unit increase in the absolute value of the biomarker at time t. It may not always be realistic to assume that changes in survival are related to an increase in absolute changes in the biomarker trajectory. In this section alternative association structures, which may be more clinically meaningful will be explored.

5.5.1 Random effects association

The original proposal for the joint model was motivated by the use of the random effects association structure to model informative dropout [206]. This includes only the random effects of the longitudinal submodel in the linear predictor of the survival submodel, shown in Equation 5.9 [207].

$$h_i(t) = h_0(t) \exp\left[\mathbf{x}_i\boldsymbol{\beta} + \alpha^T \mathbf{u}_i\right] \tag{5.9}$$

This parameterisation allows for a vector α^T of association parameters, which correspond to the random effects included in the longitudinal submodel. So for a random intercept and random slope model the survival submodel under the random effects parameterisation is presented in Equation 5.10.

$$h_i(t) = h_0(t) \exp\left[\mathbf{x}_i\boldsymbol{\beta} + \alpha_1 u_{i0} + \alpha_2 u_{i1}\right] \tag{5.10}$$

The association parameters α_1 and α_2 give the log hazard ratio for one unit increase in the subject-specific deviation from the population-average intercept and slope respectively. For a biomarker where higher values equal greater risk, α_1 describes the increase in hazard for individuals who have a higher than average biomarker value at baseline. Whilst α_2 describes the increase in hazard for those individuals who have a steeper than average slope. When only the subject-specific deviations are included in this way, the association structure is time-independent, which leads to a closed-form solution for the integral of the survival function [207]. The addition of further random effects to model the slope non-linearly can lead to difficulty in interpreting the resulting α parameters. For example, for a trajectory modelled using a linear and quadratic time term, it would be difficult to separate the effects of each of the corresponding association parameters [207].

5.5.2 First derivative association

An important structure which will be applied to the pregnancy outcome context is the first derivative or rate of change parameterisation [207].

$$h(t|M_i(t), \mathbf{x}_i) = h_0(t) \exp\left[\mathbf{x}_i\boldsymbol{\beta} + \alpha_1 m_i(t) + \alpha_2 m_i'(t)\right] \tag{5.11}$$

where

$$m_i' = \frac{d}{dt} m_i(t) = \frac{d}{dt}\left\{\mathbf{X}_i(t)\boldsymbol{\beta} + \mathbf{Z}_i(t)\mathbf{u}_i\right\}$$

Here α_2 represents the log hazard ratio for a unit increase in the change in slope of the longitudinal trajectory function $m_i(t)$ at time t. As in the case of hCG and early pregnancy loss, absolute increases in a biomarker may not explain the entire association with the survival outcome (see Chapter 7). Rather, a change in the rate of increase of the biomarker could be an indicator of progression. It has been shown that the rate of change association is sensitive to misspecification of the longitudinal trajectory, particularly if a simple linear trajectory is assumed when the relationship is more complex [52]. Subsequently, all possible attempts to appropriately model the longitudinal trajectory should be made before fitting a rate of change association. It is common for the rate of change to be modelled alongside the current value association, and introducing an additional parameter for estimation also increases computational complexity. This association structure models a change in *linear* slope, which is illustrated in Figure 5.3.

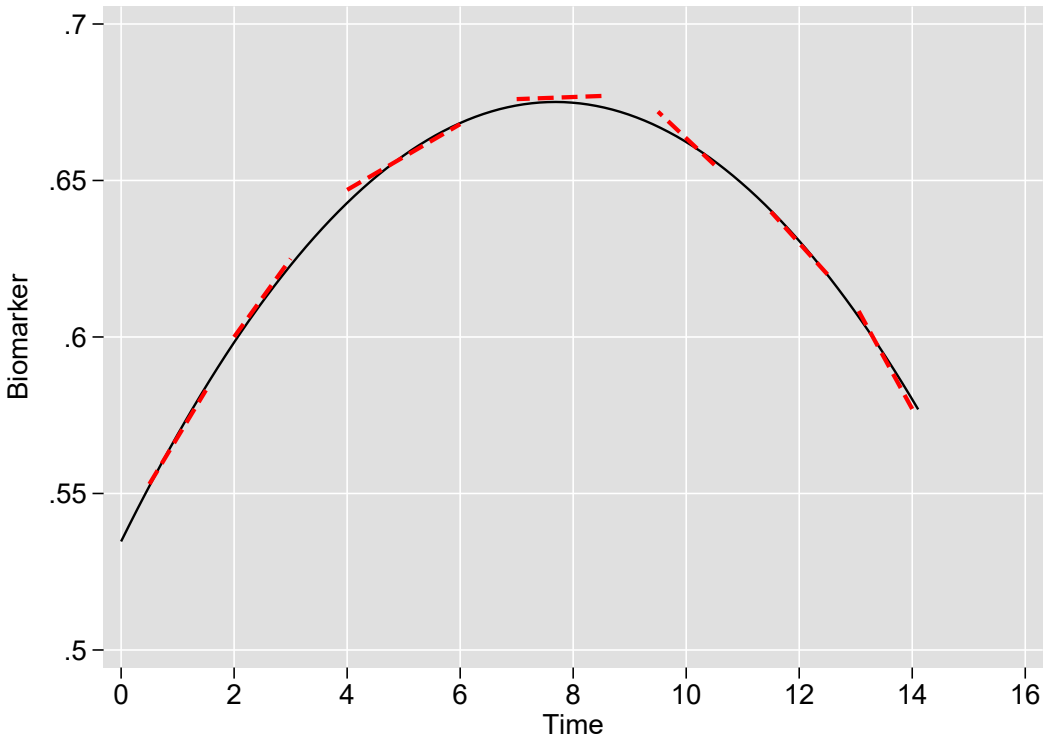


FIGURE 5.3. Conceptual representation of the first derivative association structure

5.5.3 Cumulative effects association

The current value association assumes that the risk of an event at time t depends only on the biomarker response at that same time-point. There may be a case for advocating an association structure which incorporates the cumulative effect of the longitudinal outcome up to time t. Scenarios where the cumulation of an effect may be of interest include exposures to a drug over time or possibly the long term history of smoking [207]. Including the integral up to time t of the trajectory function in the survival submodel allows the entire longitudinal history up to t to be associated with the log hazard ratio α. The survival submodel with cumulative effects association structure is presented in Equation 5.12.

$$h_i(t) = h_0(t) \exp\left\{\mathbf{x}_i\boldsymbol{\beta} + \alpha \int_0^t m_i(s)ds\right\} \tag{5.12}$$

The integral estimates the area under the curve (AUC) of the longitudinal trajectory with log hazard ratio α interpreted as the log hazard ratio for a unit increase in the AUC. A weight function $\bar{w}(.)$ that varies across time can be introduced to place greater emphasis on certain observations. This is shown in Equation 5.13.

$$h_i(t) = h_0(t) \exp\left\{\mathbf{x}_i\boldsymbol{\beta} + \alpha \int_0^t \bar{w}(t-s)m_i(s)ds\right\} \tag{5.13}$$

It would be sensible to place greater weight on more recent observations as opposed to those observed further back in time. Weight functions which achieve this end, include the normal, logistic and Student's T distributions [207]. The integrals for the weighted cumulative effect do not have closed form solutions, requiring numerical integration techniques discussed in section 5.6.

5.5.4 Non-linear association structures

The association structures discussed all estimate the association parameter for a linear change of the chosen structure. There has been little research into modelling non-linear association structures. For example, it may be more meaningful to look at the effect of an increase in the quadratic or cubic value of the biomarker trajectory on the time-to-event outcome. Blood pressure is one example of a biomarker which does not have a linear risk profile. Extreme values, both very low and very high, are expected to carry greater risk of an event, resulting in a U or J shaped hazard. This indicates a quadratic association between the biomarker trajectory and hazard and could lead to biased effect estimates if modelled assuming linearity. The simulation study conducted in Chapter 9

proposes to investigate the effect of misspecification of a non-linear association on predicted survival estimates.

5.6 Likelihood Estimation

Estimation of the joint model is a complex task, which is why simpler methods continue to be implemented (see sections 5.3.2 and 5.3.3). Much of the literature on joint models is based on the frequentist framework, although Bayesian estimation is receiving increasing attention as it bypasses the necessary numerical integration required for estimating the random effects in the likelihood method [198]. The classical shared random effects joint model will be the focus here, although the Bayesian estimation using Markov Chain Monte Carlo (MCMC) methods will be discussed in section 5.8.

The SREM is estimated through a full maximum likelihood approach. The log-likelihood is defined for the joint distribution of $\{T_i, d_i, y_i\}$, which refer to the observed survival time, event indicator and longitudinal response for the i^{th} individual where $i = 1, \cdots, n$. The observed survival time is the minimum of the true survival time S_i and the censoring time C_i, so that $T_i = min(S_i, C_i)$. The event indicator, d_i is 1 if $S_i < C_i$ and 0 otherwise [52]. Furthermore, $\mathbf{u}_i$ represents the vector of time-independent random effects which underpin the longitudinal and survival outcomes. Following the approach of Rizopoulos [208] first a parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_t, \boldsymbol{\theta}_y, \boldsymbol{\theta}_u)$ with $\boldsymbol{\theta}_t, \boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_u$ referring to the parameters for the time-to-event outcome, longitudinal outcome and the random-effects covariance matrix respectively. The key assumption here is that the longitudinal and survival processes depend jointly on the underlying vector of time-independent random effects $\mathbf{u_i}$ [196]. The random effects describe the correlation between

repeated longitudinal observations and the purported association between longitudinal and time-to-event outcomes [208]. To estimate the parameters the joint distribution function is written as follows,

$$p(\mathbf{y}_i \mid \mathbf{u}_i; \boldsymbol{\theta}) = \prod_{j=1}^{n_i} p\left\{y_i(t_{ij}) \mid \mathbf{u}_i; \boldsymbol{\theta}\right\} \tag{5.14}$$

$$p(T_i, d_i, \mathbf{y}_i \mid \mathbf{u}_i; \boldsymbol{\theta}) = p(T_i, d_i \mid \mathbf{u}_i; \boldsymbol{\theta}) p(\mathbf{y}_i \mid \mathbf{u}_i; \boldsymbol{\theta}) \tag{5.15}$$

Here j refers to the specific longitudinal observation of individual i. Independent censoring is assumed, whilst the timing of observations is considered to be non-informative. The joint likelihood is made up of the probability density functions for the conditional longitudinal submodel, random effects and conditional survival submodel. Thus the log-likelihood contribution for the i^{th} individual is given by,

$$\begin{aligned} \ln L_i &= \ln p(T_i, d_i, \mathbf{y}_i; \boldsymbol{\theta}) = \ln \int p(T_i, d_i, \mathbf{y}_i, \mathbf{u}_i; \boldsymbol{\theta}) d\mathbf{u}_i \\ &= \ln \int p(T_i, d_i \mid \mathbf{u}_i \boldsymbol{\theta}_t) \left[\prod_{j=1}^{n_i} p\{y_i(t_{ij}) \mid \mathbf{u}_i; \boldsymbol{\theta}_y\}\right] p(\mathbf{u}_i, \boldsymbol{\theta}_u) d\mathbf{u}_i \end{aligned} \tag{5.16}$$

The conditional survival and longitudinal submodel, and random-effect density functions are given in Equations 5.17, 5.18 and 5.19 respectively.

$$\begin{aligned} p(T_i, d_i \mid \mathbf{u}_i; \boldsymbol{\theta}_t) &= h(T_i \mid M_i(T_i); \boldsymbol{\theta}_t)^{d_i} \times S(T_i \mid M_i(T_i); \boldsymbol{\theta}_t) \\ &= \left[h_0(T_i) \exp\left\{\mathbf{x}_i \boldsymbol{\beta} + \alpha m_i(T_i)\right\}\right]^{d_i} \times \exp\left(-\int_0^{T_i} h_0(s) \exp\left\{\mathbf{x}_i \boldsymbol{\beta} + \alpha m_i(s)\right\} ds\right) \end{aligned} \tag{5.17}$$

$$p\left\{y_i(t_{ij}) \mid \mathbf{u}_i, \boldsymbol{\theta}_y\right\} = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{\left[y_i(t_{ij}) - m_i(t_{ij})\right]^2}{2\sigma^2}\right\} \tag{5.18}$$

$$p(\mathbf{u}_i; \boldsymbol{\theta}_u) = (2\pi)^{-q_u/2} |\mathbf{G}|^{-1/2} \exp\left(-\frac{\mathbf{u}_i^T \mathbf{G}^{-1} \mathbf{u}_i}{2}\right) \tag{5.19}$$

The random effects are assumed to be multivariate normally distributed with variance-covariance matrix $\mathbf{G}$, and q_u is the dimension of the random effects.

The overall log-likelihood $\ln L$ can be maximized using either or a combination of the Newton-Raphson and Expectation-Maximisation (EM) algorithms discussed in sections 4.6.4 and 3.5.3.

The Newton-Raphson algorithm is the default method implemented in Stata for joint modelling commands `stjm` and `merlin` (see section 5.9) [38; 140; 200]. When maximising a likelihood function to estimate a vector of parameters $\boldsymbol{\theta}$, given the observed data X, the first derivative is computed with respect to the parameter vector $\boldsymbol{\theta}$ and solved by setting it to 0. This root-finding exercise can be performed iteratively by updating a chosen set of initial values until the convergence criterion are met. The steps for the Newton-Raphson algorithm, as described by Gould *et al.* [140] are as follows,

(1) Start with an initial guess for $\boldsymbol{\theta}_i$

(2) Calculate a new guess $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \{-\mathbf{H}(\boldsymbol{\theta}_i)\}^{-1} g(\boldsymbol{\theta}_i)$, where $g(\boldsymbol{\theta}_i)$ is the gradient or first derivative vector and $-\mathbf{H}(\boldsymbol{\theta}_i)$ is the positive definite matrix of second derivatives, known as the Hessian

(3) Repeat until convergence criterion are met (in Stata this is when $g(\boldsymbol{\theta}_i)$ $\mathbf{H}(\boldsymbol{\theta}_i)^{-1}$ $g(\boldsymbol{\theta}_i)' < \epsilon_1$ where $\epsilon_1 = 1 \times 10^{-05}$)

The Expectation-Maximisation algorithm maximises the conditional expectation of the complete likelihood function, rather than the observed data likelihood

function[209]. In the instance of joint models the latent variables or random effect estimates are treated as unobserved data or missing in the expectation step [208]. Once the expectation of the log-likelihood is computed for a set of initial values, a new estimate of the parameters is determined through maximisation. The maximisation can be carried out using step 2 of the Newton-Raphson algorithm [208]. This is repeated until the chosen convergence criterion are met. The steps of the Expectation-Maximisation algorithm are presented as follows [209; 210],

(1) Start with an initial guess for the parameters θ_i
(2) Expectation step: compute $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n) = E\left[\ln L(\boldsymbol{\theta}\mathbf{W}) \mid \mathbf{X}, \boldsymbol{\theta}_n\right]$
(3) Maximisation step: obtain the new estimate $\boldsymbol{\theta}_{i+1}$ by maximising $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)$
(4) Repeat steps 2 and 3 until convergence is achieved

More specifically, the complete data likelihood for $\mathbf{W} = (\mathbf{X}\mathbf{X}^m)$ depending on parameters $\boldsymbol{\theta}$ is given by $L(\boldsymbol{\theta}\mathbf{W})$, where $\mathbf{X}$ is the observed data and $\mathbf{X}^m$ the unobserved data (random effects) [210]. $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)$ is the expectation of the complete log likelihood, conditional on the current parameter estimates $\boldsymbol{\theta}_n$.

When utilising these algorithms, the score and Hessian must be calculated. However, the integrals with respect to the random effects in the joint log likelihood or the expectation of the joint log likelihood function integrals with respect to the random effects are intractable (see Equation 5.16)[208]. Numerical integration techniques, such as Gaussian quadrature, are implemented to approximate these integrals for maximisation [208].

5.7 Numerical approximation

Simple Gauss-Hermite quadrature was first used to evaluate the log likelihood when joint models were in their infancy. The method has been used extensively in the generalised linear mixed model (GLMM) framework [211]. With time and advances in computation adaptive and pseudo-adaptive methods are now preferred for their increased accuracy and efficiency. Simple Gauss-Hermite quadrature will be discussed briefly before extending to adaptive quadrature.

5.7.1 Gauss-Hermite quadrature

Simple Gauss-Hermite quadrature can evaluate intractable integrals which take the form shown in Equation 5.20 [212].

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx \approx \sum_{q=1}^{m} w_q f(x_q) \tag{5.20}$$

x_q and w_q are the corresponding quadrature nodes and weights. The weight function is specified in Equation 5.21, with x_q the q^{th} root of the Hermite polynomial $H_m(x)$. The theory states that for ideally spaced and weighted nodes the approximation will be exact if the polynomial is of degree $2m-1$[213].

$$w_q = \frac{2^{m-1} m\sqrt{\pi}}{m^2[H_{m-1}(x_q)]^2} \tag{5.21}$$

As shown by Naylor and Smith [212], if the weight function e^{-x^2} is taken to be the normal density, $g(.)$ with mean and variance μ and σ^2 respectively, we can write an approximation of the quadrature based on the normal kernel[212]. For a single random effect Equation 5.20 can be rewritten as,

$$\int_{-\infty}^{\infty} f(x)g(x \mid \mu, \sigma^2)dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} f(x) \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx \tag{5.22}$$

Writing this in terms of r, where $r = \frac{(x-\mu)}{\sqrt{2}\sigma}$, the integral can be manipulated to take the appropriate form of Equation 5.20.

$$\int_{-\infty}^{\infty} f(x)g(x \mid \mu, \sigma^2)dx = \frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} f(\mu+\sigma\sqrt{2}r) \exp^{-r^2} dr \approx \sum_{q=1}^{m} f(\mu+\sigma\sqrt{2}r) \frac{w_q}{\sqrt{\pi}} \tag{5.23}$$

This last formulation is the quadrature approximation with nodes $\mu + \sigma\sqrt{2}x_q$ and weights $\frac{w_q}{\sqrt{\pi}}$. Equivalent nodes and weights for a standard normally distributed kernel i.e. $N(0, 1)$, are $\sqrt{2}x_q$ and $\frac{w_q}{\sqrt{\pi}}$ respectively.

The number of random effects included in the corresponding joint model usually exceeds one, particularly as it is common to include a random intercept and slope. This would then lead to the use of a multivariate normal kernel of dimension Q [213]. The vector of nodes, $\mathbf{d}_{q_1,\cdots,q_Q} = (d_{q_1}, \cdots, d_{q_Q})$ for the multivariate normal is multiplied by the Cholesky decomposition of the variance-covariance matrix of the random effects $\mathbf{G}^{1/2}$. This ensures a positive definite covariance matrix. The weighted log-likelihood is then computed at each node location for each level of the random effect before the sum is taken across the total number of random effects. The log-likelihood can be written in quadrature form as shown in Equation 5.24. The approximation depends on the number of chosen nodes m, and whether it is a polynomial of degree $2m - 1$ or less [213]. It is usual to increase the number of quadrature points/nodes until the difference between the

approximations at m and $m-1$ nodes is minimal. However increasing the number of quadrature points in turn increases computational burden, whilst every additional random effect increases the load exponentially [208].

$$\log p(T_i, d_i, y_i; \boldsymbol{\theta}) = \log \int p(T_i, d_i \mid \mathbf{u}_i; \boldsymbol{\theta}_t) \left[\prod_{j=1}^{n_i} p(y_i(t_{ij}) \mid \mathbf{u}_i; \boldsymbol{\theta}_y) \right] p(\mathbf{u}_i; \boldsymbol{\theta}_u) d\mathbf{u}_i$$
$$\approx \log \sum_{u_1=1}^{m} \cdots \sum_{u_Q=1}^{m} p(T_i, d_i \mid \mathbf{G}^{1/2} \mathbf{d}_{u_1,\cdots,u_Q}; \boldsymbol{\theta}_t) \left[\prod_{j=1}^{n_i} p(y_i(t_{ij}) \mid \mathbf{G}^{1/2} \mathbf{d}_{u_1,\cdots,u_Q}; \boldsymbol{\theta}_y) \right]$$
$$\times v_{u_1} \cdots v_{u_Q} \tag{5.24}$$

In instances where the location or spread of the integrand is vastly different from that of the weight function, even increasing the number of nodes will not improve approximation[213]. Essential to note is that for each individual the log-likelihood is evaluated at the same set of nodes centred at zero and pre-multiplied by the Cholesky factor. It is unlikely that the choice of common quadrature points will result in a good approximation for every individual, particularly where there is greater variance between individuals [214].

5.7.2 Adaptive Gauss-Hermite quadrature

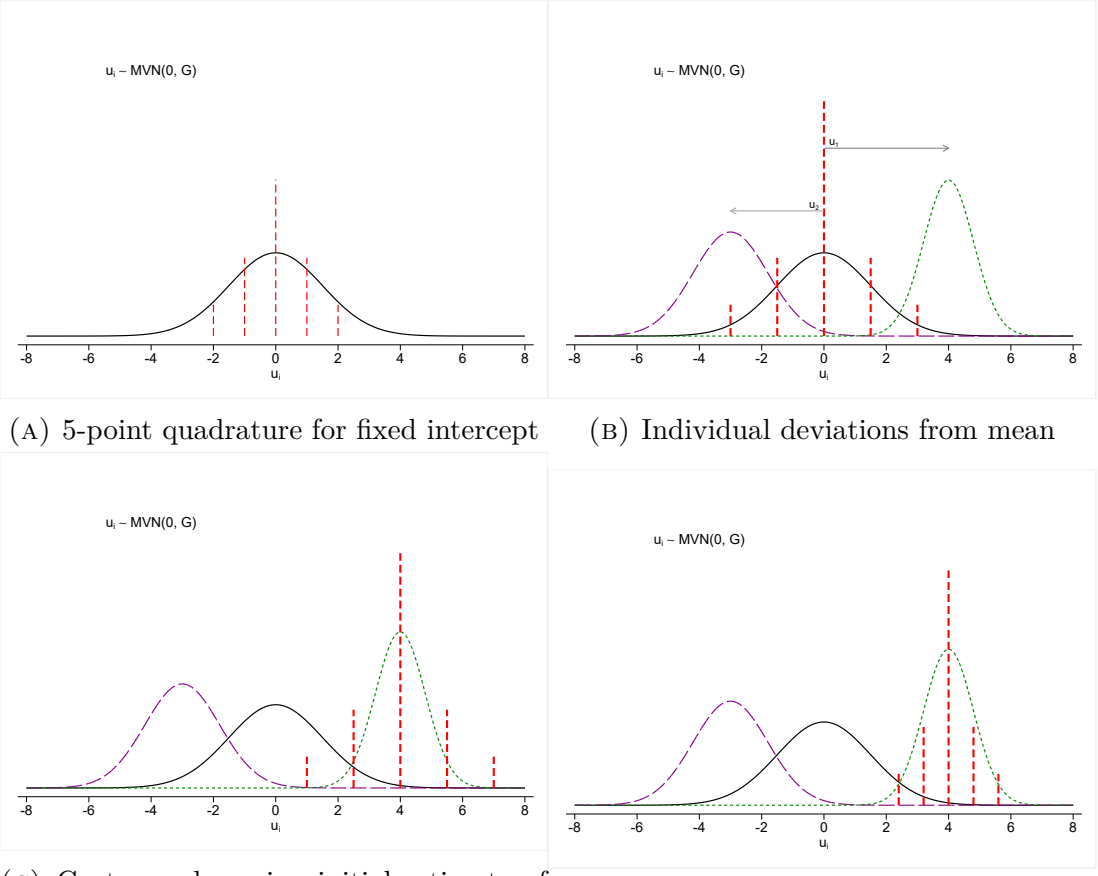
To combat these issues adaptive Gauss-Hermite quadrature developed by Pinheiro and Bates [215], has now become a mainstay in estimation of the joint likelihood. Rather than having fixed nodes centred at zero, adaptive quadrature allows the nodes for a given individual to be shifted by $\hat{u}_i$. This represents the amount an individual differs from the mean. The nodes are then scaled by the standard deviation estimate of the random effect $\hat{\sigma}$. Equation 5.25 shows the likelihood in terms of the normal kernal distribution centred at the random effect estimates

and variance-covariance matrix of the random effects; $\phi(\mathbf{u}_i \mid \hat{\mathbf{u}}_i, \hat{\mathbf{G}}_i)$ [52]. The subsequent Gauss-Hermite quadrature approximation is then given in Equation 5.26 using the transformation $\mathbf{r_i} = \hat{\mathbf{u}}_i + \hat{\mathbf{G}}_i^{1/2}\mathbf{d}_{u_1,\cdots,u_Q}$ [52].

$$\begin{aligned}\log p(T_i, d_i, y_i; \boldsymbol{\theta}) &= \log \int p(T_i, d_i \mid \mathbf{u}_i; \boldsymbol{\theta}_t \left[\prod_{j=1}^{n_i} p(y_i(t_{ij}) \mid \mathbf{u}_i; \boldsymbol{\theta}_y)\right] p(\mathbf{u}_i; \boldsymbol{\theta}_u) d\mathbf{u}_i \\ &= \log \int \frac{p(T_i, d_i \mid \mathbf{u}_i; \boldsymbol{\theta}_t)[\prod_{j=1}^{n_i} p(y_i(t_{ij}) \mid \mathbf{u}_i; \boldsymbol{\theta}_y)] p(\mathbf{u}_i; \boldsymbol{\theta}_u)}{\phi(\mathbf{u}_i \mid \hat{\mathbf{u}}_i, \hat{\mathbf{G}}_i)} \phi(\mathbf{u}_i \mid \hat{\mathbf{u}}_i, \hat{\mathbf{G}}_i) d\mathbf{u}_i\end{aligned} \tag{5.25}$$

$$\begin{aligned}\log p(T_i, d_i, y_i; \boldsymbol{\theta}) \approx (2\pi)^{Q/2} |\hat{\mathbf{G}}_i|^{1/2} \sum_{u_1=1}^{m} \cdots \sum_{u_Q=1}^{m} p(T_i, d_i \mid \mathbf{r}_i; \boldsymbol{\theta}_i) \left[\prod_{j=1}^{n_i} p(y_i(t_{ij}) \mid \mathbf{r}_i; \boldsymbol{\theta}_y)\right] \\ \times \phi(\mathbf{r}_i \mid \mathbf{0}, \mathbf{G}) \exp(\frac{1}{2}\mathbf{d}_{u_1,\cdots,u_Q}^T \mathbf{d}_{u_1,\cdots,u_Q}) \prod_{q=1}^{Q} v_{u_q}\end{aligned} \tag{5.26}$$

Where between-subject variation is high, this approach allows the nodes for individuals to be located more appropriately. This is beneficial as estimates can usually be obtained using fewer nodes overall resulting in reduced computational burden when compared with the simple quadrature approach. For selected models in later Chapters, estimation of the likelihood will be via adaptive Gauss-Hermite quadrature which is implemented in Stata packages `stjm` and `merlin`. Figure 5.4 illustrates the adaptive Gaussian quadrature concept with an example.



(A) 5-point quadrature for fixed intercept

(B) Individual deviations from mean

(C) Centre nodes using initial estimate of $\hat{u}_1$

(D) Re-scale centred nodes using $se(\hat{u}_1)$

FIGURE 5.4. Illustration of adaptive Gaussian quadrature for estimating a subject-specific deviation from the mean intercept. Adapted from Crowther [3]

5.7.3 Pseudo-adaptive quadrature

Pseudo-adaptive quadrature aims to further increase computational efficiency. This methods relies on extracting node and scale information of the posterior distribution of the empirical Bayes estimates of the random effects, by first fitting a linear mixed effects for the biomarker [208]. The adaptive quadrature approach requires the location of the mode of the random effect estimate and calculation of the Hessian matrix for each individual at each iteration. However the pseudo-adaptive approach proposes using only the output of the longitudinal model to

re-center and re-scale the integrands of the log-likelihood and score vector of the joint model [216]. This is possible by assuming that the posterior distribution of the random effects is proportional to the specification of the log likelihood [208]. This approach negates the need for a standard transformation of variable as in the adaptive Gauss-Hermite quadrature implementation. This pseudo-adaptive approach can be selected as an option for evaluation of the log-likelihood in the R package JM [201].

5.8 Bayesian estimation

Joint models can also be estimated via the Bayesian framework, with the implementation of Markov Chain Monte Carlo (MCMC) algorithms. Similarly to `merlin`, the R package JMbayes, which will be used to fit joint models in Chapter 7, employs the generalised linear mixed-effects model framework for the longitudinal part of the model [202]. In contrast to the usual distributional assumptions, the package makes use of B-splines, which are based on an alternative parameterisation of the cubic spline, to model the baseline hazard [217].

As with the frequentist approach we make the assumptions that the longitudinal and survival mechanisms are independent conditional on the random effects and that longitudinal responses of each subject are independent of another. MCMC draws from the posterior distribution of the parameters described in Equation 5.27. Here $\boldsymbol{\theta}$ refers to the full parameter vector to be estimated.

$$p(\boldsymbol{\theta}, \mathbf{u}) \propto \prod_{i=1}^{n} \prod_{l=1}^{n_i} p(y_{il} \mid \mathbf{u_i}, \boldsymbol{\theta}) p(T_i, d_i \mid \mathbf{u_i}, \boldsymbol{\theta}) p(\mathbf{u_i} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{5.27}$$

where the following holds,

$$p(y_{il} \mid \mathbf{u_i}, \boldsymbol{\theta}) = \exp \left\{ \frac{[y_{il}\psi_{il}(\mathbf{u_i}) - c\{\psi_{il}(\mathbf{u_i})\}]}{a(\phi) - d(y_{il}, \phi)} \right\} \tag{5.28}$$

Here $c(.)$, $a(.)$ and $d(.)$ are specified functions of a distribution from the exponential family for the GLMM, with $\psi_{il}(\mathbf{u}_i)$ and ϕ respectively the natural and dispersion parameters. The corresponding survival part is described in Equation 5.29.

$$p(T_i, d_i \mid \mathbf{u_i}, \boldsymbol{\theta}) = h_i(T_i \mid H_i(T_i))^{d_i} \exp\{-\int_0^{T_i} h_i(s \mid H_i(s))ds\} \tag{5.29}$$

where $h_i(t \mid H_i(t), \mathbf{w_i}(t)) = h_0(t) \exp[\boldsymbol{\gamma}^T \mathbf{w}_i(t) + f\{H_i(t), \boldsymbol{u_i}, \boldsymbol{\alpha}\}]$, with $H_i(t)$ denoting the longitudinal history up to time t, $h_0(t)$ the baseline hazard and $\mathbf{w}_i(t)$ are the vector of covariates which may vary over time. As with the frequentist approach the integral in the survival function must be approximated. The technique proposed is the Gauss-Kronrad quadrature. The rule takes a slightly different form to Gauss-Hermite quadrature detailed in Equation 5.20. For an integral $f(x)$ the $2n+1$ point Gauss-Kronrad approximation takes the form,

$$K^{2n+1} f = \sum_{i=1}^{2n+1} w_i f(x_i) \tag{5.30}$$

The rule can approximate the integral when f is a polynomial of degree $3n+1$ or less [218]. As this is an extension of the Gaussian quadrature a requirement is that n of the nodes of K^{2n+1} coincide with those of the n-point Gaussian quadrature [219]. Independent univariate normal diffuse priors are standard issue for the vector of fixed effects for the longitudinal biomarker, the regression coefficients of the survival model and the vector of spline coefficients used to model the baseline hazard and the association parameter. An inverse Wishart prior distribution is assumed for the covariance of the random effects and an inverse-Gamma prior for the variance of the residual error terms [198; 202].

5.9 Software developments

In recent years the software development for joint longitudinal-survival model has caught up with methodological advances. Both Stata and R have their own established packages for fitting a joint model with a continuous longitudinal outcome. For Stata these are `stjm` and more recently `merlin`, the latter of which utilises the generalised linear mixed-effects model framework. Both also implement adaptive Gauss-Hermite quadrature to approximate the likelihood as discussed in 5.7.2. The flexibility of the `merlin` architecture allows extensions to competing risks and multiple longitudinal markers. In R the most commonly cited software package is JM, which again estimates joint models using a frequentist approach, implementing pseudo-adaptive quadrature as an option (see section 5.7.3) [201]. Though the JM packages offers extensions to multiple markers, the R packages joineR and joineRML are specifically designed for this [204; 220]. On the Bayesian front, JMBayes allows more flexible estimation, paralleling the estimation of the conditional survival probabilities described in section 5.11.2[202]. An updated version, JMBayes2, has recently been released at the time of writing this, which now also models competing risks and multi-state processes [221]. A SAS macro equivalent to R's JM has become available relatively recently [222]. Finally JMFit, implemented in SAS, is the only fully-fledged command for joint model selection, and is discussed in detail in section 5.10.1 [223]. A review of joint modelling reporting conducted in 2016 found an increase in application of joint model methodology after 2012, which very much coincides with accelerated software output [224].

5.10 Model selection

Model selection for joint models is a work in progress. As with any model building, it is a sound idea to first refer to clinical experts for input when deciding on commonly associated variables for inclusion. This is the route which was taken for analysis of the pregnancy outcomes dataset in Chapter 7. However, comparing the significance of adding or removing a variable when deciding on the functional form of the biomarker profile or baseline hazard is useful. Joint models do not necessarily need to include the same variables within each submodel. This is because a covariate which affects the time-to-event outcome may not necessarily directly impact the biomarker response. It has become common in the literature to build each submodel separately, looking at measures such as the submodel AIC or BIC for comparison [52]. Sensitivity analyses after the model has been chosen can be used to confirm functional forms of covariates and identify outlying observations. Approaches to simultaneously submodel selection are still emerging and are presented in the following sections.

5.10.1 AIC and BIC decomposition

Zhang *et al.* [225] propose a decomposition of the full joint model AIC (or BIC) to compare the fit of each submodel in the context of the whole model. The complete joint model AIC (and BIC) is split into the AIC for the longitudinal submodel and that of the conditional (on the longitudinal data) survival submodel, so that

$$AIC = AIC_{long} + AIC_{surv|long} \tag{5.31}$$

This is currently only implementable in the SAS statistical software via the command JMFit, and is based on a polynomial longitudinal submodel [223]. The

biomarker trajectory is modelled using polynomials of degree q, with random effects associated with each polynomial term. The survival submodel includes a piecewise constant baseline hazard. This entails the construction of a finite partition of the time-axis, resulting in K intervals. Within each interval the baseline hazard is assumed to be constant. The formulation of the Zhang *et al.* [223] joint model means it requires updating for it to be applicable to the traditional joint model definition, as implemented in Stata (e.g. `stjm` or `merlin`).

5.10.2 Penalized likelihood

As an alternative to AIC (BIC) decomposition, a penalized likelihood approach has been suggested [226]. To simultaneously select fixed and random effects adaptive least absolute shrinkage and selection operator (ALASSO) penalty functions are employed. This relies on a reparameterisation of the random effects using the Cholesky decomposition. The LASSO is a commonly used regression tool in machine learning, which implements regularisation and variable selection [227]. Its aim is to shrink parameter estimates, some to zero, with the resulting non-zero terms included in the model. The adaptive LASSO does further by assigning different weights to coefficients [228]. The random effect selection is based on group penalties, ensuring that the m^{th} and l^{th} row vectors of the Cholesky decompositions of the covariances matrices are either all zero, or at least one is non-zero [226]. The penalised likelihood is then approximated via Gaussian quadrature and maximised via the EM algorithm . Estimation bias is however introduced when evaluating the penalized likelihood which requires a two-stage model approach to address this. The method also increases computational complexity, as the E-step in the EM algorithm encounters analytically intractable integrals.

Modelling outcomes such as miscarriage requires clinical input in conjunction with some form of model selection. For practical reasons, model selection was

confined to selecting functional forms of biomarker trajectories and baseline hazards, as well as the addition of random effects. This was carried out separately for each submodel for analyses in Chapters 6 and 7.

5.11 Predictions

The greatest advantage of joint longitudinal-survival models is the range of predictions which can obtained from the fitted model. The inclusion of fixed and random effects allow population-average predictions as well as subject-specific predictions. The latter especially have clinical relevance as medical care becomes increasingly tailored to the individual. Fixed effect predictions for covariate patterns can be obtained by holding the random effect estimates $\mathbf{u_i}$ fixed at zero, whilst subject-specific predictions require estimation of the $\mathbf{u_i}$.

5.11.1 Random effects

The posterior means for the random effects are obtained via the empirical Bayes framework as discussed in section 3.5.5 [30]. As a brief reminder the prior for the random effects is assumed to follow a multivariate normal distribution, which along with the likelihood, conditional on the random effect parameters, are utilised to estimate the posterior density of the random effects using Bayes Theorem (Equation 5.32) [128].

$$\begin{aligned} p(\mathbf{u}_i \mid T_i, d_i, \mathbf{y}_i; \boldsymbol{\theta}) &= \frac{\psi(\mathbf{u}_i; \boldsymbol{\theta}) p(T_i, d_i \mid \mathbf{u}_i; \boldsymbol{\theta}) p(\mathbf{y}_i \mid \mathbf{u}_i; \boldsymbol{\theta})}{p(T_i, d_i, \mathbf{y}_i; \boldsymbol{\theta})} \\ &\propto \psi(\mathbf{u}_i; \boldsymbol{\theta}) p(T_i, d_i \mid \mathbf{u}_i; \boldsymbol{\theta}) p(\mathbf{y}_i \mid \mathbf{u}_i; \boldsymbol{\theta}) \end{aligned} \tag{5.32}$$

The prior distribution for the random effects is given by $\psi(\boldsymbol{u_i}; \boldsymbol{\theta})$, with the conditional data likelihood given by $p(T_i, d_i \mid \boldsymbol{u_i}; \boldsymbol{\theta})$. The BLUPs $\hat{\boldsymbol{u}}_{\boldsymbol{i}}$ are equivalent

to the mean (or mode) of the posterior distribution with corresponding variance of $\boldsymbol{u_i}$ utilised to obtain standard error estimates (see Equation 3.17).

5.11.2 Conditional survival predictions

The appeal of implementing the joint model comes from the risk prediction framework that naturally evolves from modelling changes of a biomarker over time in relation to the impact on survival [43; 229]. This will be exploited in analyses carried out in Chapters 6 and 7. The shared random effect dependency between outcomes within the joint model setting, allow for the prediction of individual survival probabilities, conditional on the longitudinal observations to date and subject-specific covariate patterns. In a clinical setting, where a patient's condition is being monitored it would be useful to predict the risk of an event at a given time-point in order to provide intervention if necessary. Alternatively the pattern of longitudinal measurements for a given patient, can inform the timing of the next observation [230]. If predictions forsee a deterioration then observations need to become more frequent, whilst patient improvements would justify less stringent monitoring. An example of a plotted conditional survival probability curve is shown in Figure 5.5.

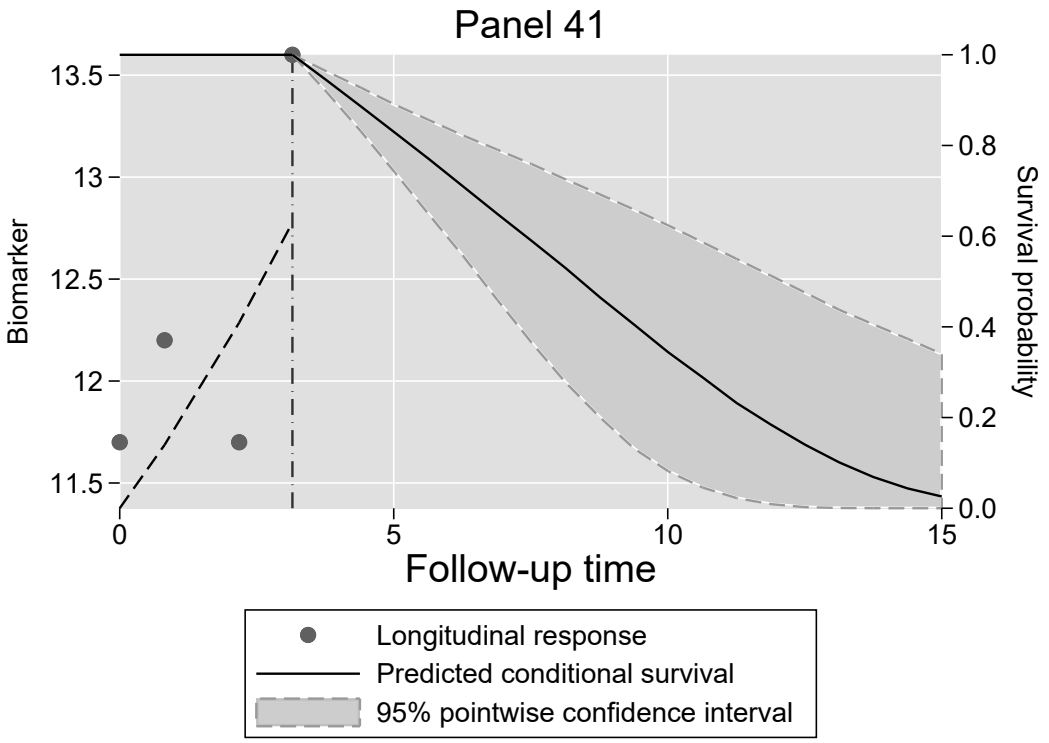


FIGURE 5.5. An example of a conditional survival probability prediction curve based on longitudinal biomarker observations

In predicting the survival probability for a future time $t + \Delta t$ for individual i, we condition on the patient having survived up to time t. In addition, the individual must provide longitudinal observations $Y_i(t) = \{y(s); 0 \leq s < t\}$ up to time t, with corresponding vector of baseline covariates X_i.

Based on the sample $D_n = \{T_i, d_i, y_i; i = 1, \cdots, n\}$ on which the joint model was fitted and for joint model parameters θ^* the conditional survival probability is then given by,

$$\pi_i(t + \Delta t \mid t) = P(T_i \geq t + \Delta t \mid T_i > t, Y_i(t), X_i, D_n; \theta^*), \quad t > 0 \tag{5.33}$$

This definition of the conditional survival probability is dynamic in nature and can be updated as new longitudinal measurements are observed. This results in predictions that are always current. Using the shared random effects framework underpinning both the longitudinal and survival processes, Equation 5.33 can be written as,

$$\begin{aligned}
&P(T_i^{\geq} t + \Delta t \mid T_i > t, Y_i(t); \boldsymbol{\theta}) \\
&= \int P(T_i \geq t + \Delta t \mid T_i > t, Y_i(t), \mathbf{u}_i; \boldsymbol{\theta}) p(\mathbf{u}_i \mid T_i > t, Y_i(t); \boldsymbol{\theta}) d\mathbf{u}_i \\
&= \int p(T_i \geq t + \Delta t \mid T_i > t, \mathbf{u}_i; \boldsymbol{\theta}) p(\mathbf{u}_i \mid T_i > t, Y_i(t); \boldsymbol{\theta}) \\
&= \int \frac{S\{t + \Delta t \mid M_i(t + \Delta t, \mathbf{u}_i, \boldsymbol{\theta}); \boldsymbol{\theta}\}}{S\{t \mid M_i(t, \mathbf{u}_i, \boldsymbol{\theta}); \boldsymbol{\theta}\}} p(\mathbf{u}_i \mid T_i > t, Y_i(t); \boldsymbol{\theta}) d\mathbf{u}_i
\end{aligned} \tag{5.34}$$

Here, $S_i(.)$ refers to the relevant survival function and M_i is the unobserved longitudinal trajectory as estimated by the linear mixed effects model. Using the aforementioned empirical Bayes estimate for u_i (see section 3.5.5), a first order estimator for $\pi_i(t + \Delta t \mid t)$ is shown in Equation 5.35.

$$\pi_i(t + \Delta t \mid t) = \frac{S_i\{t + \Delta t \mid M_i(t + \Delta t, \hat{\mathbf{u}}_i{}^{(t)}, \hat{\boldsymbol{\theta}}); \hat{\boldsymbol{\theta}}\}}{S_i\{t \mid M_i(t, \hat{\mathbf{u}}_i^{(t)}, \hat{\boldsymbol{\theta}}); \hat{\boldsymbol{\theta}}\}} + O(n_i^{-1}) \tag{5.35}$$

As before $\hat{\boldsymbol{\theta}}$ are the maximum likelihood estimates of the joint model, and $\hat{u}_i^{(t)}$ is the mode of the conditional distribution of $\log p(u_i \mid T_i > t, Y_i(t); \hat{\theta})$, and finally $n_i(t)$ denotes the number of longitudinal observations. To calculate corresponding standard errors a Monte Carlo simulation scheme has been proposed by Rizopoulos [43] which uses the following sampling scheme,

For $l = 1, \cdots, L$ repetitions

(1) Draw $\boldsymbol{\theta}^{(l)} \sim N(\hat{\boldsymbol{\theta}}, \hat{var}(\hat{\boldsymbol{\theta}}))$

(2) Draw $\mathbf{u}_i^{(l)} \sim \{\mathbf{u}_i \mid T_i > t, Y_i(t), \boldsymbol{\theta}^{(l)}\}$

(3) Calculate $\pi_i^l(u \mid t) = \frac{S_i\{t+\Delta t|M_i(t+\Delta t,u^{(l)},\boldsymbol{\theta}^{(l)});\boldsymbol{\theta}^{(l)}\}}{S_i\{t|M_i(t,u^{(l)},\boldsymbol{\theta}^{(l)});\boldsymbol{\theta}^{(l)}\}}$

By calculating the first-order estimates using the Monte Carlo estimates of θ and u_i from each repetition, this produces an overall standard deviation estimate, accounting for the combined uncertainty of the likelihood and empirical Bayes estimates. In the context of early pregnancy outcomes, particularly for women who experience recurrent loss, these subject-specific conditional survival predictions provide an adjunct to established clinical practice. For example in women who do not exhibit adequately increasing hCG levels, monitoring could become more frequent to prepare for a loss sooner.

5.12 Model assessment

Building a prediction model requires careful testing for us to understand how well it will perform in practice. It is usual to investigate how well a model is calibrated, i.e. how it performs in predicting event rates. Additionally it is important for a prognostic model to be able to correctly discriminate between individuals who are likely and unlikely to experience the event of interest. Calibration is estimated utilising an extension of the Brier score. Whilst discrimination is estimated by extending the receiver operating characteristic (ROC) curve framework to the joint model setting.

5.12.1 Discrimination

Model discrimination is a measure of how well a model distinguishes or 'discriminates' between individuals who are likely to experience an event or not in the time-frame of interest. This usually involves estimation of the sensitivity and specificity for an optimal threshold value, which estimate the true positives

and true negatives respectively. In other words whether the model can correctly identify an individual who experiences an event and one who does not.

5.12.1.1 Sensitivity and specificity

Based on propositions by Rizopoulos [30] and Pencina *et al.* [231], an estimate of the sensitivity, utilising the predicted probability of an event is shown in Equation 5.36. Here, if the survival probability for an interval $t + \Delta t$, conditional on having survived up to time t, for a subject j falls below a threshold value $c \in 0, 1$ then the model predicts an event.

$$SN_t^{\Delta t}(c) = Pr\{\pi_l(t, \Delta t) \leq c \mid T_l \in (t, t + \Delta t]\} \tag{5.36}$$

The corresponding estimate for specificity follows naturally. The definition assumes that a predicted survival probability over a threshold c indicates that an individual did not experience an event.

$$SP_t^{\Delta t}(c) = Pr\{\pi_l(t, \Delta t) > c \mid T_l > t + \Delta t] \tag{5.37}$$

To account for censoring, weights are included in the formulation of the sensitivity, specificity and the ROC AUC estimates. The two main approaches are through inclusion of model based weights, or inverse probability of censoring weighting (IPCW). The latter involves using non-parametric estimators, usually Kaplan-Meier in the survival context. IPCW provides unbiased estimates even when the model is misspecified, however a requirement is that the model used to specify the weights is correct. This can be difficult to achieve, particularly as the dependency between the mode of censoring and biomarker can be complex. Model based weights, on the other hand, allow for the dependency between the

longitudinal observations and censoring, however the model must be well calibrated to begin with. In the literature, particularly for implementation in R via the JM(bayes) packages, model based weights are proposed due to censoring necessarily depending on the longitudinal biomarker [201; 202]. The proposed weighted estimate for the sensitivity is shown in Equation 5.38.

$$\widehat{SN}_t^{\Delta t}(c) = \frac{\sum_{i:T_i \geq t} I\{\hat{\pi}_i(t + \Delta t \mid t) \leq c\} \times \Omega_i}{\sum_{i:T_i \geq t} \Omega_i} \tag{5.38}$$

with

$$\Omega_i = \begin{cases} 1, & \text{if} \quad T_i \leq t + \Delta t \quad \text{and} \quad d_i = 1 \\ 1 - \hat{\pi}_i(t + \Delta t \mid T_i), & \text{if} \quad T_i \leq t + \Delta t \quad \text{and} \quad d_i = 0 \end{cases}$$

The corresponding estimate for the specificity is,

$$\widehat{SP}_t^{\Delta t}(c) = \frac{\sum_{i:T_i \geq t} I\{\hat{\pi}_i(t + \Delta t \mid t) > c\} \times \Phi_i}{\sum_{i:T_i \geq t} \Phi_i} \tag{5.39}$$

with

$$\Phi_i = \begin{cases} 1, & \text{if} \quad T_i > t + \Delta t \\ \hat{\pi}_i(t + \Delta t \mid T_i), & \text{if} \quad T_i \leq t + \Delta t \quad \text{and} \quad d_i = 0 \end{cases}$$

5.12.1.2 ROC AUC

The ROC AUC follows on from measures of sensitivity and specificity to discriminate between those who will experience an event and those who will not. So for a pair of randomly chosen individuals (l_1, l_2) the AUC is given by,

$$AUC(t, \Delta t) = Pr[\pi_{l_1}(t, \Delta t) < \pi_{l_2}(t, \Delta t) \mid \{T_{l_1} \in (t, t + \Delta t]\} \cap \{T_{l_2} > t + \Delta t\}] \tag{5.40}$$

For a time-frame of interest $(t, t+\Delta t]$, if individual l_1 experiences an event and l_2 does not, then we expect a higher predicted conditional survival probability for l_2 when compared with l_1. To also incorporate censoring, the AUC is decomposed in Equation 5.41.

$$\widehat{AUC}(t, \Delta t) = \sum_{w=1}^{4} \widehat{AUC}_w(t, \Delta t) \tag{5.41}$$

The decomposition involves the following pairs of patients, where $\Omega^{(1)}_{l_1 l_2}$ are those comparable individuals, whilst $\sum_{w=2}^{4} \widehat{AUC}_w(t, \Delta t)$ are not comparable due to one or both individuals being censored $(\Omega^{(w)}_{l_1 l_2}, w = 2, 3, 4)$.

$$\Omega^{(1)}_{l_1 l_2}(t) = [\{T_{l_1} \in (t, t + \Delta t]\} \cap \{d_{l_1} = 1]\} \cap \{T_{l_2} > t + \Delta t\}]$$

$$\Omega^{(2)}_{l_1 l_2}(t) = [\{T_{l_1} \in (t, t + \Delta t]\} \cap \{d_{l_1} = 0]\} \cap \{T_{l_2} > t + \Delta t\}]$$

$$\Omega^{(3)}_{l_1 l_2}(t) = [\{T_{l_1} \in (t, t + \Delta t]\} \cap \{d_{l_1} = 1\}] \cap [\{T_{l_1} < T_{l_2} \leq t + \Delta t\} \cap \{d_{l_2} = 0\}]$$

$$\Omega^{(4)}_{l_1 l_2}(t) = [\{T_{l_1} \in (t, t + \Delta t]\} \cap \{d_{l_1} = 0\}] \cap [\{T_{l_1} < T_{l_2} \leq t + \Delta t\} \cap \{d_{l_2} = 0\}]$$

The AUC (Equation 5.42), where $I(.)$ is the indicator function, estimates the proportion of individuals of those comparable at time t who are deemed "concordant" [205]. This essentially means that for a randomly selected pair of individuals if the one with a greater survival probability does not experience the event and the one with a lower probability does, they are said to be concordant. When $w = 1$ then $\hat{K}_1 = 1$, as this represents a comparable pair. For the other

three pairs, the AUC estimate is weighted by the extent of the comparability of the pair. The indicated weights are $\hat{K}_2 = 1 - \hat{\pi}_{l_1}(t, \Delta, t)$, $\hat{K}_3 = \hat{\pi}_{l_2}(t, \Delta t)$ and $\hat{K}_4 = \{1 - \hat{\pi}_{l_1}(t, \Delta t)\} \times \hat{\pi}_{l_2}(t, \Delta t)$.

$$\widehat{AUC}_w(t, \Delta t) = \frac{\sum_{l_1=1}^{n} \sum_{l_2=1; l_2 \neq l_1}^{n} I\{\hat{\pi}_{l_1}(t, \Delta t) < \hat{\pi}_{l_2}(t, \Delta t)\} \times I\{\Omega_{l_1 l_2}^{(w)}(t)\} \times \hat{K}_w}{\sum_{l_1=1}^{n} \sum_{l_2=1; l_2 \neq l_1}^{n} I\{\Omega_{l_1 l_2}^{(w)}(t)\} \times \hat{K}_w} \tag{5.42}$$

$$\widehat{AUC}(t, \Delta t) = \sum_{w=1}^{4} \hat{AUC}_w(t, \Delta t) \tag{5.43}$$

5.12.2 Calibration

The most common measure used to assess model calibration in survival contexts is the Brier score. This compares the predicted survival probability for a given individual with whether an event was observed or not, using a suitable loss function [232]. The sum across individuals then tells us how well the model is predicting event rates. In the joint model context, the dynamic nature of the longitudinal marker is also incorporated through estimation of subject-specific survival predictions. This essentially means the model can be assessed for how well it is calibrated at various time time-points for a given individual. The following is the expected error of prediction for an individual, utilising a quadratic loss function.

$$BS(t + \Delta t \mid t) = E\{N_i(t + \Delta t) - \pi_i(t + \Delta t \mid t)\}^2 \tag{5.44}$$

Here $N_i(t) = I(T_i^* > t)$ is the observed event status at time t, with $\pi_i(t+\Delta \mid t)$ is the conditional probability of survival at time $t + \Delta$ having survived up to t. This estimate however, does not account for censoring. Henderson *et al.* [233] instead have proposed an estimate which incorporates both right censoring and

takes into account the longitudinal marker information up to t. Following the notation used by Andrinopoulou *et al.* [205] the prediction error estimate is shown in Equation 5.45.

$$\begin{aligned}\widehat{PE}(t+\Delta t \mid t) = \{R(t)\}^{-1} \sum_{i:T_i \geq t} \Big\{ & I(T_i > t+\Delta t)\{1-\hat{\pi}_i(t+\Delta t \mid t)\}^2 \\ & + d_i I(T_i < t+\Delta t)\{0-\hat{\pi}_i(t+\Delta t \mid t)\}^2 \\ & + (1-d_i) I(T_i < t+\Delta t)\Big[\hat{\pi}_i(t+\Delta t \mid T_i)\{1-\hat{\pi}_i(t+\Delta t \mid t)\}^2 \\ & + \{1-\hat{\pi}_i(t+\Delta t \mid T_i)\}\{0-\hat{\pi}_i(t+\Delta t \mid t)\}^2\Big]\Big\}\end{aligned} \tag{5.45}$$

The number of individuals at risk at time t is denoted by $R(t)$. The term d_i indicates whether an individual was censored ($d_i = 0$) or experienced an event ($d_i = 1$). $I(T_i > u)\{1-\hat{\pi}_i(t+\Delta t \mid t)\}^2$ refers to the contribution of those individuals who remain event-free after the time of interest $t+\Delta t$. Those who experience an event before $t+\Delta t$ are represented by $d_i I(T_i < t+\Delta t)\{0-\hat{\pi}_i(t+\Delta t \mid t)\}^2$. The final terms indicate those individuals who are censored in the interval $[t, t+\Delta t]$. This has been adopted as a measure of model calibration for the analysis in Chapter 7.

5.13 The joint latent class model

An alternative approach to the SREM joint model is the joint latent class model (JLCM) which favours a latent class structure. The latent classes are assumed to describe all of the correlation between the biomarker profiles and the risk of the event [234]. The resulting assumption is that subgroups, or latent classes, of individuals have similar longitudinal marker trajectories and a similar risk of experiencing an event [188; 235]. The JLCM can provide a less computationally

expensive alternative to predict, or investigate the association between the longitudinal and time-to-event outcomes without making specific assumptions about the dependency structure [234].

A latent class probability, π_{ig} is first defined, estimating the probability of an individual i belonging to a particular class g. This probability is conventionally modelled via a multinomial logistic regression model with explanatory variables. Each latent class has a corresponding class-specific trajectory function and risk of event. As with the SREM, a linear mixed model as described in 3.5 is utilised to model the repeated biomarker observations for each class.

$$Y_i(t_{ij})|_{c_i=g} = \mathbf{X}_{li}(t_{ij})^T\boldsymbol{\beta}_g + \mathbf{Z}_i(t_{ij})^T\mathbf{u}_{ig} + \eta_i(t_{ij}) \tag{5.46}$$

The vector of random effects are normally distributed with $\mathbf{u}_i \sim \sum_{g=1}^{G} \pi_{ig} N(\mu_g, \mathbf{B}_g)$, with variance-covariance matrix B_g. This can be the same over all classes or specific to a class. The vector of measurement errors are similarly normally distributed, where $\boldsymbol{\eta}_i = (\eta_i(t_{i1}), \cdots, \eta_i(t_{in_i}))^T \sim N(0, \boldsymbol{\Sigma}_i)$, with a diagonal variance-covariance matrix $\boldsymbol{\Sigma}_i$ of homoscedastic independent errors.

The class-specific proportional hazards survival submodel is shown below. The baseline hazard can either be stratified by the class structure or baseline hazards across classes are assumed proportional.

$$h_i(t \mid c_i = g; \zeta_g, \boldsymbol{\phi}_g) = h_{0g}(t; \zeta_g) \exp\{\mathbf{V}_{ei}(t)^T\boldsymbol{\phi}_g\} \tag{5.47}$$

Estimation of the JLCM is via maximum likelihood, using the Marquardt algorithm and does not require numerical integration as in the SREM context [196; 236]. However to select the number of classes and to ensure convergence the estimation of the JLCM must be performed several times. For the SREM a single random-effects structure is implemented to explain the within-subject

biomarker correlation and the underlying dependency between between the longitudinal and survival outcomes. In the JLCM definition, however, the random effects only describe the correlation between repeated measurements whilst the latent class structure models the dependency between the two outcomes [236]. The association between the outcomes in the SREM is predetermined by the chosen association structure (see section 5.5). The chosen structure should be correct to produce appropriate predictions. The JLCM makes less stringent assumptions about the dependency between longitudinal and survival outcomes. Though stratification over latent classes allows the baseline hazard to vary with the marker, this can also lead to a substantial increase in the number of parameters estimated. Using a greater number of latent classes can lead to fewer individuals per class, which can also affect validity of predictions [236].

Though many aspects of the JLCM could be applied to the early pregnancy outcomes data introduced in Chapters 6 and 7, the focus for this thesis is the SREM. This is not only for prediction purposes but also to appropriately quantify the association between longitudinal hCG and time-to-miscarriage. With relatively small sample sizes splitting the data further into classes seems inadvisable and so the JLCM will not be considered further here, though classification of trajectories has received attention in the literature [45].

5.14 Landmark models

Up to now we have discussed joint models with varying underlying dependency structures, centring around the LMM and a proportional hazards survival model. Another arguably simpler approach is landmarking.

Landmark models are useful when the focus of the analysis is to predict the future risk of an event. The initial step is to select a landmark reference time at

which predictions will be made - for landmark-age models, this might be an age of 50 years for example [109]. In the basic landmarking framework a series of Cox models (see section 4.6.1) are fitted for various prediction times so that only the sample of subjects still at risk at the landmark reference time are considered, along with a summary of the longitudinal biomarker up to that point [33; 229]. The naive landmark model utilises the most recent observation prior to the landmark time, or last observation carried forward (LOCF) within the Cox model [237]. A slightly more complex approach, designed to also incorporate measurement error of the observed biomarker, includes predicted values from a linear mixed effects model in the landmark Cox model, paralleling the two-stage model approach discussed in section 5.3.3 [229]. Even with this improvement however there is the issue of the age of the biomarker value when compared with the prediction time[237]. Survival predictions provide an approximation of risk for a fixed future period or horizon time t_{hor}. Administrative censoring is applied at this point to encourage robust estimates even when the proportional hazards assumption is not met [237]. The basic landmarking approach estimates several of these landmark models for varying landmark reference points s [238]. It is this repeated model estimation which makes landmarking an inefficient method when compared with joint models [229].

A super landmark models fits a lone stratified Cox survival model to stacked landmark datasets which are created for each landmark time[239]. A continuous model is specified for the regression parameters, allowing them to vary with the landmark times smoothly [238; 240]. Linear models, polynomials and linear mixed effects models have been suggested as the continuous model [109; 237; 238]. The stratified Cox model on landmark times s is then estimated via a pseudo partial-likelihood, which is equivalent to the sum of the partial likelihoods of the

Cox models at each single landmark time-point [238]. A variation of the super landmark model also allows landmark specific baseline hazards [239].

Proponents of landmarking over joint models prefer the relative simplicity of its implementation, which avoids the computationally taxing numerical integration over the the random effects. However, landmark models remain inefficient when compared with the joint model due to the repeated estimation of models at different landmark times. The greatest issue concerns the acknowledged 'staleness' of the biomarker observation as more time elapses between the landmark time and the last observed measurement, as this results in inaccurate conditional survival predictions [237]. When the joint model is correctly specified, which admittedly can be difficult to achieve, then it outperforms the landmark model. However, as the level of misspecification increases the landmark model performance outranks that of the joint model [237].

5.15 Discussion

When it is hypothesized that a particular biomarker is associated with a time-to-event endpoint there is an incentive to estimate both the longitudinal profile of the biomarker and the risk of event occurrence. Although established methods can be used to model the association between the two types of outcomes, they come with their disadvantages. Neither the survival model with time-varying covariate, nor the two-stage model can avoid the pitfalls of measurement error and/or intermittently measured biomarkers. It is clear then that a joint model should be considered for their ability to appropriately describe the association between a longitudinally measured biomarker and time-to-event outcome with efficiency. Joint models are an attractive proposition when the ultimate aim is

for prediction, particularly to enable dynamic monitoring of subjects via up-to-date event risk estimates based on individual longitudinal profiles and data. The shared parameter model has an intuitive dependency structure and though more computationally burdensome than alternatives, such as the latent class structure or landmarking, a well defined model can give good association estimates in tandem with predictions. In fact, software developments within joint models across platforms now allow the fitting of increasingly complex models with relative ease. Methods for model assessment can also now be used to evaluate the developed joint model, which brings them one step closer to being used in clinical practice. This naturally brings us to application of these models to the pregnancy setting, the results (and challenges) of which will be detailed in Chapters 6 and 7.

Chapter 6

Application of joint modelling methods to early pregnancy outcomes

6.1 Chapter overview

In this chapter I apply the joint longitudinal-survival model introduced in Chapter 5 to a dataset of 129 pregnant women. The dataset was collected by SPD Development Company Ltd, with the view to prospectively follow up women attempting to conceive, from conception through to pregnancy outcome at the close of the study. The data has been analysed previously utilising the simpler two-stage model approach (section 5.3.3) [44]. The aim here is to replicate this analysis and formally extend to the shared parameter framework of the joint model. Ultimately this will give an estimate of the association between hCG and early miscarriage, which accounts for measurement error and gives more realistic uncertainty estimates [28; 38]. Comparisons between association parameters from naive methods and the joint model will be made. Focus will be on appropriately modelling the longitudinal trajectory for hCG utilising the LMM (section 3.5) and addressing the non-linear slope. The baseline hazard function of the survival submodel will be modelled using various distributional assumptions as well as restricted cubic splines to establish best fit (see Chapter 4). Model selection procedures will be carried out to identify variables associated with miscarriage

that will subsequently be included in the survival component model. Through estimation of conditional survival probabilities from the joint longitudinal-survival model, the potential of dynamic monitoring in early pregnancy will be assessed.

This analysis has been published in Nature Scientific Reports (see Appendix A) [241].

6.2 Introduction

It has been established in Chapter 2 that human chorionic gonadotrophin, colloquially the 'pregnancy hormone', is not routinely collected during pregnancy. Once pregnancy has been confirmed by means of a hCG observation above 25 mIU/ml, there is usually little call for any further use of it, unless an imminent early miscarriage is suspected. In such cases, changes in hCG can serve as a first port of call to establish whether the pregnancy is ongoing (see section 2.4) [17]. However, this involves a short term, usually two-day, ratio of hCG values to confirm the level of rise or decrease. There is potential, however, for hCG to be utilised to a much fuller extent in the early pregnancy outcome setting. With more frequent observation of hCG, a longitudinal profile of pregnancy progression can be obtained. Certainly the literature has shown that profiles of failing and viable pregnancies are distinct (see section 2.4)[18; 242]. The question lies in whether hCG can aid prospective monitoring in early pregnancy to predict an eventual outcome. Research into whether biomarkers can be monitored over an extended period of time with a view to predicting miscarriage remain in their infancy [15]. Where this has been explored, the data analysed has been from an assisted pregnancy setting, with focus on serum hCG and fewer observations per individual [11; 45; 46]. Furthermore, the focus of these analyses was to classify similar abnormal and normal hCG trajectories utilising the Bayesian framework,

rather than prediction. For this thesis, comprehensive hCG data for women who were prospectively followed up from the beginning of their menstrual cycle, through conception and to early pregnancy outcome, are available for analysis. This dataset is unique in its focus on prospective collection of urinary hCG observations, which are arguably cheaper and less invasive to collect. The singular nature of the data on which the methods will be applied allow the investigation of whether urinary observations can be used for meaningful monitoring in early pregnancy on an individual level. If hCG can be exploited in this way, tracking could lead to earlier identification of a miscarriage, with the potential to save women from unnecessary intervention (see section 2.3.2).

The repeated collection of a continuous biomarker, such as hCG, over time gives rise to intermittently observed longitudinal data which are subject to measurement error as discussed in section 3.5. Separate analyses of the two types of data utilise linear mixed effects models, with time-to-event outcomes analysed using survival models (see sections 3.5 and 4.6.3 respectively). However, when interest lies in quantifying the association between the repeatedly measured biomarker and time-to-event outcome, separate analyses ignore the dependency between the longitudinal and time-to-event processes [38]. The dependency can be addressed through the shared parameter structure of the joint longitudinal-survival model introduced in Chapter 5. This allows the association to be appropriately modelled, whilst taking into account the intermittent nature of observations and measurement error. The model allows for the prediction of subject-specific conditional survival probabilities (see section 5.11.2), which can aid monitoring and be used to plan the schedule of observations [28].

6.3 Methods

To quantify the association between urinary hCG and time-to-miscarriage a joint longitudinal-survival model was fitted to early pregnancy outcome data, combining the LMM and a PH survival submodel (see section 5.4.2). Subject specific predictions, discussed in section 5.11.2, were obtained from the fitted model to establish the predictive capabilities of the fitted model.

6.3.1 Data source

The data for this analysis were collected by SPD Development Company Ltd as part of a study which recruited women who were attempting to conceive. Participants were asked to collect daily early morning urine samples for their entire menstrual cycle and up to 28 days after the day of their missed period if they successfully conceived. Women recruited were aged between 18-45 years and actively trying to conceive and were allowed to remain in the study for a maximum of three cycles. Women were not excluded on the basis of challenged fertility, although no women recruited to the study were undergoing fertility treatment. All women in the study were offered complimentary use of Clearblue ovulation strips in order to time intercourse to maximise conception. Not all urine samples from women who conceived underwent laboratory testing. All women who miscarried as part of the study and a random sample of healthy pregnancies (double the number of losses) were selected for testing and subsequently analysed.

Other maternal data collected included age in years, ethnicity, level of education, occupation, previous number of pregnancies, previous number of live births, number of months trying to conceive, binary previous miscarriage variable, whether they were on any infertility treatment and shortest, usual and longest

menstrual cycle length in days. Pregnancy status was further categorised into three groups, biochemical miscarriage, early miscarriage and healthy pregnancy.

6.3.2 Time variables

Several time variables were associated with the day hCG was observed. These were,

- day in current cycle, where day one is the first day of the period
- day relative to expected period, where day zero is the beginning of the next cycle
- day relative to conception, where day zero is the estimated day of ovulation

Each was assessed for its suitability for analysis. It was essential that the variable was common amongst the women and so the analysis was conducted on the time since conception scale. Although this is not a realistic anchor in the natural pregnancy setting, it serves a purpose for this analysis. For one it allows the full scope of hCG measurements to be included in the survival model by maintaining a time variable of positive integers, something which would be impossible if day relative to the missed period was utilised. Secondly, the day of ovulation can be estimated reasonably well as the day after the LH surge, or approximately two weeks prior to the day of the expected period [243; 244; 245]. This would be tracked in an assisted pregnancy setting, which may be a route to conception for women who have experienced multiple miscarriages. This provides motivation for modelling on the time since conception scale. Figure 6.1 below illustrates how these timelines relate to each other.

Day in cycle																											
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Day relative to missed period																											
-28	-27	-26	-25	-24	-23	-22	-21	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1
Day relative to conception																											
-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

FIGURE 6.1. Time variables used for analysis

6.3.3 Modelling longitudinally observed hCG

When hCG observations are observed for the entirety of a pregnancy they approximately follow a normal distribution (see Figure 2.2). However, as samples were collected for only part of each pregnancy the hCG observations were negatively skewed. As a result hCG was modelled on the natural log scale.

A longitudinal hCG trajectory was built by fitting an LMM, discussed in Chapter 3, to the log hCG response variable. This was utilised as the basis of the longitudinal component submodel when fitting the two-stage model (see section 5.3.3) and joint longitudinal-survival model (see section 5.4.2). A linear and quadratic time term were chosen to model log hCG over time, consistent with the Marriott *et al.* [44] analysis. A random intercept and linear slope were included to allow for varying baseline log hCG values and slopes. Where possible a random quadratic slope term was also included. The unstructured random effects covariance structure was chosen (see section 3.5.1) to allow estimation of unique variances for each random effect. After each addition of fixed or random effect, models were compared using likelihood ratio tests and the AIC and BIC (see section 4.6.6). Consistent with the analysis conducted by Marriott *et al.* [44] no further covariates were included in the longitudinal submodel.

6.3.4 Building the survival submodel

The association between miscarriage and longitudinal hCG was modelled using a PH survival model. In order to capture the lag between theoretical conception at ovulation and the first measurable log hCG observation, a delayed-entry model was fitted to model time-to-miscarriage. Various distributional assumptions for the baseline hazard function were explored, including an exponential, Weibull distribution and the use of RCS functions via an FPM (see Chapter 4). Ideally, all models would have captured the shape of the baseline hazard using RCS functions, however due to increasing model complexity at the joint model stage Weibull joint models were fitted where the flexible parametric joint model did not converge. Joint models were fitted with both current value and the first derivative association structures (see section 5.5). Estimates from the joint longitudinal-survival model were compared to estimates from the time-varying covariate and two-stage model approaches introduced in sections 5.3.2 and 5.3.3.

The original analysis by Marriott *et al.* [44] included continuous age, longest cycle length, and time taken for hCG to reach 25 mIU/ml in the survival sub-model. For the re-analysis a forward stepwise procedure for model selection was carried out at the 5% significance level. Models were compared using the AIC and BIC. The final survival model included covariates age and usual cycle length, as well as the longitudinal log hCG observations. Non-linear effects of age and log hCG were considered. Interactions between log hCG and age as well as age and cycle length were investigated. The PH assumption was tested by including an interaction for age and usual cycle length with log time in the model and comparing to the main model using a likelihood ratio test. Martingale residuals, estimated for the null and full model were plotted against age, usual cycle length

and log hCG in turn to establish non-linear effects of each variable. Deviance residuals were examined for outlying observations.

6.3.5 Subject-specific conditional survival probabilities

Conditional subject-specific survival probabilities for pregnancy viability were obtained from the fitted current value association model for selected pairs of biochemical losses, early losses and healthy pregnancies. For each pregnancy, conditional survival was predicted for two or three-day intervals from the last observed follow-up time. These were illustrated graphically.

6.4 Results

A total of 1505 eligible US women were recruited to the study. Of these women, 250 became pregnant during the course of the study. The majority (n=178, 71%) of these pregnancies were viable singleton pregnancies with a single (0.04%) multiple pregnancy. A total of 44 (17.6%) women suffered miscarriages. During the study 14 women withdrew prior to completion and 13 women were subsequently lost to follow-up. The dataset used for analysis consists of 85 randomly selected viable singleton pregnancies and 44 miscarried pregnancies.

6.4.1 Data exploration

Baseline demographics are presented in Table 6.1 as mean(SD) and n(%) for continuous and categorical variables respectively, unless otherwise stated.

TABLE 6.1. Maternal demographics for healthy and miscarried pregnancies

Variables	Healthy pregnancies (n=85)	Miscarried pregnancies (n=44)	Overall (n=129)
Age, years	29.95 (4.15)	32.34 (4.60)	30.77 (4.44)
Ethnicity, n(%)			
White	75 (88.24)	34 (77.27)	109 (84.50)
Black	3 (3.53)	7 (15.91)	10 (7.75)
Asian	4 (4.71)	2 (4.55)	6 (4.65)
Mixed	3 (3.53)	1 (2.27)	4 (3.10)
Education, n(%)			
High School	4 (4.71)	2 (4.55)	6 (4.65)
Graduate	69 (81.18)	28 (63.64)	97 (75.19)
Postgraduate	12 (14.12)	14 (31.82)	26 (20.16)
Occupation, n(%)			
Homemaker	12 (14.12)	3 (6.82)	15 (11.63)
Student	1 (1.18)	1 (2.27)	2 (1.55)
Skilled Labourer	2 (2.35)	2 (4.55)	4 (3.10)
Office Administrator	8 (9.41)	5 (11.36)	13 (10.08)
Professional	60 (70.59)	31 (70.45)	91 (70.54)
Other	2 (2.35)	2 (4.55)	4 (3.10)
Cycle length	29.94 (2.95)	28.66 (3.21)	29.50 (3.09)
Previous pregnancies	1.00 (1.05)	1.11 (1.15)	1.04 (1.08)
Previous live births	0.62 (0.76)	0.70 (0.88)	0.65 (0.80)
Time to conceive, months, median(Q_1, Q_3)	3 (1, 5.5)	3 (1.25, 5.75)	3 (1, 5.5)
Previous miscarriage, n(%)	11 (12.94)	4 (9.76)	15 (11.90)

All values are mean (SD) unless otherwise stated
Q_1: lower quartile Q_3: upper quartile

On average, women who miscarried were slightly older (mean(SD): 32.3(4.6) years) than women who went on to have viable pregnancies (mean(SD): 30.0(4.2) years). For both pregnancy groups the majority of women were from a white ethnic background (88.2% and 77.3% for healthy and miscarried pregnancies respectively). The majority of women in both groups achieved a graduate or postgraduate level of education, (95.3% in the healthy and 95.5% in the miscarried group). However this was distributed differently across the groups with more women who experienced healthier pregnancies gaining graduate degrees (81.2%) than women who miscarried (63.6%). A similar percentage of women classed their occupations as professional across the healthy (70.6%) and miscarriage (70.5%) groups. The average usual cycle was slightly longer at 29.9 days (SD: 2.95) for women who had healthy pregnancies than the 28.7 days (SD: 3.21) for pregnancies which ended in loss. The mean number of previous pregnancies was approximately 1 (mean(SD): 1.00(1.05), 1.11(1.15) for healthy and miscarried pregnancies respectively) in both groups. The median time to conception was 3 months (IQR: 4.5) for both groups. A slightly higher percentage of women who had viable pregnancies experienced a previous miscarriage than those who miscarried (12.9% versus 9.8%).

Of the women who miscarried 18 (14.2%) suffered biochemical losses, a loss after implantation but before the day of the missed period, and 24 (57.1%) women suffered early miscarriages, after the day of the missed period. Two women who miscarried did not complete the urine collection and were not included in the joint modelling analysis.

6.4.2 hCG Trajectories

The remaining 127 women contributed repeated hCG measurements. The median number of hCG observations for women who miscarried was 19.5 (Q_1, Q_3:

9.25, 26.0) and higher at 25 (Q_1, Q_3: 22.0, 26.5) for women who experienced viable pregnancies. Individual hCG trajectories plotted by pregnancy status are shown in Figure 6.2.

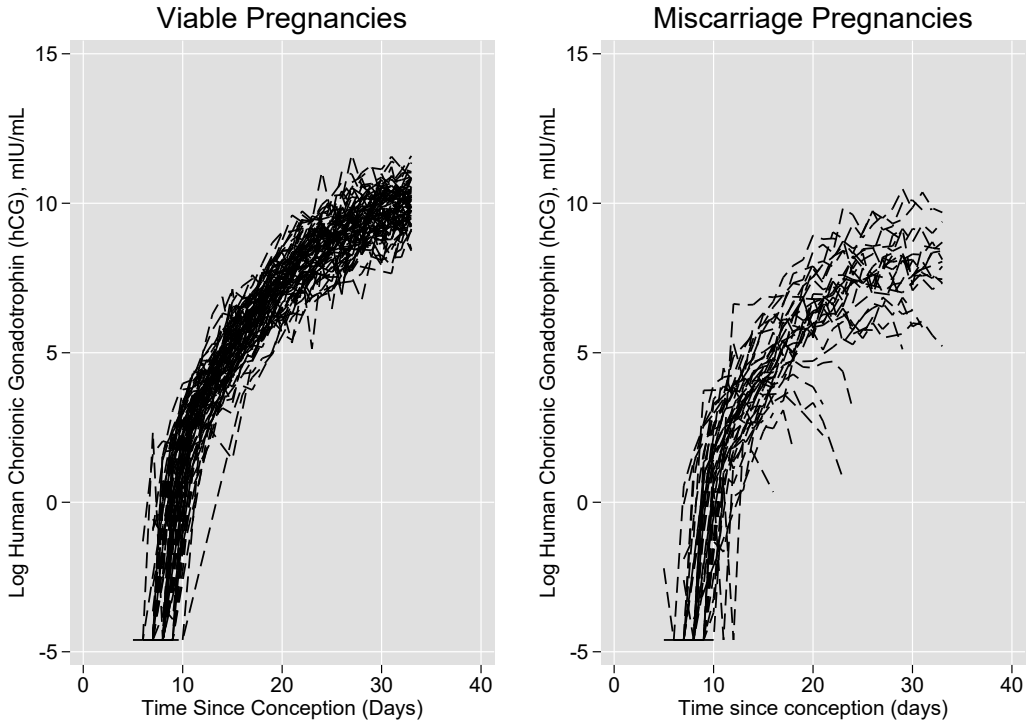


FIGURE 6.2. Log hCG trajectories for women who experienced healthy pregnancies or miscarriage

Generally the profile of log hCG measurements for viable pregnancies were distinguishable from those who miscarried. The hCG measurements for viable pregnancies tended to follow the same general trajectory; an initial rise after conception which continued through the first three weeks of pregnancy before slowing in rise. There was greater variation in hCG profiles for women who miscarried. Women who miscarried presented with an initial rise after conception, however some women experienced a sharp drop in hCG, others simply experienced a more gradual rise in hCG in comparison with women who had healthy pregnancies.

Figure 6.3 shows the log hCG profiles for women who experienced biochemical pregnancies and early losses. The trajectories for biochemical pregnancies show a drop in hCG before the miscarriage occurs, and though this is also the case for some early miscarriages, the majority of trajectories follow a slower rate of increase.

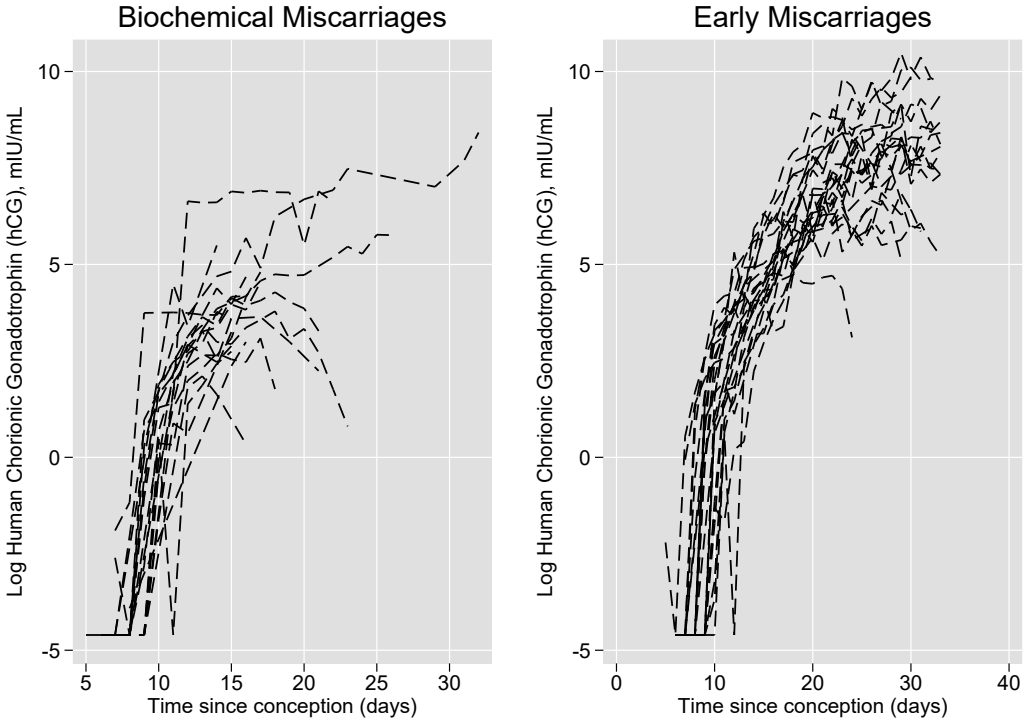


FIGURE 6.3. Log hCG trajectories for women whose pregnancies resulted in a biochemical or early miscarriage)

6.4.3 Survival modelling with a time-varying covariate

Overall Kaplan-Meier survival estimates are shown in Figure 6.4. The first hCG observation (and therefore confirmed pregnancy) was observed at day four, prompting the first participant to enter the study. A further ten days elapsed

before the first event occurred at day 15 relative to conception. The last observed exit was at day 33.

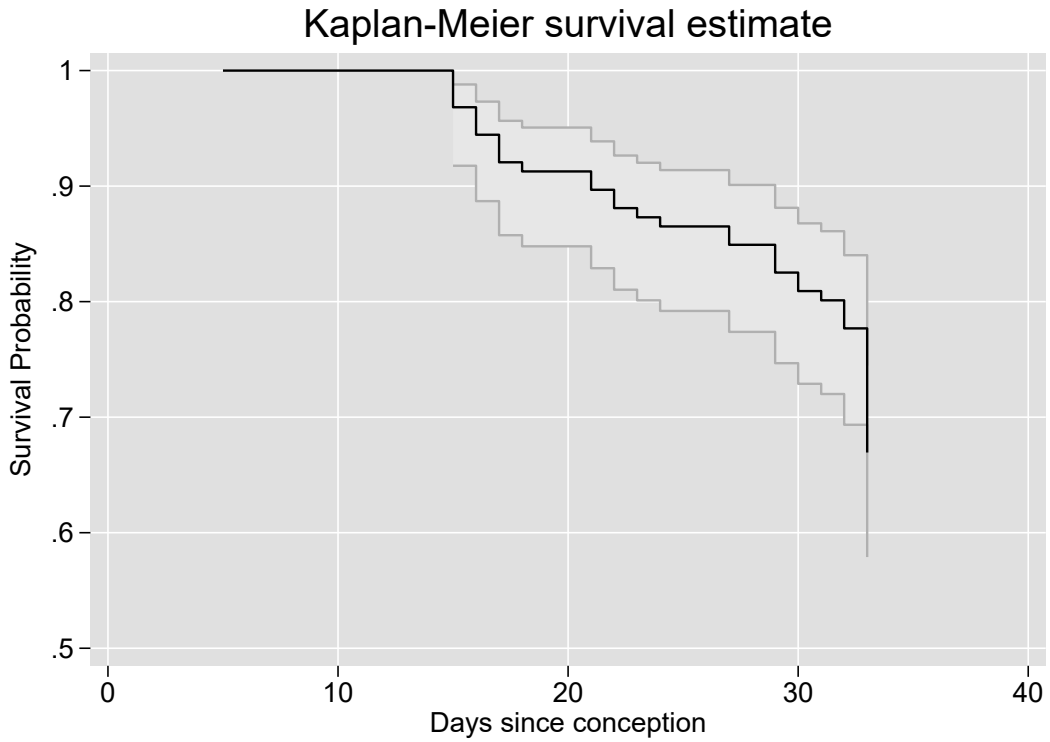


FIGURE 6.4. Kaplan-Meier Curve showing Survival Probability across all Women

Figure 6.5 shows the predicted conditional survival curves for several baseline hazard functions, which account for delayed entry. A three-RCS function was most effective in modelling the underlying baseline hazard. This is confirmed by the model fit statistics presented in Table 6.2. The AIC was very similar for the Weibull and two-RCS models (AIC of 141.723 and 141.510 respectively), whereas the AIC for the three-RCS model was lower at 119.17. Ideally, all models would have captured the shape of the baseline hazard using RCS functions, however due to convergence issues at the joint model stage Weibull joint models were fitted where models with a flexible baseline hazard did not converge.

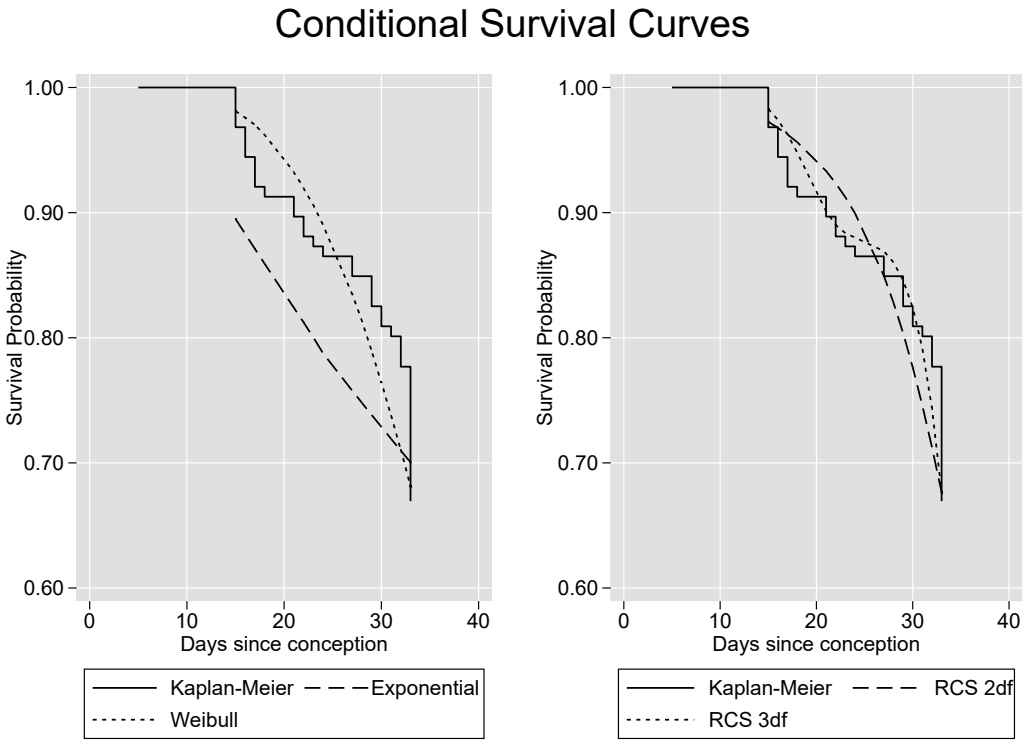


FIGURE 6.5. Survival probability curves for various distributions and restricted cubic spline modelled baseline hazards

TABLE 6.2. Model fit estimates for the exponential, Weibull, and flexible parametric baseline hazard survival models

Baseline hazard	Log likelihood	df	AIC	BIC
Exponential	-84.423	1	170.846	173.682
Weibull	-68.862	2	141.723	147.396
Two RCS*	-67.755	3	141.510	150.019
Three RCS∼	-55.585	4	119.170	130.515

*1 internal and 2 boundary knots
∼2 internal and 2 boundary knots
df: degrees of freedom

6.4.3.1 Model covariates

Age and usual cycle length were significantly associated with miscarriage at the 5% significance level and were included in the survival model. No further covariates were found to be statistically significant at the 5% level. Model fit statistics for each addition to the Weibull and FPM are given in Table 6.3.

TABLE 6.3. Model fit estimates for addition of age and usual cycle length

Weibull model	Log likelihood	df	AIC	BIC
Null model	-68.579	2	141.158	153.012
Log hCG model	-24.633	3	55.266	72.890
Add age, years	-22.036	4	52.073	75.572
Add cycle length, days	-19.290	5	48.580	77.954
Flexible parametric model*	Log likelihood	df	AIC	BIC
Null model	-52.431	4	112.862	136.570
Log hCG model	-6.181	5	22.363	51.736
Add age, years	-3.234	6	18.467	53.716
Add cycle length, days	0.991	7	12.018	53.141
*2 internal and 2 boundary knots df: degreees of freedom				

Additions of interactions between cycle length and age as well as log hCG and age did not improve the fit of the model when compared to the basic model, based on likelihood ratio tests applied to nested models. Model fit statistics for each interaction addition are shown in Table 6.4.

TABLE 6.4. Model fit estimates for the addition of interaction terms

Weibull model	Log likelihood	df	AIC	BIC
No interactions	-19.290	5	48.580	77.954
Age x cycle length	-18.940	6	49.880	85.128
Age x log hCG	-18.940	6	49.880	85.128
Flexible parametric model*	Log likelihood	df	AIC	BIC
No interactions	0.991	7	12.018	53.141
Age x cycle length	1.391	8	13.218	60.216
Age x log hCG	1.005	8	13.990	60.988
*2 internal and 2 boundary knots				
df: degress of freedom				

Non-linear effects of age and log hCG were also investigated. Model fit statistics are shown in Table 6.5. A quadratic age term did not appreciably improve fit, however addition of the quadratic log hCG term suggests that the relationship between miscarriage and log hCG is not linear. Even so log hCG was modelled linearly at this stage, to allow for comparisons at the joint model stage, where the software used (`stjm`) did not allow for non-linear associations for a biomarker to be incorporated.

TABLE 6.5. Model fit estimates for the addition of non-linear effects of age and log hCG

Weibull model	Log likelihood	df	AIC	BIC
Linear age	-19.290	5	48.580	77.954
Quadratic age	-19.152	6	50.304	85.553
Quadratic log hCG	-13.210	6	38.421	73.669
Restricted cubic splines*	Log likelihood	df	AIC	BIC
Linear log hCG	0.991	7	12.018	53.141
Quadratic age	1.056	8	13.887	60.885
Quadratic log hCG	7.694	8	0.612	47.609
*2 internal and 2 boundary knots				
df: degrees of freedom				

Estimates from the fitted Weibull model and FPM are presented in Table 6.6.

TABLE 6.6. Model estimates from Weibull and flexible parametric* models modelling time to miscarriage using a time-varying log hCG covariate

Weibull model	Hazard Ratio	95% CI
Age, years	1.082	1.008, 1.163
Usual cycle length, days	0.880	0.769, 1.008
Time varying log hCG	0.419	0.357, 0.491
Flexible parametric model*	**Hazard Ratio**	**95% CI**
Age, years	1.091	1.014, 1.175
Usual cycle length, days	0.828	0.713, 0.961
Time varying log hCG	0.392	0.334, 0.460

*2 internal and 2 boundary and knots
CI: Confidence Interval

Age and log hCG were significantly associated with time-to-miscarriage at the 5% level across both models, with cycle length also statistically significant when modelled using the FPM. Estimates were similar across the two models. A one-year increase in maternal age at conception inferred a 8.3% increase (HR: 1.083 95% CI: $1.008, 1.163$) in the rate of miscarriage for the Weibull model and a 9.1% increase (HR: 1.091 95% CI: $1.014, 1.175$) in rate of miscarriage for the FPM. A one-day increase in a woman's usual cycle length was associated with a 12.0% decrease (HR: 0.880 95% CI: $0.769, 1.008$) in the rate of miscarriage for the Weibull model and a corresponding 17.2% decrease (HR: 0.828 95% CI: $0.713, 0.961$) in the rate of miscarriage for the FPM. Increases in log hCG produced the greatest change in the rate of pregnancy loss. A one-unit increase in log hCG resulted in a corresponding 58.1 % (HR: 0.419 95% CI: $0.357, 0.491$) and 60.8% reduction (HR: 0.392 95% CI: $0.334, 0.460$) in the rate of miscarriage for the Weibull and flexible parametric models respectively. This shows strong evidence of the association between hCG and time-to-miscarriage, whilst using the maximum available amount of hCG biomarker information.

6.4.3.2 Model checks

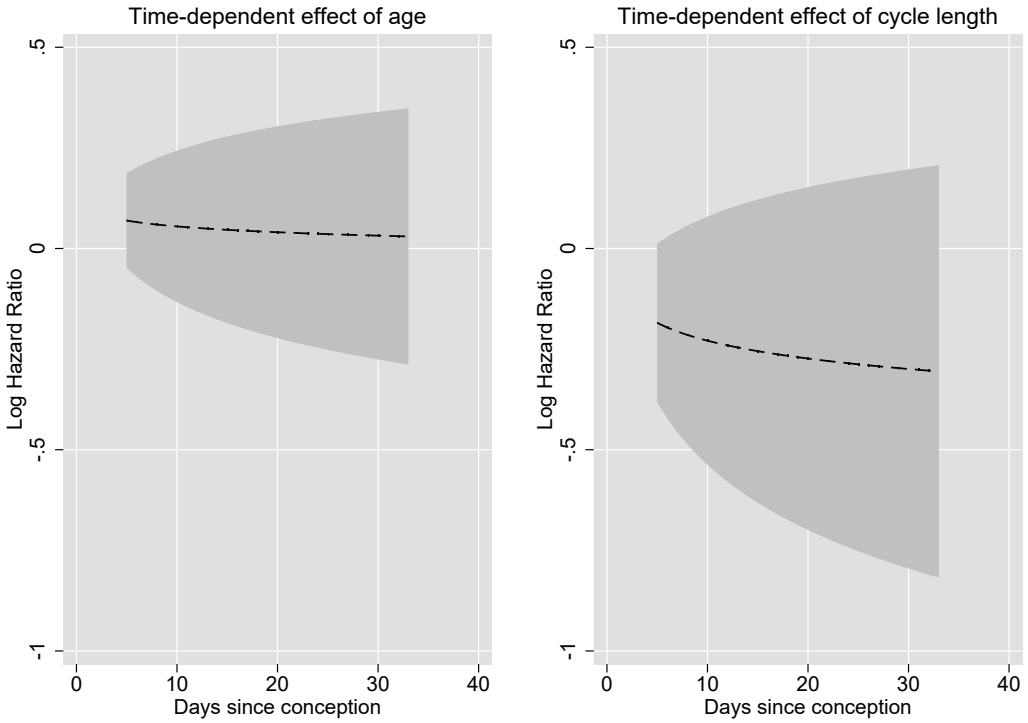


FIGURE 6.6. Time-dependent log hazard ratio plots for log hCG and age for the Weibull model

The PH assumption was checked by including a time-dependent effect for each covariate in the model and carrying out a likelihood ratio test against the main-effects model. There was no evidence of a time-dependent effect of age (p=0.7186) or cycle length (p=0.4827) for the Weibull model. A corresponding time-dependent FPM could not be fitted. Plots of the time-dependent log hazard ratios for age and cycle length from Weibull models are shown in Figure 6.6.

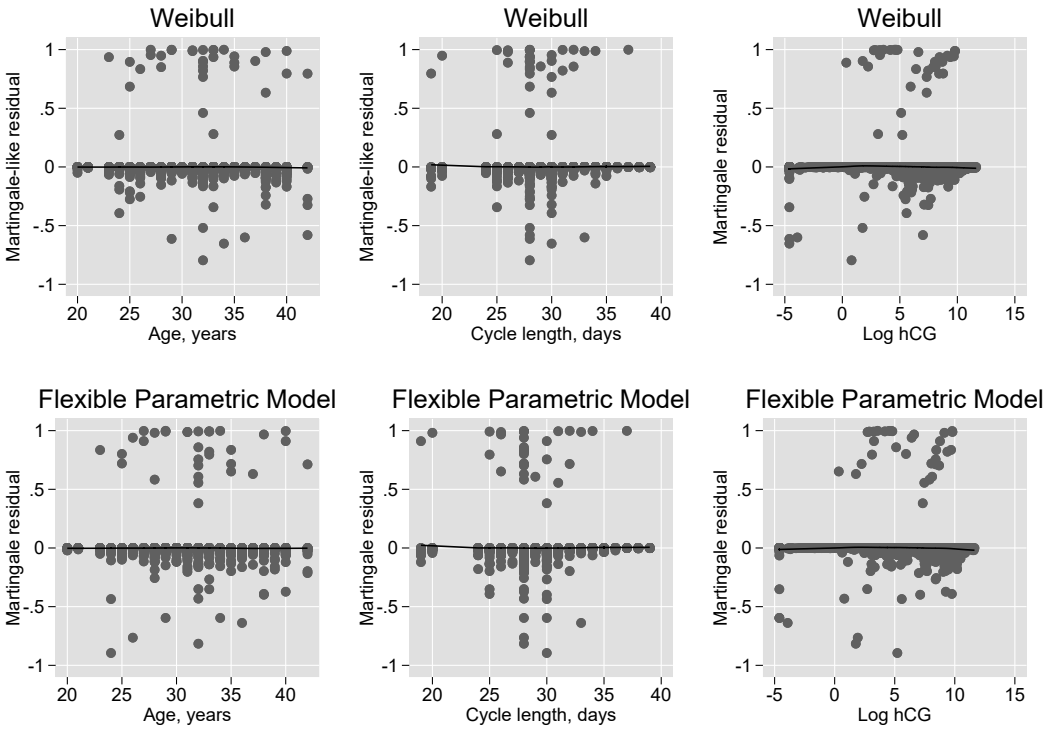


FIGURE 6.7. Observation level Martingale residuals from the full Weibull and flexible parametric models against age, cycle length and log hCG

Martingale residuals estimated from the fitted Weibull and flexible parametric models were plotted against age, cycle length and log hCG and are shown in Figure 6.7. Despite evidence of a decrease in AIC associated with the addition of a quadratic log hCG term shown in Table 6.5, this is not reflected in the Martingale residual plot. This is probably due to residuals being predicted for each observation for each individual. Martingale residuals predicted for each individual show curvature, suggesting a non-linear relationship for log hCG (see Figure 6.8).

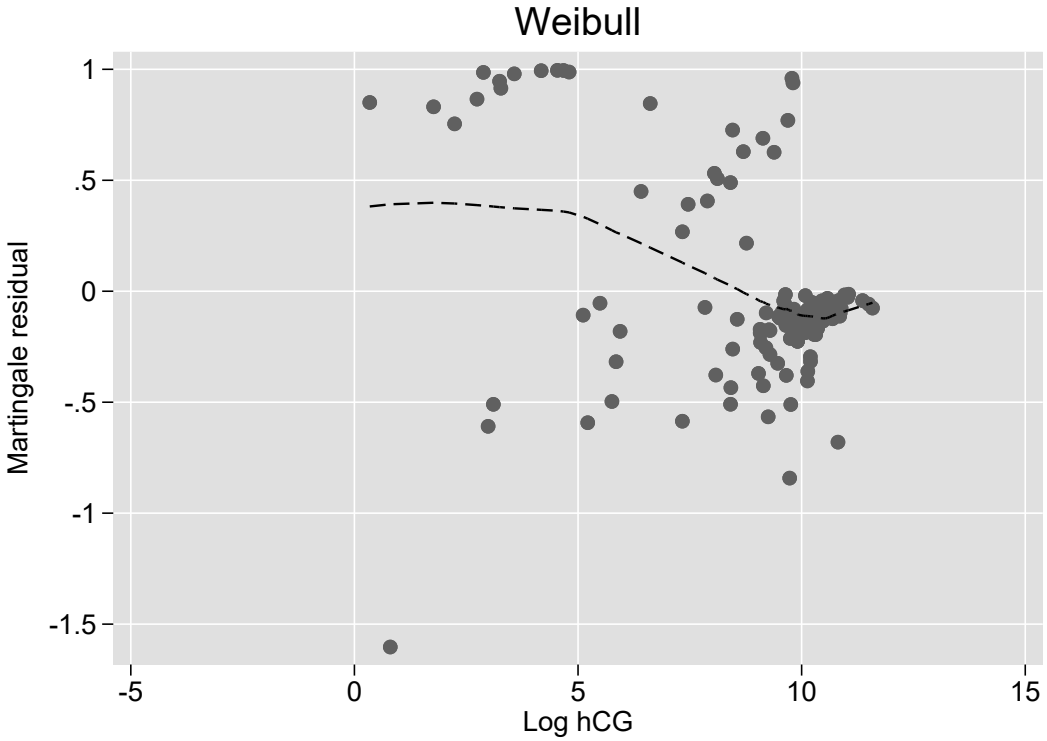


FIGURE 6.8. Individual level Martingale residuals from the full model plotted against log hCG

Deviance residuals were predicted and plotted against time as shown in Figure 6.9. All residuals above 1.96 were losses, ranging from 15 to 32 days post conception. Five women were under 30 years of age and one woman was 40 years of age. Cycle lengths for these women ranged from 20 to 37 days. Women who miscarried before day 16 had had log hCG values of less than 4, women who miscarried later reached log hCG values of close to 10.

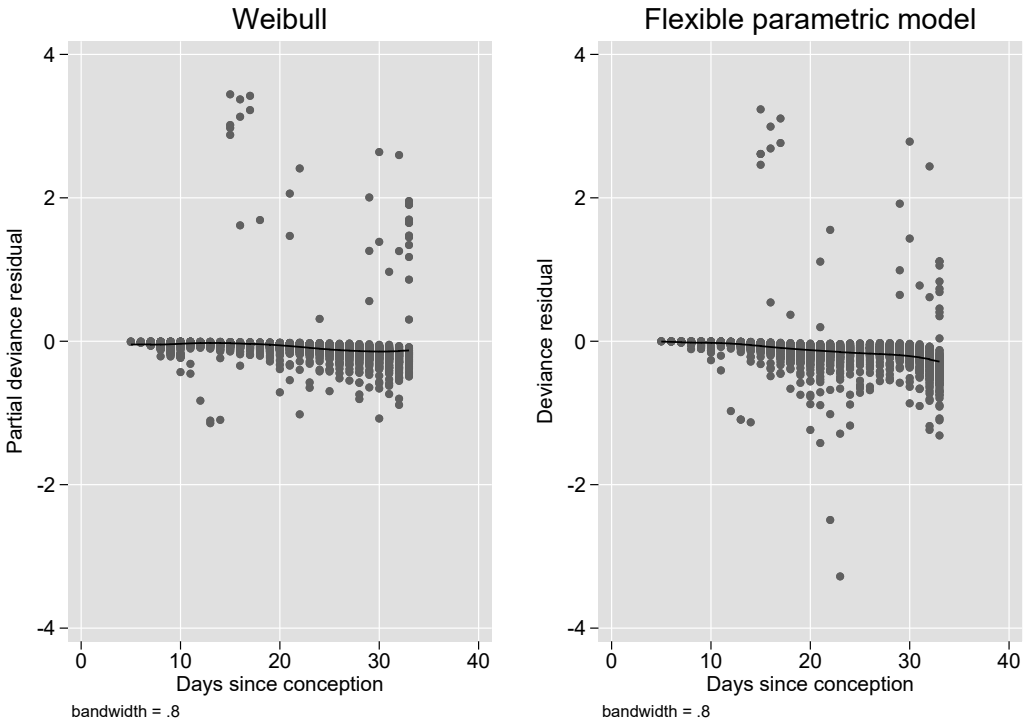


FIGURE 6.9. Deviance residuals from the full model plotted against time

As previously discussed in section 5.3.3 of Chapter 5, these estimates do not take into account measurement error for the hCG observations, resulting in overly precise estimates. In addition, the model is estimated on the assumption that observed hCG values do not change between measurements. We can improve estimates by using a two-stage model to begin to address the issue of measurement error.

6.4.4 Two-stage model

A two-stage model was fitted to the data (section 5.3.3), first modelling log hCG via an LMM and then incorporating the subject-specific predictions into both Weibull model and FPM.

6.4.4.1 Modelling the longitudinal profile

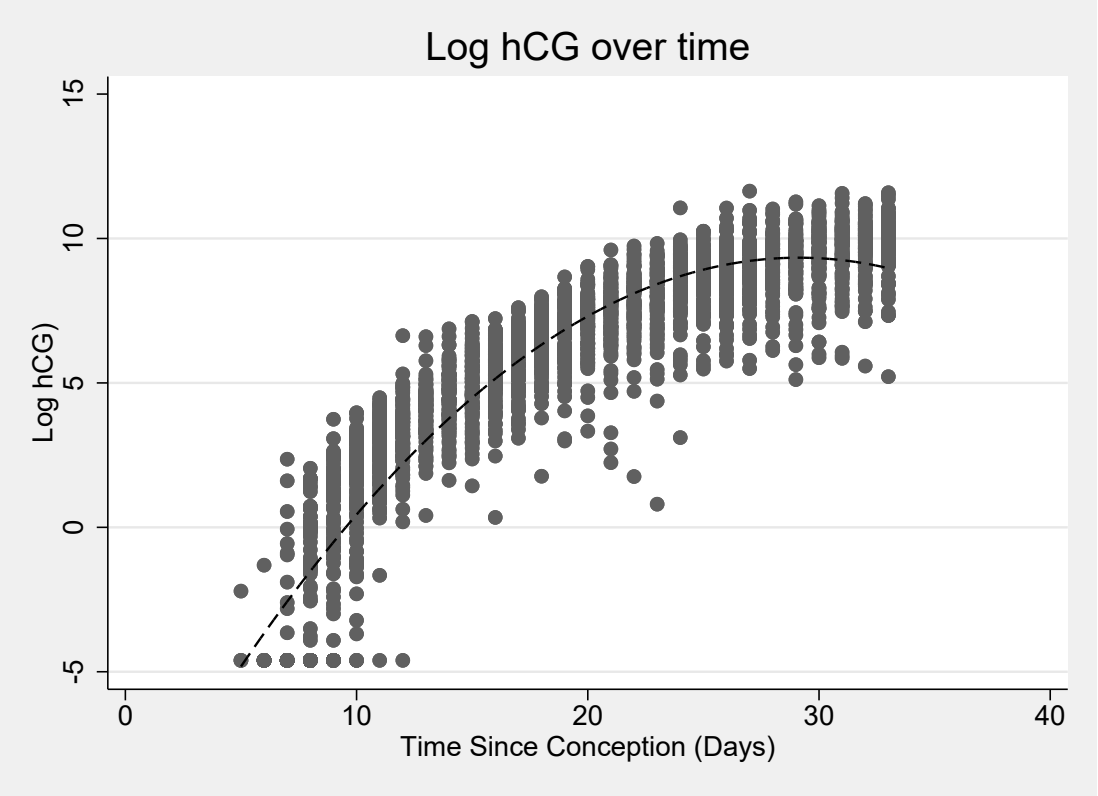


FIGURE 6.10. Log hCG over time with quadratic line of best fit

Log hCG was modelled using the LMM discussed in detail in section 3.5. A scatter plot of log hCG over time with fitted quadratic curve of best fit is given in Figure 6.10. A quadratic time term was included to more effectively model the change of log hCG over time, when modelling hCG as a response variable. The reasons for this are twofold. Biologically a complete hCG profile from the start to end of pregnancy is shaped like an inverted U, suggesting a quadratic may be a good fit. Assuming a quadratic curve does however, assume a dip in log hCG towards the end of follow-up, which may be unrealistic. This is especially so, as hCG for a normal pregnancy would not plateau and decrease until the end of the first trimester. Even so, compared with the linear slope the quadratic

demonstrated a reduction in AIC from 11022.39 for the linear slope to 9446.586 as shown in Table 6.7.

TABLE 6.7. Model fit estimates for fixed linear and quadratic slopes

Variable	Log likelihood	df	AIC	BIC
Linear slope	-5508.193	3	11022.39	11040.01
Quadratic slope	-4719.293	4	9446.586	9470.085
df: degrees of freedom				

Results from an initial fitted LMM, including a grouping variable for pregnancy outcome, confirmed that mean log hCG was -1.66 (95% CI -2.14, -1.18) lower in the biochemical pregnancy group and -1.13 (95% CI -1.48, -0.78) lower in the early miscarriage group, when compared with the healthy pregnancies. Results are presented in Table 6.8.

TABLE 6.8. Model estimates from a linear mixed effects model for log hCG mIu/ml

Variable	Mean	95% CI	
Time since conception, days	1.431	1.396,	1.466
Quadratic time since conception, days	-0.025	-0.026,	-0.024
Healthy	Reference		
Biochemical Loss	-1.656	-2.135,	-1.176
Early Loss	-1.132	-1.484,	-0.781
CI: Confidence Interval			

The hormone hCG is not detectable in the urine until around eight days after conception. This means that at the assumed day of conception ($t = 0$), hCG is zero and correspondingly log hCG is undefined. If modelling untransformed hCG then there would be no variation in hCG intercepts across woman and so a random intercept would be biologically implausible. However on the log scale the intercept can be retained, as a negative parameter estimate for the intercept would give a log hCG value asymptotically close to zero. Additionally retention of the intercept provides the flexibility for increasingly complex models to converge,

hence both fixed and random intercepts were retained for the longitudinal part of the model. The predicted fixed effect log hCG profiles from the fitted model retaining both intercepts is shown in Figure 6.11.

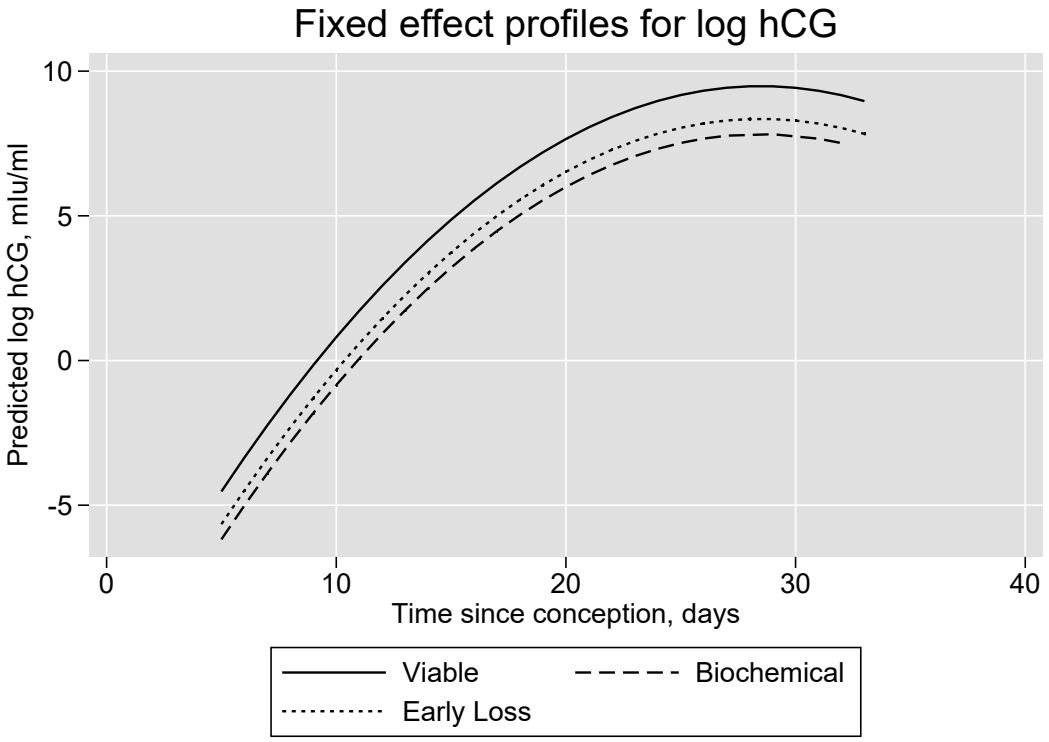


FIGURE 6.11. Predicted fixed log hCG profiles for model including both fixed and random intercept

The hCG slopes vary across women and so a random linear slope is important to include in the model. A fixed quadratic term for time was incorporated into the model, as well as a random quadratic slope. Table 6.9 shows the AIC estimates for each random effect addition. From this we see that the addition of the random quadratic slope reduces the AIC from 8334.067 to 8288.999, with likelihood ratio test giving $p<0.001$.

TABLE 6.9. Model fit estimates for the addition of random effects to the longitudinal submodel

Model	Log likelihood	df	AIC	BIC
Random intercept	-43144.233	5,	8638.466	8667.84
Random linear slope	-4160.033	7	8334.067	8375.19
Random quadratic slope	-4134.499	10	8288.999	8347.746
df: degrees of freedom				

TABLE 6.10. Model estimates from a linear mixed effects model for log hCG with random quadratic time term

Fixed-effect parameters	Hazard Ratio		95% CI
Time since conception, days	1.464	1.416,	1.512
Quadratic time since conception, days	-0.026	-0.027,	-0.025
Intercept	-11.687	-12.232,	-11.142
Random-effect parameters	Estimate		95% CI
$\sigma^2_{u_1}$	0.044	-0.028,	0.068
$\sigma^2_{u_2}$	0.000021	0.00001,	0.00004
$\sigma^2_{u_0}$	7.126	5.036,	10.084
$cov(u_1, u_2)$	-0.001	-0.001,	-0.0005
$cov(u_1, u_0)$	-0.524	-0.736,	-0.313
$cov(u_2, u_0)$	0.010	0.005,	0.015
σ^2_e	1.047	0.987,	1.111
$\sigma^2_{u_1}$, $\sigma^2_{u_2}$: slope variances			
$\sigma^2_{u_0}$: intercept variance			
cov(,): covariances between pairs of variances			
σ^2_e: measurement error variance			
CI: Confidence Interval			

Table 6.10 shows model estimates for the random quadratic slope model, which indicates a very small but statistically significant variance for the random quadratic slope term. The random quadratic effect was ultimately not included in the model to allow for comparability to the later joint model, which could not be fitted with a random quadratic time effect.

Final longitudinal model estimates are given in Table 6.11. Baseline mean log hCG was -11.612 (95%: -12.012, -11.21), corresponding to a hCG very close to zero. For a unit increase in time there was an overall 1.426 increase in mean log hCG. The variance terms also justify the inclusion of the random effects terms.

TABLE 6.11. Model estimates from a linear mixed effects model for log hCG

Fixed-effect parameters	Hazard Ratio		95% CI
Time since conception, days	1.451	1.418,	1.485
Quadratic time since conception, days	-0.025	-0.026,	-0.025
Intercept	-11.612	-12.012,	-11.213
Random-effect parameters	**Estimate**		**95% CI**
$\sigma^2_{u_1}$	0.005	0.004,	0.007
$\sigma^2_{u_0}$	2.619	1.949,	10.084
$cov(u_1, u_0)$	-0.095	-0.127,	-0.062
σ^2_e	1.113	1.051,	1.179

$\sigma^2_{u_1}$: slope variance
$\sigma^2_{u_0}$: intercept variance
$cov(u_1, u_0)$: covariance of u_1 and u_0
σ^2_e: measurement error variance
CI: Confidence Interval

Fitted values from the longitudinal model were incorporated into a Weibull survival model and FPM. Estimates are shown in Table 6.12. A one-unit increase in log hCG inferred a 65.1% reduction (HR: 0.349 95% CI: 0.277, 0.439) in the rate of pregnancy loss at time t for the Weibull model and a larger 69.0% (HR: 0.310 95% CI: 0.240, 0.399) reduction in the rate of pregnancy loss for the FPM. Measurement error for the log hCG observations has been accounted for as the fitted model creates a log hCG response at any time t. As noted in section 5.3.3, inputting predictions into the survival model assumes that values of log hCG do not change between observations, as we have predictions for each time point as opposed to a complete continuous trajectory of observations. Estimates are also still too precise as the uncertainty in estimates from the mixed model stage are

not carried through to the survival model stage, and may be characterised by bias and poor coverage [28].

TABLE 6.12. Survival model estimates from a two-stage model, modelling time to miscarriage with the inclusion of log hCG predictions

Weibull Model	Hazard Ratio	95% CI
Age, years	1.075	1.010, 1.145
Usual cycle length, days	0.862	0.762, 0.976
Predicted log hCG	0.349	0.277, 0.439
Flexible Parametric Model*		
Age, years	1.067	0.999, 1.140
Usual cycle length, days	0.869	0.762, 0.990
Predicted log hCG	0.310	0.240, 0.399
*2 boundary and 2 internal knots		
CI: Confidence Interval		

6.4.5 Fitting a joint longitudinal-survival model

In order to now account for both measurement error and the intermittently measured hCG measurements a joint longitudinal-survival model was fitted to the dataset (see section 5.4.2). To link the expected value of the log hCG to the event time, a trajectory of log hCG measurements was built via the longitudinal part of the joint model. Figure 6.12 reiterates that steeper trajectories of log hCG are consistent with viables pregnancies.

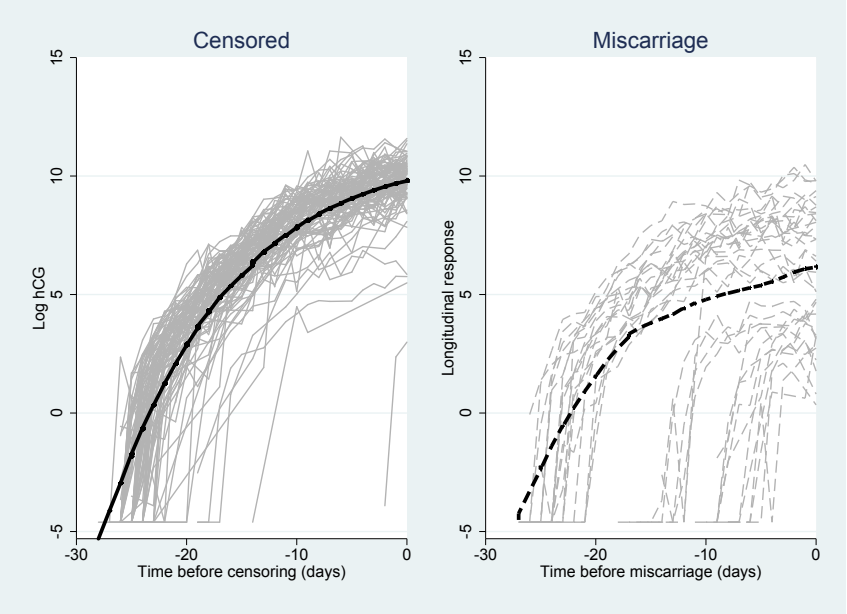


FIGURE 6.12. Log hCG trajectories prior to censoring or miscarriage

The longitudinal submodel mirrored the model built in section 6.4.4.1 and in this instance contained no further covariates, which is consistent with the Marriott *et al.* [44] analysis. Joint model estimates including random intercept and random linear slope term, and Weibull baseline hazard are given in Table 6.13.

A strong association between log hCG and miscarriage was observed, through linking the expected log hCG response to the event time. A one-unit increase in expected current value of log hCG resulted in a 66.1% decrease (HR: 0.339 95% CI: 0.257, 0.447) in the rate of miscarriage. Maternal age at conception was no longer significantly associated with miscarriage, although usual cycle length remained statistically significant. A one-day increase in a woman's usual cycle length represented a 15.6% decrease (HR: 0.844 95% CI: 0.739, 0.965) in the rate of miscarriage. It was not possible to fit a joint model incorporating an FPM,

or a quadratic random effect for the slope in the longitudinal submodel. This is most likely due to the limitations of the sample size.

TABLE 6.13. Model estimates from a joint longitudinal-survival model with current value association structure

Survival submodel	Hazard Ratio		95% CI
Age, years	1.076	0.998,	1.159
Usual cycle length, days	0.844	0.739,	0.965
Expected current value of log hCG	0.339	0.257,	0.447
Scale and shape parameters	Estimate		95% CI
log λ	-26.716	-33.299,	-20.132
log γ	2.371	2.178,	2.563
Longitudinal submodel	Mean		95% CI
Time	1.431	1.396,	1.466
Quadratic time	-0.025	-0.026,	-0.024
Intercept	-10.171	-10.540,	-9.801
Random-effect parameters	Estimate		95% CI
σ_{u_1}	0.072	0.061,	0.084
σ_{u_0}	1.538	1.327,	1.782
$\rho(u_1, u_0)$	-0.781	-0.851,	-0.683
σ_e	1.054	1.024,	1.084

σ_{u_1}, σ_{u_0}: slope and intercept SD
$\rho(u_1, u_0)$: correlation between u_1 and u_0
σ_e: measurement error SD
CI: Confidence Interval

6.4.6 Comparing estimates across models

The association between log hCG and time-to-miscarriage was evident across all models. Association parameter estimates for each of the methods are shown in Table 6.14. To allow for comparison, the two-stage and joint model approaches were both fitted with longitudinal models that included a random linear slope only. The largest association was estimated by the joint model, with the two-stage model following closely behind. The survival model with time-varying covariate gave the smallest association. Standard errors were much smaller at 0.034 and 0.041 for the time-varying covariate and two-stage models respectively, compared

to 0.142 for the joint model. This emphasises that joint models when appropriately utilised can better gauge uncertainty than their simpler counterparts [28].

TABLE 6.14. Model estimates for log hCG from each applied method of analysis

Model	Hazard Ratio	95% CI	Standard Error
Time-varying covariate	0.419	0.357, 0.491	0.034
Two-stage model	0.349	0.277, 0.439	0.041
Joint Model	0.339	0.257, 0.447	0.142
CI: Confidence Interval			

6.4.7 Alternative survival submodels and association structures

Restricted cubic splines with three internal knots (see Figure 6.5) more aptly described the baseline hazard function for the survival submodel, compared to the Weibull distribution. However, despite altering the number of integration points and quadrature nodes, as well as using a variety of initial values, it was not possible for model convergence to be achieved using RCS functions to describe the baseline hazard. Subsequent joint models were fitted using a Weibull baseline hazard. The exploratory plots suggest that the rate of change of log hCG may be an important factor affecting progression in early pregnancy. This was investigated through inclusion of both the current and slope association structures in the model.

Results in Table 6.15 show that together the current value and slope of the log hCG trajectory are strongly associated with early pregnancy loss.

TABLE 6.15. Model estimates from a joint longitudinal-survival model with current and first derivative association structures

Survival submodel	Log Hazard Ratio		95% CI
Age, years	0.084	0.005,	0.163
Usual cycle length, days	-0.187	-0.328,	-0.045
Expected current value of log hCG	-1.574	-2.036,	-1.113
Expected rate of change of log hCG	9.377	2.810,	15.944
log λ	-66.874	-96.884,	-36.864
log γ	3.173	2.758,	3.587
Longitudinal submodel	**Mean change**		**95% CI**
Time	1.431	1.396,	1.466
Quadratic time	-0.025	-0.026,	-0.024
Intercept	-10.179	-10.548,	-9.809
Random effect paramters	**Estimate**		**95% CI**
σ_{u_1}	0.072	0.061,	0.085
σ_{u_0}	1.542	1.332,	1.787
$\rho(u_1, u_0)$	-0.790	-0.857,	-0.696
σ_e	1.054	1.024,	1.084

σ_{u_1}: slope SD
σ_{u_0}: intercept SD
$\rho(u_1, u_0)$: correlation between u_1 and u_0
σ_e: measurement error SD
CI: Confidence Interval

For women who had the same value of expected log hCG the log hazard ratio for a unit increase in the slope of the log hCG trajectory was 9.923 (95% CI:3.264, 16.582). This suggests that an increase in log hCG has a detrimental effect on pregnancy, compared to a woman who experiences a decrease in their log hCG slope, provided that their current log hCG value is the same. This was paired with a consistently positive association between a unit increase in the expected current value of log hCG at time t and the rate of miscarriage at the same time. A unit increase in expected log hCG at time t inferred a corresponding 78.9% decrease (HR: 0.211 95% CI: 0.331, 0.724) in the rate of miscarriage at time t. This effect was larger than was seen in the current value joint model. The results for the slope association do not reflect what we would expect of hCG

trajectories and so requires further investigation. This may be a symptom of using a quadratic to model log hCG which dips driving down the trajectories for healthy pregnancies. Inclusion of a random quadratic slope would have allowed greater variation, preventing this.

6.4.8 Subject-specific conditional survival predictions

Subject-specific predictions from the current value model are illustrated below. Figure 6.13 presents longitudinal response data as well as conditional survival probabilities for two participants who experienced biochemical losses.

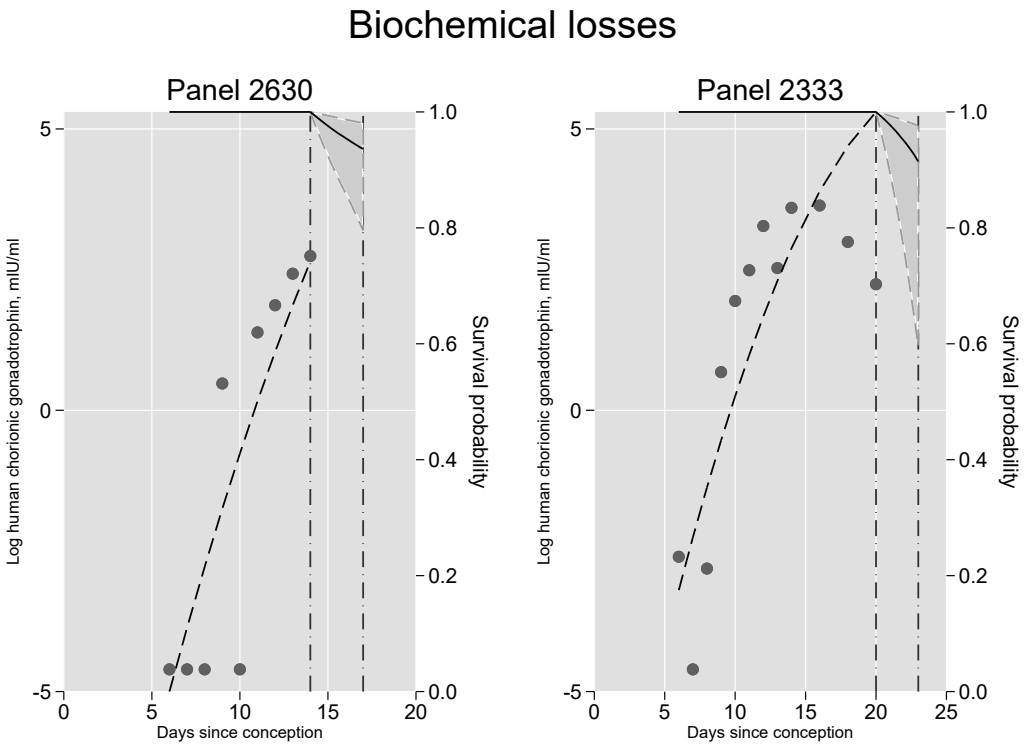


FIGURE 6.13. Conditional survival probability curves for participants who experienced biochemical losses

For panel 2630, five log hCG measurements were observed. Prediction of pregnancy survival at day 17, two days after the final longitudinal observation, gave a pregnancy survival probability of approximately 95%. This is because the observed data did not indicate a decrease in log hCG. Panel 2333 included 12 log hCG measurements and indicated a pregnancy survival of slightly less than 95% at day 23, despite the decrease in log hCG. The uncertainty was also greater for the second panel with wider confidence intervals.

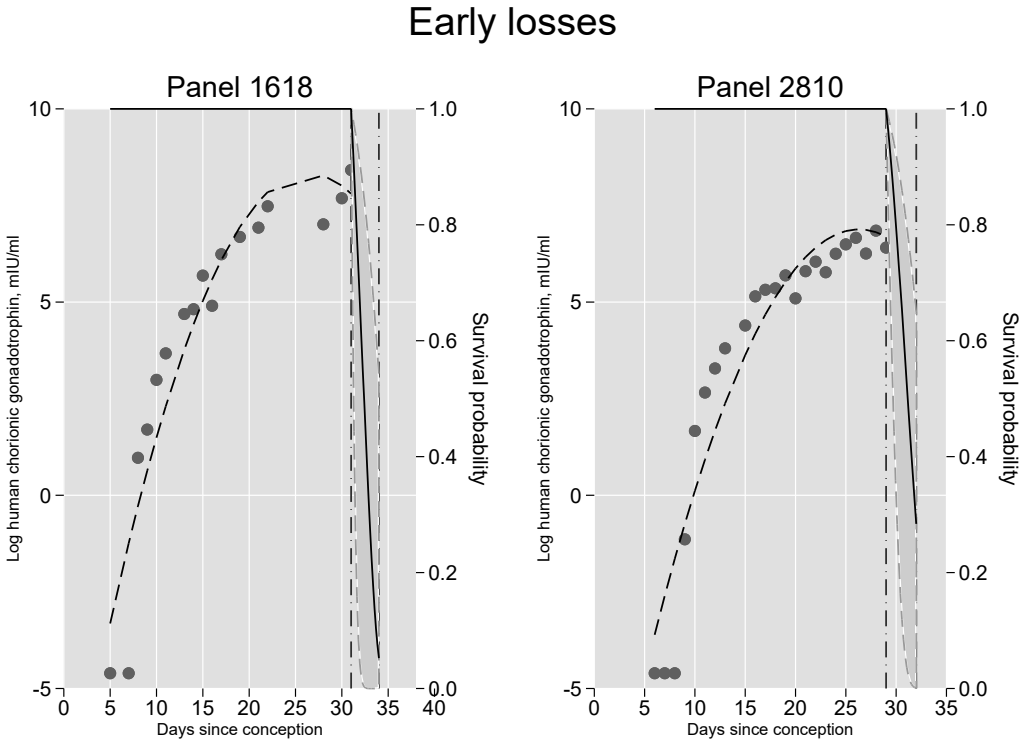


FIGURE 6.14. Conditional survival probability curves for participants who experienced biochemical losses

Subject-specific conditional survival probabilities for two early losses are shown in Figure 6.14. Panel 1618 included seventeen log hCG observations. The trajectory began to plateau and drop from day 22 onwards. The corresponding two-day conditional survival prediction at day 34 was approximately 5%, confirming a loss.

Panel 2810 contributed 23 biomarker observations. The longitudinal profile did not increase as steeply as panel 1618, but indicated a more consistent rise. The conditional survival probability at day 32 was just over 30%.

Figure 6.15 shows conditional survival probabilities for two healthy pregnancies. Panel 1186 consisted of 21 log hCG observations. The longitudinal trajectory showed variability in hCG over follow-up time, rather than consistent increases. The resulting two-day subject-specific survival prediction at day 35 was close to 70%. Panel 1303, on the other hand, showed a more consistent and steeper rise in log hCG over 18 observations. Observed values were clustered close together, resulting in a predicted pregnancy survival of over 95%.

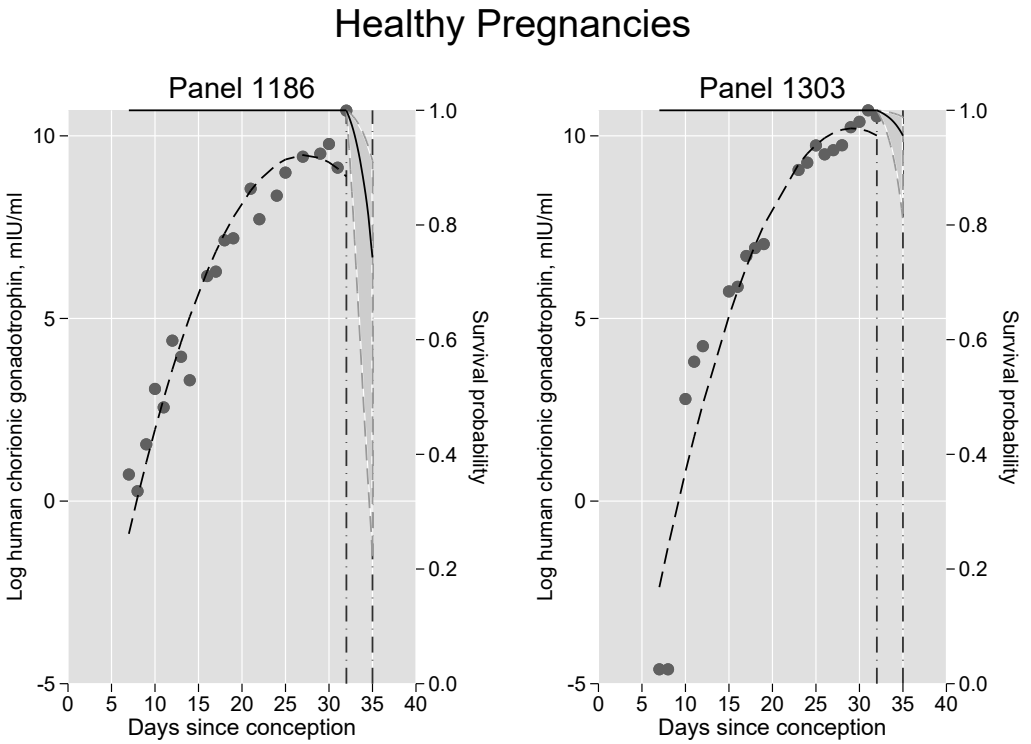


FIGURE 6.15. Conditional survival probability curves for participants who experienced biochemical losses

6.5 Discussion

In this analysis the association between longitudinal log hCG and time-to-miscarriage was appropriately modelled utilising the more advanced joint longitudinal-survival model framework. This allowed the association between the two outcomes to be modelled whilst accounting for measurement error and the intermittent nature of observations. A unit increase in the current value of log hCG was associated with a 66.1% decrease in the rate of miscarriage. This is a slighter larger effect than observed in the two-stage model estimates (HR: 0.349), but much greater than the 4% decrease in rate observed in the two-stage model analysis conducted by Marriott *et al.* [44]. The latter model assumed day zero was the estimated day of the missed period, thereby excluding hCG observations prior to this day. At this point hCG would likely be above 25 mIU/ml, which is the threshold for standard pregnancy tests. The model fitted in this chapter included lower values of hCG from implantation (based on day zero at conception), and subsequently resulted in a much larger association between log hCG and time-to-miscarriage. This perhaps indicates that absolute increases in hCG around implantation are important in determining the viability of a pregnancy. Another key difference between models was the use of the Weibull PH model here and the Cox PH model by Marriott *et al.*. Given sufficient sample size, a fully parametric modelling approach can yield more efficient estimates than the Cox model. However, assuming a Weibull baseline hazard function opens up the possibility of misspecification, which the Cox model avoids [246]. It is important to note, that association estimates for log hCG and time-to-miscarriage across the two-stage and joint models fitted in this chapter were consistent.

The survival submodel included cycle length and age in the model. Advanced maternal age at conception is a known risk factor associated with miscarriage [247; 248]. Furthermore evidence suggests that a shorter cycle length may represent a surrogate marker for advanced reproductive age, particularly as the initial transition to peri-menopause can be characterised by a shortening of cycle length [249; 250]. Further known risk factors are number of previous miscarriages as well as lifestyle related aspects such as an increased alcohol consumption, smoking and maternal obesity to name a few, however not all of these were recorded as part of this study [251]. In this sample of women, whether a woman had experienced a previous miscarriage was not significantly associated with pregnancy loss. This may be because the variable included was binary and did not allow quantification of the number of previous miscarriages. Unlike the Marriott *et al.* [44] study, this analysis did not include the time taken to reach a hCG value of 25 mIU/ml from LH surge in the model. Though this was not a statistically significant addition at the survival modelling stage for this analysis, Marriott *et al.* [44] reported a 30% increase in the rate of miscarriage for every additional day it took for hCG to reach 25 mIU/ml from the day of the LH surge. From a clinical perspective this may have been an important omission as there is evidence to suggest that slow rising hCG is indicative of miscarriage or ectopic pregnancy [252; 253]. For this data the day of surge was measured, however in natural pregnancy settings, perceived slow rises in hCG may be an artefact of a misdated pregnancy [254].

The rate of miscarriage in this dataset was 34.1%, which is significantly greater than the 14-24% recorded in the literature [5]. This is a consequence of randomly sampling double the number of healthy pregnancies compared with miscarriages, and may have introduced sampling bias, impacting subsequent model estimates.

Unique to this analysis were the prediction of subject-specific survival probabilities. Prediction window intervals were kept narrow to minimise uncertainty around survival probability estimates. For individuals who contributed consistent and greater number of hCG observations, there was less uncertainty around risk predictions and they were more likely to match the recorded outcome. More often these women experienced healthy pregnancies and so by design contributed more measurements. Predictions for biochemical pregnancies were the poorest, because observations were fewer and trajectories were not characterised by a decrease in hCG as they were with the early losses.

6.5.1 Current research

Here, the potential of prospectively monitoring pregnancy from first detection of urinary hCG was investigated. It is more likely that tracking hCG at this early stage would present as an adjunct to diagnosis by ultrasound later on in the pregnancy. This echoes research suggesting declines in hCG can be noted even prior to other symptoms presenting [255]. There is also potential for this monitoring to occur prior to conception, with a recent study finding that a lag between the luteal phase and hCG production can be indicative of a biochemical pregnancy, possibly due to early or delayed implantation [15]. Tracking of hCG by pregnant women is also practicable, as demonstrated by Foo *et al.* [15] who employed a fertility monitor that also provided semi-quantitative analysis of hCG levels on pregnancy tests that were used daily in women who conceived. Retrospective analysis of the semi-quantitative data indicated that non-viable pregnancies had different hCG profiles to viable pregnancies.

Serial tracking has the potential to cause stress, although women using tests to track ovulation for fertility purposes do not appear to have higher stress levels than those not employing tests [256; 257]. Even so, it is likely that tracking would

initially be of benefit in high risk pregnancies, where anxiety levels are already high and there would be a willingness and reason to track.

Monitoring from first detection has the potential to be useful in cases of recurrent miscarriage, particularly as research into treatment gains traction. A recently published feasibility study assessing the effectiveness of the diabetes drug sitagliptin as a treatment for recurrent miscarriage, presented promising findings [23]. This trial builds on previous research, which found that in some cases of recurrent miscarriage, it is the deterioration of stem-like cells in the uterus, which contribute to pregnancy loss. When adjusted for age and baseline colony forming unit (CFU) counts, the CFU count was higher (RR: 1.52, 95% CI: 1.32, 1.75) in the sitagliptin group compared to placebo, pointing to successful regeneration of cells. These findings could revolutionise treatment for unexplained recurrent miscarriage, with tracking serving as a complement.

Not all miscarriage is likely to be predictable due to the diverse aetiology of the condition. Some causes can be directly related to reduced hCG levels, e.g. conditions that affect rate of embryonic development such as chromosomal abnormalities, or an inadequate placenta. Other causes, for example, where infectious agents or trauma are involved, may have no forewarning.

6.5.2 Strengths and limitations

The joint model improved upon the two-stage approach carried out by Marriott *et al* [44]. Fitting the current and slope association structures exemplified the strength of the association between the hCG trajectory slope and time-to-miscarriage, which is a sensible result. This also represented the first application of these advanced modelling techniques to early pregnancy outcomes , which utilised a frequentist framework. Recent papers have both utilised Bayesian approaches to classify each type of pregnancy based on longitudinal profiles [45; 46].

The final joint model could not be fitted using the more flexible restricted cubic splines, falling back instead on the Weibull baseline hazard. Model selection was carried out using forwards stepwise selection, which is known to introduce bias [258]. Alternative selection methods should be considered in future, specific to the joint modelling context. Selection based on the log likelihood contribution for the longitudinal part and conditional survival model have been proposed, but are currently only implemented in the SAS statistical software[223]. The example dataset was relatively small, and so fitting a model as complex as the joint model was challenging. Results must therefore be interpreted with caution.

Sacrifices were made when modelling the baseline hazard for the survival submodel, as well as for the choice of random components in the longitudinal submodel, in order to achieve convergence. Despite changes to integration points, quadrature nodes and initial values, an arguably better fitting model could not be obtained. The choice to model log hCG quadratically resulted in healthy trajectories being driven down at the boundaries of the data, most likely explaining why the slope association structure produced conflicting results to the current value model. Using RCS functions to model the trajectory function may be a solution, which will be explored further in Chapter 7. Overall, the complexity of the joint model and the limited observed data prevented exploration of more advanced association structures. This speaks to how advanced these models really are and the subsequent computational complexity. Of interest is the possible non-linear association between log hCG and time-to-miscarriage which was ultimately modelled linearly. If indeed the association between the two is quadratic then misspecification may produce biased estimates. This will be investigated in a simulation study in Chapter 9.

As previously noted the data had a higher prevalence of women who miscarried due to the way samples were chosen for testing. Another source of selection bias may come from insights into the type of women who would have been interested in taking part in a study with such an intense collection protocol. Though any underlying medical information affecting fertility was not recorded, it is not inconceivable that women recruited to the study may have been experiencing difficulty in conceiving. Furthermore, subfertility has also been shown to be associated with a greater incidence of miscarriage [259]. Certainly, miscarriage is more prevalent amongst women with polycystic ovary syndrome (PCOS), with 30% to 50% of PCOS sufferers experiencing early pregnancy loss [260].

As this was a retrospective analysis of data with limited follow-up measurements, it was not possible to update predictions as measurements were observed. Uncertainty was therefore greater for predictions for wider time periods. Predictions were also based on current value models, ignoring changes in hCG slopes, which may be important in modelling the association between hCG and miscarriage. If models are incorrectly calibrated then resulting predictions will be affected. For prediction windows which are chosen arbitrarily and without continued follow-up, it is natural that increasingly lower survival probabilities will be predicted as the time between the last observation and prediction time elapses further. Though feasible to produce, these predictions require further scrutiny to be meaningful in practice.

Attempting to utilise data for diagnostic or monitoring purposes also requires careful consideration of the potential for false positives. This analysis did not take into account the sensitivity and specificity of the fitted model, however this is an important component of the analysis in Chapter 7, in line with developments in

joint model methodology [43; 261]. With the small sample size, it was also not viable to split the dataset for development and validation of the model.

A large enough sample size is required to achieve convergence when fitting a joint model. However employing two separate modelling techniques for longitudinal and survival data requires even larger sample sizes. The increased efficiency of simultaneously modelling the two outcomes has the advantage of maintaining desired power at a lower sample size [29]. This makes designing clinical trials around a joint model framework an attractive prospect.

6.5.3 Conclusions

The analysis of this dataset required careful consideration of the biological aspects of timing and level of hCG production, and how this related to the statistical aspect of modelling. This resulted in a number of compromises being made whilst modelling to allow the software to fit these models. Despite challenges, this analysis demonstrated that an association exists between hCG and early pregnancy outcome, and there is a need to model these outcomes simultaneously.

The novel extension to this analysis concerns the output of subject-specific pregnancy survival predictions. Without prospective follow-up data, updating predictions was not possible. Though the effectiveness of possible treatments, particularly for recurrent miscarriage, remain uncertain, the joint model is well placed for dynamic monitoring. However long term follow-up observations are required, along with access to a larger dataset for a model to be developed and subsequently validated. Further analyses should also consider the sensitivity and specificity of the fitted predictive model, to minimise the likelihood of false diagnoses of miscarriage. Issues of model calibration and discrimination as it pertains to joint models (see section 5.12) will be considered in the analysis of a larger dataset of longitudinal hCG and pregnancy outcomes presented in Chapter 7.

Chapter 7

The General Cycle Collection study

7.1 Chapter overview

This chapter presents a detailed analysis of the general cycle collection study data. The analysis builds on the previous application of the joint longitudinal-survival model to the early pregnancy setting in Chapter 6, with the availability of a larger dataset and extended follow-up. A carefully considered prediction model was built, for which individualised predictions were obtained. A key extension was the assessment of the discriminative capabilities of the fitted joint models, as well as the calibration. In conducting this analysis the aim was to judge the feasibility of fitting such models in practice against the constraints of estimation.

7.2 Introduction

The analysis performed in Chapter 6 demonstrated that a joint longitudinal-survival model more appropriately models the association between longitudinal hCG and time-to-miscarriage. The inclusion of the biomarker in the survival model as a continuous trajectory function addresses measurement error, the intermittent nature of observations, uncertainty and links the two outcomes through a shared parameter framework. The resulting association indicated that a unit increase in log hCG corresponded to a 66.1% decrease in the rate of miscarriage. These results however should be interpreted cautiously, as several barriers were

encountered in the modelling process, concerning size of the available data and its distribution of viable and miscarried pregnancies. The limitations of the data introduced in Chapter 6, and the computational complexity of fitting joint models impacted the ability to build a well-calibrated model. In this chapter, the joint model building for longitudinal hCG and time-to-miscarriage will be approached again with a larger dataset, including additional maternal demographic data and extended follow-up. Computational hurdles encountered previously will be overcome by utilising alternative software, which has since been developed (see section 5.9).

7.3 Data

The general cycle collection study (clinical trial number: NCT01577147) was carried out by SPD Development Company Ltd with the view to collect and maintain a bank of urine samples for product development and validation. The study followed fertile, non-pregnant women from the beginning of their cycles, to establish levels of key hormones; FSH3, LH, P3G and hCG, for women who conceived (see section 2.2). The probability of conception within one cycle, however, is estimated to be between 30 to 40% with an approximate cumulative 90% probability of conception only achieved after twelve complete cycles [262]. To evaluate pregnancy-related products women were followed up over multiple cycles, with intercourse timed to maximise chance of conception and reduce the length and expense of collection.

Recruitment began in January 2012, with the final observations collected in September 2017. Volunteers were recruited through online advertising, targeting women who were actively attempting to conceive. Women included in the study were aged between 18 to 45 years, experienced menstrual bleeding and could

provide written, informed consent. Those women diagnosed with a condition recognised as a significant barrier to successful conception were excluded. Women were excluded if they were currently undergoing fertility treatment, as fertility drugs often trigger abnormally elevated levels of the very same hormones intended to be studied. Treatments can involve drugs which stimulate the production of reproductive hormones known collectively as gonadotrophins, of which hCG is one example [263]. As a result, any hCG observations from these women would not follow the typical path of a natural pregnancy.

Participation in the study was incentivised through provision of Clearblue ovulation prediction products. Natural family planning (NFP) methods rely on recording basal body temperatures and/or monitoring cervical secretions. A very small increase in temperature is usually observed in the three days post ovulation. However, this temperature postdates optimal fertility, which would hinder conception. Changes in the quantity and viscosity of cervical mucus can help to pinpoint the days leading up to ovulation. Yet, these NFP methods all require consistent observation over several cycles to succeed in identifying personalised fertile windows. Similarly many cycle apps utilise the calendar method to pinpoint the fertile window. This is usually unilaterally based on a fixed average 28 day cycle, and even when allowances are made for an alternative cycle length, the maximal probability of calendar apps correctly predicting the day of ovulation is still only 21% [264]. The prediction products offered to volunteers are more sophisticated in that they measure hormones in early morning urine to identify the optimal window for intercourse, specifically LH and oestrogen.

Women remained in the study for up to three cycles. It was originally estimated that 600 women would need to be recruited to achieve 100 pregnancies. This was a conservative estimate as evidence has suggested that by the end of

three cycles, the cumulative probability of conception is 68% [262], though this did not consider fertility aids. Recruitment was expanded to recruit 4000 volunteers in total. The final number of women recruited was 4025.

Volunteers collected early morning urine samples from the first day of their cycle (first day of bleed) to up to seven days of the next cycle if they did not become pregnant, or up until day 60 if they did become pregnant. Continued collection for women who self-reported that they did not conceive allowed the capture of possible biochemical pregnancies. The day of LH surge, an indicator of ovulation, was identified via sample testing. This was utilised to calculate the estimated date of conception, assumed to be the day after the LH surge [265].

In addition to observing daily urinary hCG measurements, FSH3 and P3G were also measured. LH and FSH3 are intertwined in that they both play a role in triggering ovulation[266]. P3G encourages successful implantation of the fertilised ovum into the endometrial wall and low levels of this have been linked to recurrent miscarriage [21]. The roles of hormones during the cycle are discussed in more detail in Chapter 2.

7.4 Methods

Joint longitudinal-survival models were fitted to longitudinal log hCG and time-to-miscarriage. The hCG measurements were modelled on the natural log scale due to the skewed nature of the data. As hCG rises exponentially, raw observations can become large very quickly. A natural log transformation allows the observations to be scaled for better comparison, and brings the data closer to being normally distributed.

Previously all hCG observations were included in the linear mixed effects model. For this analysis observations which fell below the 2 mlU/ml detection limit of the assay were excluded.

Modelling approaches were discussed with colleagues at SPD Development Company Ltd. This included biological and statistical input from experts who have been working in the early pregnancy field for a number of years. The longitudinal trajectory was built using a combination of model selection procedures and consideration of biological plausibility. When deciding on the general shape of the profiles, the addition of non-linear terms and the inclusion of random effect terms, the Akaike's Information Criterion (AIC) was utilised as a means for aiding model selection. Model covariates were chosen based on evidence of an association between the covariate and hCG or miscarriage.

7.4.1 Time scale and time zero

To allow for a common timeline amongst women, it was necessary to anchor day zero to a particular day in the cycle that would be meaningful to the analysis. As with the analysis performed in Chapter 6 miscarriage was modelled from the estimated date of conception. This was based on the assumption that ovulation takes place approximately 24 hours after the LH surge [265]. The lag between conception and implantation when hCG begins to be observed and pregnancy can be confirmed, resulted in delayed entry.

7.4.2 Modelling longitudinal hCG

In Chapter 6 log hCG was modelled non-linearly using a quadratic time term. The quadratic failed to capture the hCG plateau, instead causing a dip much earlier than would usually be observed. To address this, restricted cubic splines were fitted to model the log hCG trajectory. Varying numbers of splines and knot

positions were considered, based on the centiles of the time variable. Random effects were chosen based on likelihood ratio tests for each addition to the model following the approach discussed in section 3.5.6. A random intercept and random linear slope were included in the final model allowing for variation in individual baseline log hCG and slope. Inclusion of additional non-linear random effect terms were also considered.

The longitudinal model developed in Chapter 6 did not include any additional variables. With input from collaborators at SPD, maternal body mass index (BMI) and average cycle length were included. Higher BMI values correspond to lower hCG values. A study investigating reference ranges and the determinants of total serum hCG values during pregnancy found that women with a BMI of between 34 to 46 kg/m^2 exhibited on average 9369 IU/L lower (SD: 729) lower hCG than women in the lowest BMI range of 15 to 25 kg/m^2 [267]. The length of the cycle determines when the LH surge occurs and subsequently when conception and implantation take place[10]. Ethnicity was considered for inclusion, but was excluded due to the lack of ethnic variation in the data. BMI and cycle length were centred at 25 kg/m^2 and 28 days respectively to allow for meaningful interpretation.

7.4.3 Survival modelling for time-to-miscarriage

A Weibull distribution was utilised to model the baseline hazard, though restricted cubic splines provided a better fit. This was owing to the restrictions presented by the software in Stata, which did not allow estimation of individualised survival predictions for models including spline terms. For similar reasoning the survival submodel again included a Weibull baseline hazard. Delayed entry models were fitted. As with the longitudinal trajectory model, BMI was

included, as a known confounder, along with smoking status, maternal age, number of previous miscarriages, and P3G and FSH3 hormone levels on the day of implantation. Smoking status was formed of two categories; never smokers and a combined current and previous smokers category, due to the low numbers in the current smokers category. This is unsurprising in a trying to conceive population, who are advised of cessation beforehand. Smoking has been shown to be associated with miscarriage, with passive smoke inhalation also a risk factor for miscarriage [72]. Age is an established risk factor for miscarriage with women aged 25 to 29 conferring the lowest risk and women aged 45 and over the highest [68]. Following inspection of the age distribution across participants a quadratic age term was included in the model. Women who have experienced a miscarriage previously are also at higher risk of incurring another loss than women who have not previously experienced a loss [67]. BMI was centred at 25 kg/m^2 and age was centred at 30 years.

Higher levels of P3G tend to signal a positive pregnancy outcome, as it is vital to establish and maintain the pregnancy. Looking at P3G on the day of implantation may allow evaluation of the 'health' of the corpus luteum [268]. Elevated levels of FSH3 have been linked to a shortening of cycle length and is thought to act as a surrogate for reproductive ageing [269].

Non-linear effects of continuous variables were considered and interactions between covariates were tested for inclusion in the model. The proportional hazards assumption was assessed by plotting log analysis time against the negative log of the negative logarithm of the survival probability for categorical variables. For continuous variables a time-dependent effect for the covariate was included in the model and tested for significance using a likelihood ratio test. Deviance residuals were estimated to identify outlying observations. Sensitivity analyses were

conducted by removing observations and refitting models to assess the effect on model estimates.

Entry into the study was dependent on pregnancy confirmation by detection of hCG greater than 2 mlU/ml. Women who did not have hCG observations above this threshold were excluded from the analysis. For women who experienced an event or were censored on the same day as entry, a day was added to their survival time to allow for inclusion in the analysis.

7.4.4 Joint longitudinal-survival model

Both frequentist and Bayesian joint longitudinal-survival models were fitted to the data, using Stata and R (see section 5.4.2). Subject-specific predictions could only be obtained utilising Bayesian estimation via the R package JMBayes [202]. The baseline hazard was modelled using cubic b-splines by default, with knots placed at equally spaced percentiles of the observed event times. Independent univariate normal diffuse priors were assumed for the vector of fixed effects for the longitudinal biomarker, the regression coefficients of the survival model and the vector of spline coefficients used to model the baseline hazard and the association parameter. An inverse Wishart prior distribution was assumed for the covariance of the random effects and an inverse-Gamma prior for the variance of the residual error terms. Proposal distributions (taken from separately fitted LMM and Cox PH models) were tuned for 3,000 iterations. The burn-in length was 3,000 iterations, and the number of repetitions was a default of 20,000. Chains were thinned to 2,000 iterations. More details on the estimation can be found in section 5.8.

Models with current value and first derivative association structures were fitted to the data. This allowed investigation of the change in absolute value and rate of change of log hCG.

7.4.5 Subject-specific survival predictions

Subject-specific survival probabilities were predicted for a selection of biochemical losses, early losses and healthy pregnancies. Estimation of these conditional survival probabilities is described in detail in section 5.11.2.

7.4.6 Discrimination and calibration

The predictive capabilities of the model were assessed using extensions to calibration and discrimination measures tailored to joint longitudinal-survival models. Methods for estimating these measures are described in section 5.12. Model sensitivity and specificity were calculated and ROC curves were plotted. Corresponding ROC AUCs were estimated for specified prediction windows from plotted ROC curves. Cut-points were identified utilising both the F score and Youden's index. Model calibration was assessed by estimating the prediction error in the predicted event rates using both a square and absolute loss function [205].

7.4.7 Multivariate joint model

A multivariate joint longitudinal-survival model was fitted including log P3G as the second longitudinal biomarker along with log hCG. Log P3G was modelled utilising three restricted cubic splines, a random intercept and a two-spline random slope. The P3G trajectory model included BMI, as obesity is known to be associated with low P3G levels in early pregnancy [270]. The K longitudinal outcomes were modelled using a GLMM, shown in Equation 7.1 [271]. Here, $\mathbf{y}_{ki}$ is the vector of longitudinal responses for the k^{th} outcome for individual i. The conditional distribution of $\mathbf{y}_{ki}$ given the vector of random effects $\mathbf{u}_{ki}$ follows a distribution from the exponential family. The link function is denoted by $g_k(.)$ The linear predictor ν_{ki} is made up the design vectors $\mathbf{x}_{ki}(t)$ and $\mathbf{z}_{ki}(t)$ of the

fixed and random effects $\boldsymbol{\beta}_k$ $\mathbf{u}_{ki}$ respectively. This reduces to the LMM for a continuous longitudinal response.

$$g_k\left[E\left\{\mathbf{y}_{ki}(t) \mid \mathbf{u}_{ki}\right\}\right] = \nu_{ki}(t) = \mathbf{x}_{ki}(t)\boldsymbol{\beta}_k + \mathbf{z}_{ki}(t)\mathbf{u}_{ki} \tag{7.1}$$

The longitudinal outcomes are linked by assuming all of the random effects $\mathbf{u}_i$ across the two outcomes are normally distributed with mean zero and variance-covariance matrix $\mathbf{D}$. The corresponding survival submodel is shown in Equation 7.2.

$$h_i(t \mid H_i(t), \mathbf{w}_i(t)) = h_0(t)\exp\left[\boldsymbol{\gamma}\mathbf{w}_i(t) + \sum_{k=1}^{K}\sum l = 1^{L_k} f_{kl}\left\{H_{ki}(t), \mathbf{w}_i(t), \mathbf{u}_{ki}, \boldsymbol{\alpha}_{kl}\right\}\right] \tag{7.2}$$

The model is conditional on the longitudinal history up to time t such that $H_{ki}(t) = \{\nu_{ki}(s), 0 \leq s < t\}$. The vector of covariates is given by $\mathbf{w}_i(t)$ with associated parameters $\boldsymbol{\gamma}$. The association structures for each longitudinal outcome is determined by the functions $f_{kl}(.)$ which are parameterised by the vector $\boldsymbol{\alpha}_{kl}$.

For the longitudinal outcomes, independent normal prior distributions were assumed for the fixed effects, and inverse gamma prior distributions for the scale parameters. For the survival submodel independent normal prior distributions were assumed for the covariate and association parameter vectors. The variance-covariance matrix for the random effects was parameterised in terms of a correlation matrix and vector of standard deviations. A Lewandowski-Kurowicka-Joe correlation prior distribution was assumed for the correlation matrix and a half-Student's t prior distribution for each standard deviation.

All models were fitted using `merlin` in Stata IC version 15.1 and/or JMBayes in R version 4.01 [200; 202].

7.5 Results

A total of 4,025 participants were recruited into the study. Of these, 3,869 women met the initial eligibility criteria and were consented. A total of 1,829 individuals were excluded before beginning the trial. On testing 572 women conceived pre-trial onset, a further two participants were peri-menopausal and no longer eligible. A number of individuals withdrew consent (n=253). The most prevalent reason was a change in personal circumstances which meant they were no longer trying to conceive (n=87) or a lack of interest (n=88). Only 27 women withdrew due to the stress or concerns related to the study itself (n=27). The number of participants withdrawn by the trial team totalled 325. The majority of these women did not complete or comply with collections (n=125). Sixty-seven participants did not receive required study materials. A further 677 participants were lost to follow-up after consent was given but before study onset.

The 2,040 participants who began the study completed 3,904 cycles. Of these 2,040 individuals, 376 women conceived. One ectopic pregnancy was excluded from analysis as evidence suggests that hCG profiles for ectopic pregnancies are not always discernible from viable intrauterine or miscarried pregnancies [272]. Data on hCG was not available for five participants. Two withdrew consent, one was withdrawn by the study team due to non-compliance, another was lost to follow-up and finally one volunteer was diagnosed with polycystic ovary syndrome (PCOS). A full breakdown of participant retention and flow is presented in Figure 7.1.

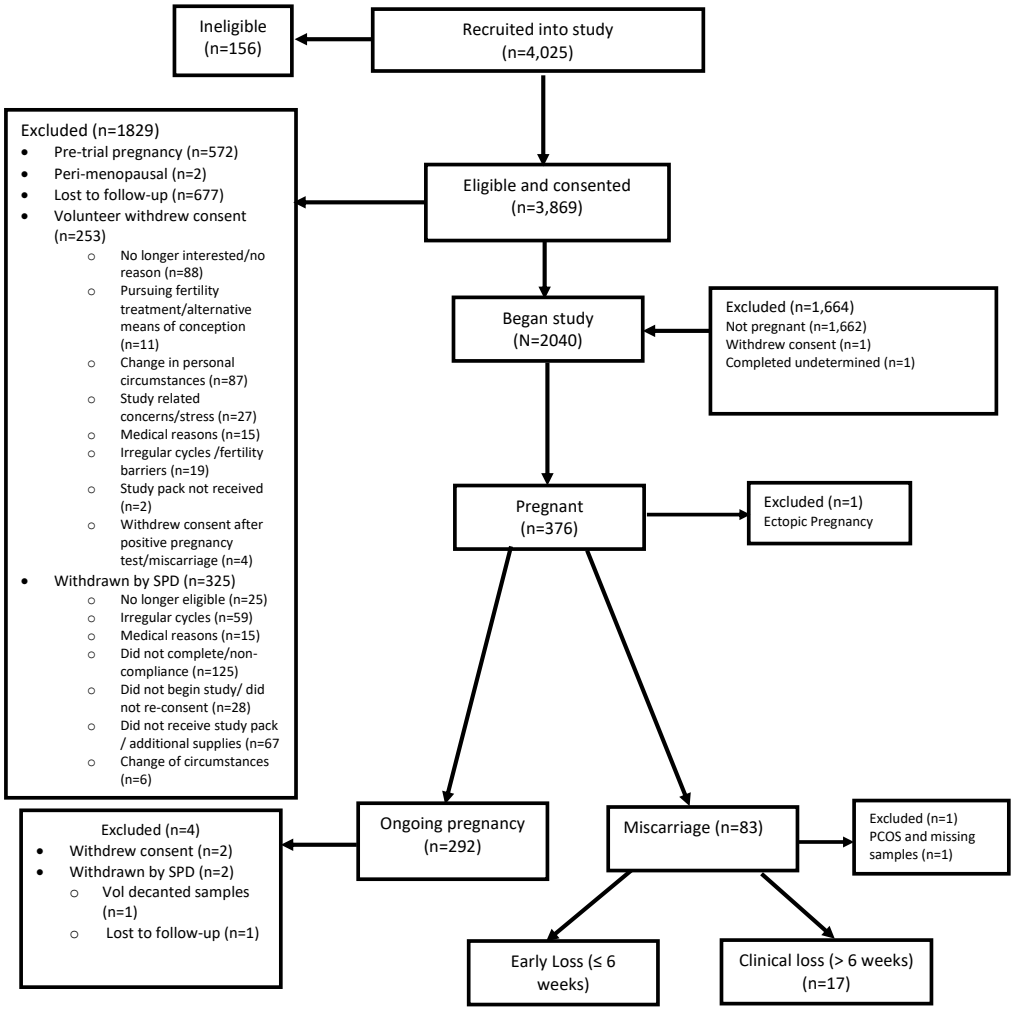


FIGURE 7.1. Flowchart for patient retention and flow

Three of the 370 pregnancies were twin pregnancies, and were excluded from analysis. This is because hCG levels tend to peak at a higher level for twin pregnancies in comparison to single pregnancies, pointing towards an alternative hCG trajectory [273]. The final data consisted of 285 (77.7%) viable and 82 (22.3%) miscarried pregnancies. Of the miscarried pregnancies, 65 (79.3%) were classified as early losses ($\leq$ 6 weeks) and 17 (20.7%) as clinical losses ($>$ 6 weeks).

Baseline characteristics are presented in Table 7.1. Maternal characteristics were similar across healthy and miscarriage groups. Women who experienced viable pregnancies were slightly younger on average than those who miscarried, 30.23 (SD: 4.66) and 31.37 (SD: 5.84) respectively. Mean BMI was higher at 26.87 (SD: 5.97) for women who had viable pregnancies than the mean of 26.37 (SD: 5.30) for those who experienced losses. The average cycle length was similar across both groups at approximately 29 days (29.26 (SD: 3.20) and 29.32 (SD: 3.68) for viable and miscarriage groups respectively). Women who experienced healthy pregnancies had been trying for slightly longer on average at 8.16 months (SD: 9.61) than women who miscarried at 6.33 months (SD: 7.91). This tended to vary greatly amongst the women within each group, with a median of 5.0 months (Q_1, Q_3: 3.0, 10.0) for those who experienced viable pregnancies and 4.0 months (Q_1, Q_3: 2.0, 7.0) for women who suffered a loss. Sixty-one percent (n=174) of individuals who had a healthy pregnancy had experienced a previous live birth compared with a slightly higher 69.5% (n=57) of women who miscarried. For both groups a higher proportion of women described themselves as non-smokers, 174 (61.1%) and 51 (62.2%) for viable and miscarried pregnancy groups, respectively. Almost 95% of individuals across both groups were of a White-European ethnicity, 268 (94.0%) for viable and 76 (92.7%) for miscarried groups.

TABLE 7.1. Maternal demographics for healthy and miscarried pregnancies

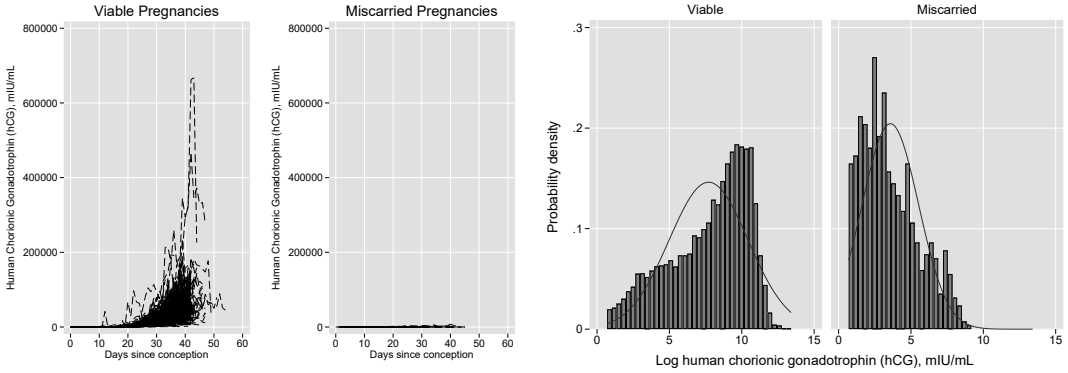
Variable	Healthy pregnancies (n=285 (77.7%))	Miscarried pregnancies (n=82 (22.3%))	Overall (n=367 (100.0%))
Age, years	30.23 (4.66)	31.37 (5.84)	30.49 (4.96)
BMI, kg/m^2	26.87 (5.97)	26.37 (5.30)	26.76 (5.82)
Height, m	1.65 (0.07)	1.64 (0.07)	1.65 (0.07)
Weight, kg	73.42 (16.89)	71.17 (15.70)	72.92 (16.64)
Live births	174(61.05)	57(69.51)	231(62.94)
Shortest cycle length	26.98 (3.44)	27.02 (3.73)	26.99 (3.50)
Longest cycle length, days	32.30 (5.60)	32.13 (7.10)	32.26 (5.95)
Average cycle length, days	29.26 (3.20)	29.32 (3.68)	29.28 (3.31)
Ectopic pregnancies	0.03 (0.18)	0.05 (0.22)	0.04 (0.19)
Time trying to conceive, months	8.16 (9.61)	6.33 (7.91)	7.75 (9.28)
Previous miscarriages	0.77 (1.22)	0.65 (0.92)	0.74 (1.16)
FSH3 level on day of implantation	1.85 (1.83)	3.37 (3.10)	2.16 (1.20)
P3G level on day of implantation	23.85 (16.62)	17.27 (13.21)	22.53 (16.20)
Smoking status, n(%)			
Non smoker	174 (61.1%)	51 (62.2%)	225 (61.3%)
Current/previous smoker	111 (38.9%)	31 (37.8%)	142 (38.7%)
Ethnicity, n(%)			
White	268 (94.0%)	76 (92.7%)	344 (93.7%)
Black	5 (1.8%)	1 (1.2%)	6 (1.6%)
Asian	7 (2.5%)	3 (3.7%)	10 (2.7%)
Mixed	5 (1.8%)	2 (2.4%)	7 (1.9%)
Education, n(%)			
Level 1	4 (1.4%)	0 (0.0%)	4 (1.1%)
GCSE	51 (18.0%)	16 (19.5%)	67 (18.3%)
A-Levels	75 (26.4%)	28 (34.1%)	103 (28.1%)
Higher Education	21 (7.4%)	6 (7.3%)	27 (7.4%)
Degree	83 (29.2%)	18 (22.0%)	101 (27.6%)
Postgraduate Degree	34 (12.0%)	9 (11.0%)	43 (11.7%)
Other	16 (5.6%)	5 (6.1%)	21 (5.7%)
Occupation, n(%)			
Student	6 (2.1%)	2 (2.4%)	8 (2.2%)
Manager/Senior Official	17 (6.0%)	5 (6.1%)	22 (6.0%)
Professional Occupation	75 (26.3%)	18 (22.0%)	93 (25.3%)
Associate Professional / Technical	44 (15.4%)	11 (13.4%)	55 (15.0%)
Administrative/Secretarial	45 (15.8%)	20 (24.4%)	65 (17.7%)
Skilled Trade	5 (1.8%)	1 (1.2%)	6 (1.6%)
Personal Service	49 (17.2%)	13 (15.9%)	62 (16.9%)
Sales/Customer Service	26 (9.1%)	8 (9.8%)	34 (9.3%)
Process, Plant, Machine Operative	3 (1.1%)	0 (0.0%)	3 (0.8%)
Elementary	15 (5.3%)	4 (4.9%)	19 (5.2%)

All values are mean(SD) unless otherwise stated

Most women who experienced healthy pregnancies had completed higher education (n(%): 138(48.6%)), with degree level education the most prevalent response (n(%): 34(12.0%)). Higher education overall was the most popular response for women who miscarried (n(%): 33(40.3%)), again with degree the most common response (n(%): 12(22.0%)). A quarter of women in the viable pregnancy group were educated to A-Level standard (n(%): 75 (26.4%)), and a slightly higher number of women who miscarried were similarly educated (n(%): 28 (34.1%)). A similar proportion of participants had experienced a miscarriage previously in both groups, 124(43.5%) and 37(45.1%) for the viable and miscarried pregnancy groups respectively. The mean number of prior miscarriages for women who experienced healthy pregnancies was slightly higher at 0.77 (SD: 1.22) than the 0.65 (SD: 0.92) for those who suffered losses. The median number of previous losses was 0 (Q_1, Q_3 : 0, 1) in both groups. The average FSH3 level on the day of implantation was lower in the viable pregnancy group at 1.85 (SD: 1.83) than 3.37 (SD: 3.10) in the miscarriage group. In contrast the average level of P3G on the day of implantation was higher for viable pregnancies (23.85, SD: 16.62) than for miscarried pregnancies (17.27, SD: 13.21).

7.5.1 hCG trajectories

Absolute values of hCG and log transformed by pregnancy viability group are shown in Figure 7.2. There were 531 hCG observations which fell below the detection limit of 2 mIU/ml. Seven women had no recorded hCG observations above this limit. After exclusion of these observations and the log transformation the data was closer to normally distributed.



(A) hCG trajectories for viable pregnancies and miscarried pregnancies

(B) Histogram for log hCG trajectories by viable and miscarried pregnancy groups

FIGURE 7.2. Raw and log transformed hCG observations

Women who miscarried contributed a median of 15 (Q_1, Q_3: 8, 25) hCG observations, whilst women who experienced viable pregnancies contributed 32 (Q_1, Q_3: 30, 34) observations on average. For early loss pregnancies the median number of observations was 10 (Q_1, Q_3: 6, 15) and for clinical losses the median number was 25 (Q_1, Q_3: 16.5, 33).

Plots of individual log hCG trajectories by pregnancy outcome are displayed in Figure 7.3. There was a marked difference in profiles of log hCG for viable and failing pregnancies. In general, there was a consistent rise in hCG for healthy pregnancies. The initial rate of increase of hCG was steep through the first three weeks after conception, before slowing in rise. Women who miscarried, presented with an initial rise after conception. Due to the length of follow-up, it was possible to observe complete profiles for the majority of women who miscarried, with hCG levels declining. This is a contrast to the partial miscarried trajectories from Chapter 6 (see Figure 6.2 for comparison). The point at which a drop in hCG was observed spanned as early as 2 weeks post conception to as late as 6 weeks.

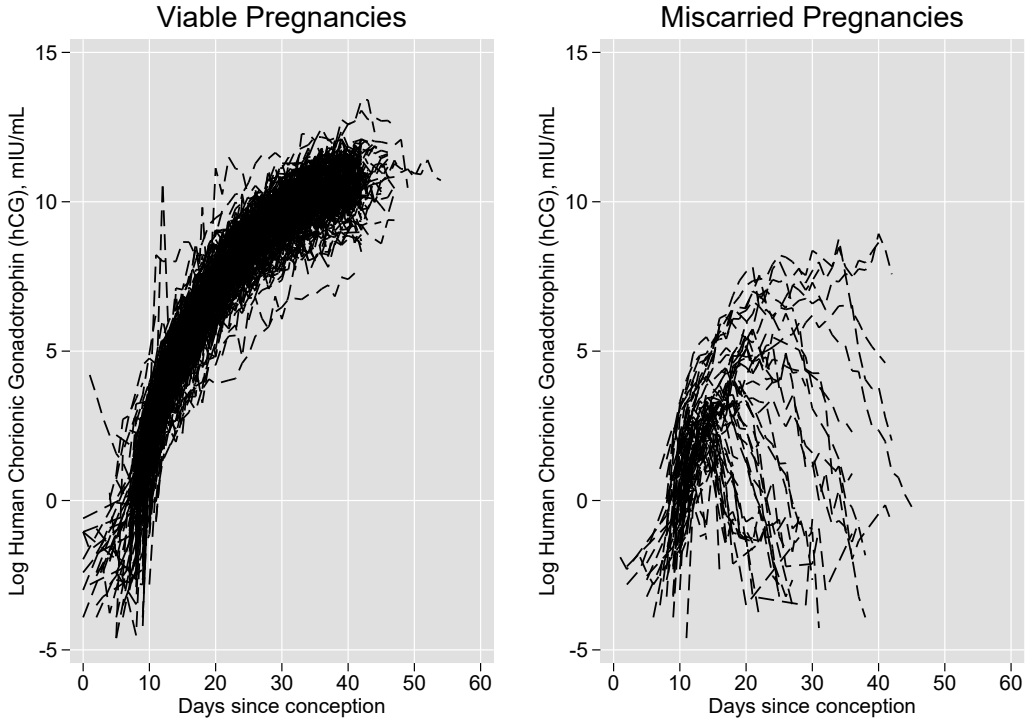
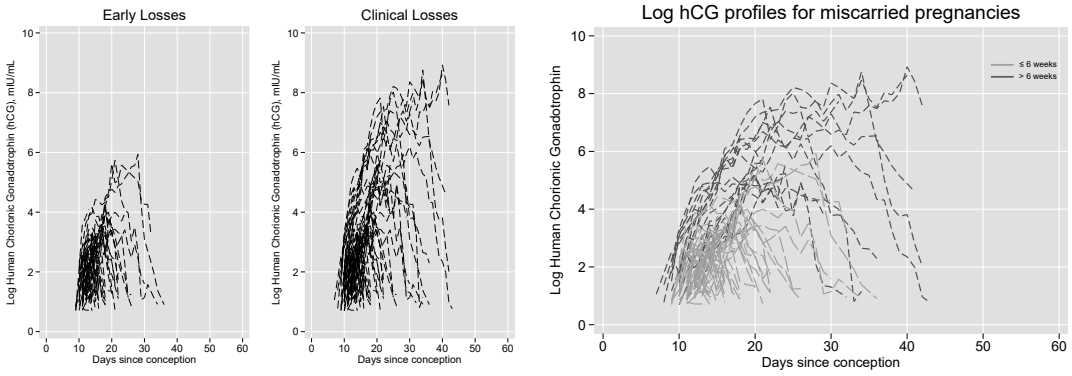


FIGURE 7.3. Log hCG trajectories for viable pregnancies and miscarried pregnancies



(A) By early and clinical losses) (B) Overlaid early and clinical losses

FIGURE 7.4. Log hCG trajectories for women who miscarried

Figure 7.4 show trajectory plots for the early and clinical losses. Trajectories for early losses follow the same shape as the clinical losses, however, they differ in the height of the log hCG peak.

7.5.2 Data exploration

Variables of interest and potential associations between pairs of variables were explored. This was to determine the possibility of interactions to be included at the modelling stage. Histograms and boxplots for age, BMI, cycle length and previous number of miscarriages are shown in Figure 7.5.

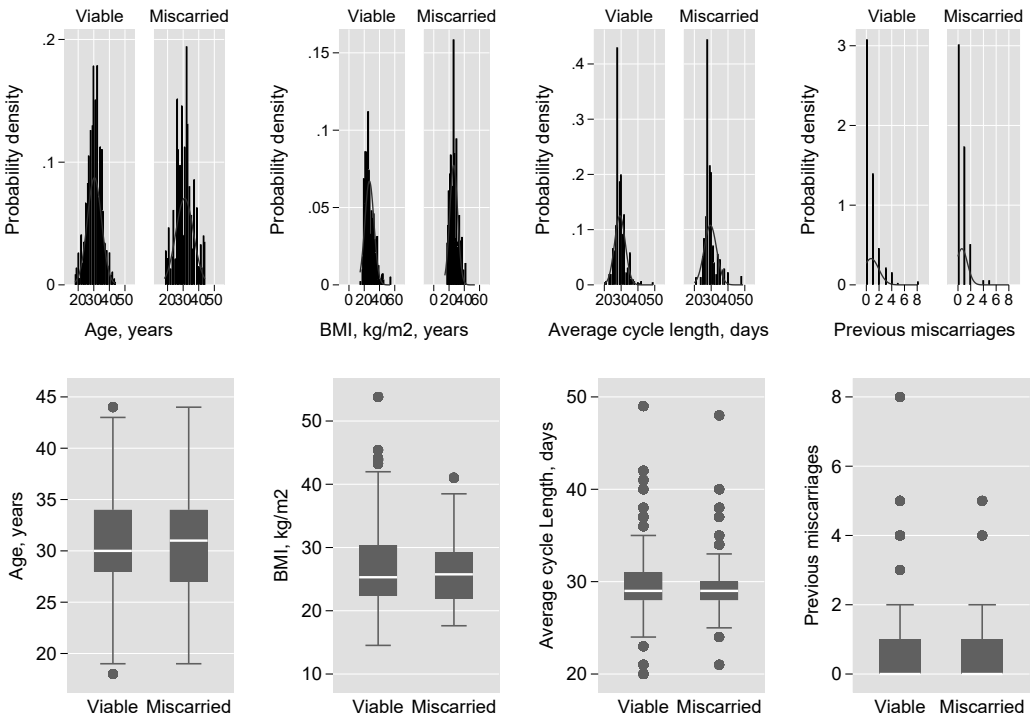


FIGURE 7.5. Histograms and box plots for age, BMI, cycle length and previous number of miscarriages by pregnancy viability group

Age was fairly normally distributed, although there was slightly more variation amongst women who miscarried. A larger inter-quartile range of $33 - 27 = 6$ was

observed for those who miscarried than the $33 - 28 = 5$ for those who had viable pregnancies. The range was similar across both groups with a minimum age of 18 and 19 years for the viable and loss groups respectively and an upper limit of 44 years of age across both groups.

There were several abnormally high and low BMI values amongst participants. Overall there was less variability in BMI amongst women who miscarried with an inter-quartile range of 5.9kg/m^2 compared to 8.0kg/m^2 for viable pregnancies. Six women who experienced healthy pregnancies had a BMI above 42.4kg/m^2, more than 1.5 times the interquartile range, placing them in the extremely obese category. One woman who miscarried had a BMI greater than 36.8, more than 1.5 times the interquartile range. A single participant was classed as extremely underweight with a BMI of 14.5 kg/m^2 in the viable pregnancy group.

Overall cycle length for both viability groups were fairly normally distributed. The inter-quartile range for viable and non-viable pregnancies spanned 28 to 31 and 30 days respectively. Select women in both groups experienced unusually long and short cycle lengths. A cycle length greater than 40 days was observed for three participants who experienced healthy pregnancies, with a maximum observed cycle length of 49 days. One woman reported an average cycle length of 48 days amongst women who eventually miscarried. Only one women had an average cycle length of less than 21 days (20 days), and was in the viable pregnancy group.

The number of previous miscarriages were positively skewed, with the median of no previous losses in both groups. The maximum number of losses in the viable pregnancy group was higher at 8 than the 5 in the failing pregnancy group. Twelve women who experienced healthy pregnancies, reported having more than 3 previous losses compared with two women in the failing pregnancy group. More

than three losses is defined as recurrent loss. A three-group categorisation was considered, no previous loss, $\leq$ 3 losses and $>$ 3 losses, however with small numbers in the final category, the variable was modelled continuously.

7.5.2.1 Variable associations

Associations between pairs of variables were investigated where there was evidence of one in the literature. Scatter plots of pairs of continuous variables are presented in Figure 7.6.

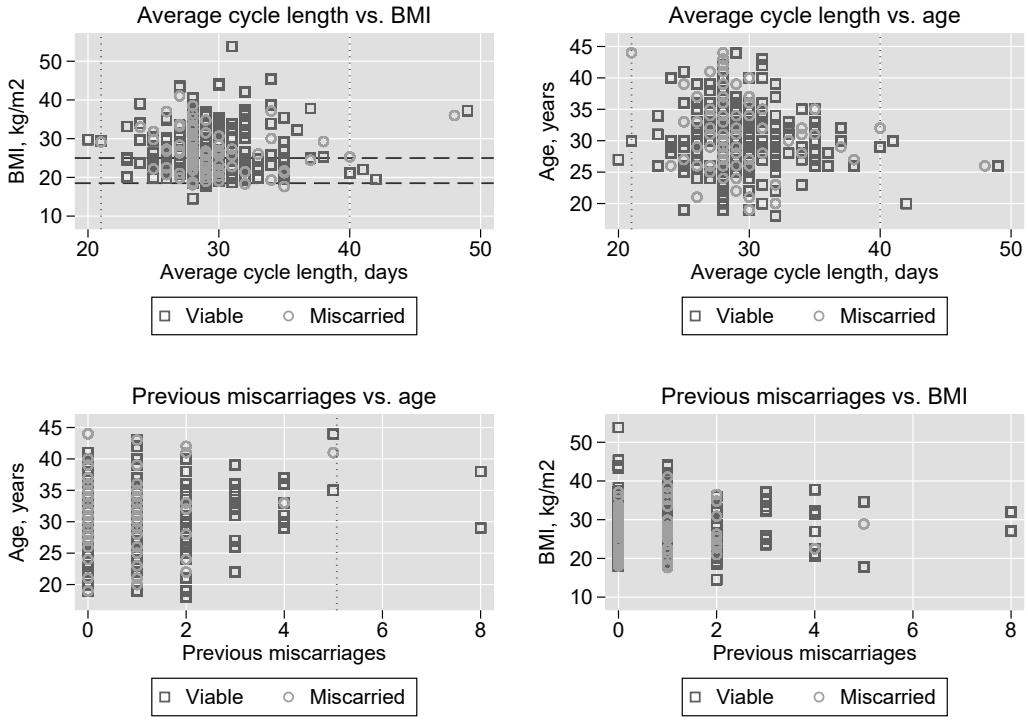


FIGURE 7.6. Scatter plots for associations between average cycle length against BMI and age and number of previous miscarriages against age and BMI, by pregnancy viability group

Being underweight or overweight can contribute to irregular menstrual cycles [274]. The majority of women in the dataset were classed as overweight or obese according to BMI. However, despite their high BMIs, most women had an average

cycle length within the normal range. Two participants who exceeded the 40 day cycle length were within the healthy BMI range, whilst the single participant for whom a cycle length of 20 days was observed was clinically overweight according to BMI. One individual who was very underweight with a BMI kg/m^2 of 14.53 had an average cycle length of 28 days. There was no evidence to suggest that cycle length was associated with BMI.

Cycle length has been suggested as a surrogate for reproductive age [275]. Women who are peri-menopausal experience a shortening of cycle length due to a shortening of the follicular phase; the time preceding ovulation. This means ovulation occurs more often. The majority of women in the dataset regardless of age had a cycle length approximately between 23 and 36 days. Women with a cycle length greater than 36 days were mostly aged between 25 to 35 years, with the individuals with the longest cycle lengths of 48 and 49 days both aged 26 years. Again there was no evidence in this case that cycle length and age were associated.

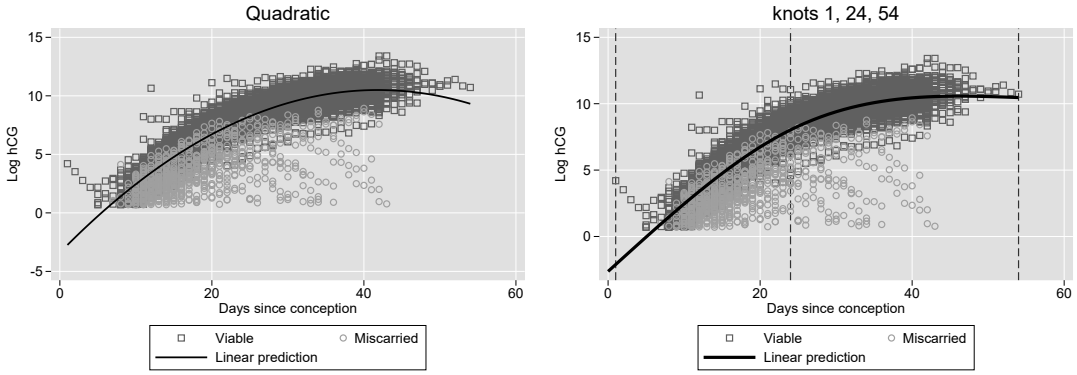
The odds of a subsequent miscarriage increases 2-fold for women who have experienced one previous loss and almost four-fold for those who have experienced three previous losses [68]. The majority of women in the data had not experienced a previous loss and this spanned the entire age range of the dataset. Multiple previous losses were also spread out across women of all ages. Four of the five women who had experienced five or more miscarriages previously were 35 years or over. Women who had previously experienced eight losses were heavier.

7.5.3 Longitudinal Modelling

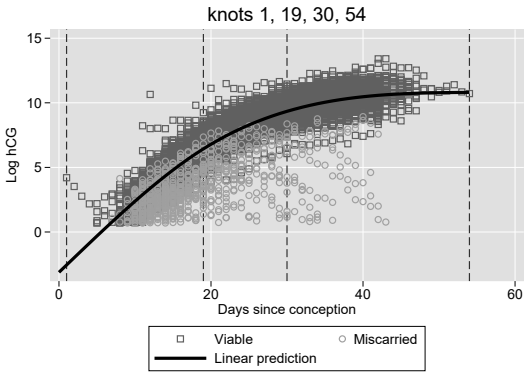
A quadratic time since conception term was utilised to describe the longitudinal trajectory in Chapter 6. This was problematic as the quadratic did not adequately

model the longitudinal trajectory. In addition, covariates with standing associations with hCG were not considered for inclusion in the longitudinal trajectory model.

Figure 7.7 shows the various functional forms of time since conception which were evaluated for modelling log hCG. The quadratic trajectory dips at the tail end of follow-up, however hCG observations for a viable pregnancy would plateau at this point.



(A) Linear and quadratic time term (B) Two restricted cubic spline time terms



(C) Three restricted cubic spline time terms

FIGURE 7.7. Modelling longitudinal log hCG with polynomials and restricted cubic splines

To improve upon this specification restricted cubic splines (RCS) were used to capture the shape of the log hCG profile. Initially two RCS were utilised,

with knots placed at day 1, 24 and 54 post conception. This was compared to a three-RCS model, with knots placed at 1, 19, 30, and 54 days post conception. The overall fit of the RCS models was similar up until the end of follow-up. The two-spline model implied a plateau, perhaps even a marginal dip at the day 54 mark, the three-spline model suggested a continued rise in hCG.

Model fit statistics for these models are presented in Table 7.2. The quadratic and two-spline models were an improvement on the basic linear model with reductions of AIC from 31761.99 (BIC: 31783.3) to 29066.47 (BIC: 29094.89) and 29014 (BIC: 29042.59) respectively. The three-spline model did not noticeably improve upon the two-spline model. For this reason the two-spline model was selected.

TABLE 7.2. Model fit statistics for the functional form of the longitudinal trajectory

Model	Log likelihood	df	AIC	BIC
Basic Model	-15877.99	3	31761.99	31783.3
Quadratic model	-14529.24	4	29066.47	29094.89
Two-RCS model*	-14503.08	4	29014.17	29042.59
Three-RCS model	-14498.83	5	29007.66	29043.19

df: degrees of freedom
*Final functional form

The addition of random effects was based on comparison of the log likelihood and the AIC. As models were fitted using ML estimation, nested models could be compared using likelihood ratio tests (see section 3.5.6). Details of fit estimates are given in Table 7.3. Addition of a random intercept to allow for individual variation at baseline, resulted in a reduction in AIC from 29014.17 to 20909.89. Addition of the random linear slope further reduced the AIC to 15805.29. For both random effect additions the likelihood ratio test gave a p-value of less than 0.0001. The addition of the random non-linear slope was statistically significant ($p<0.0001$). This was ultimately rejected, however, due to firstly the added

complexity of computation at the joint model stage and secondly the improvement in fit did not compensate for the addition of computing an extra three parameters.

TABLE 7.3. Model fit statistics for the addition of random effects to the longitudinal trajectory model

Model	Log likelihood	df	AIC	BIC
Fixed effects model	-14503.08	4	29014.17	29042.59
Random intercept model	-10449.94	5	20909.89	20945.41
Random linear slope	-7895.647	7	15805.29	15855.03
Random non-linear slope	-7355.079	10	14730.16	14801.21
df: degrees of freedom				

An LMM fitted to the longitudinal log hCG data, including a viability group variable showed profiles were significantly different (Table 7.4). Women who suffered a biochemical miscarriage had a 2.703 lower (95%: -2.914, -2.493) log hCG level on average than women who went on to have healthy pregnancies. Women who miscarried after six weeks had a 1.025 lower (95% CI: -1.362, -0.688) mean log hCG level than women who had viable pregnancies.

TABLE 7.4. Longitudinal model with pregnancy viability grouping variable

Longitudinal model	Mean change	95% CI	
RCS 1 time, days	0.508	0.495,	0.521
RCS 2 time, days	0.0003	0.0003,	0.0003
Group			
Healthy	Reference		
Biochemical loss	-2.703	-2.914,	-2.493
Early loss	-1.025	-1.362,	-0.688
Constant	-2.148	-2.370,	-1.927
CI: Confidence Interval			

The results for the longitudinal model including the centred covariates BMI and average cycle length are given in Table 7.5. A one kg/m^2 increase in BMI from 25kg/m^2, the upper limit of the normal range of BMI, resulted in a statistically significant 0.034 decrease (95% CI: -0.049, -0.019) in log hCG. A minimal 0.004

increase (95%: -0.022, 0.031) in log hCG was observed for each one-day increase on a 28-day cycle. This was not statistically significant at the 5% level. The overall variance at baseline was 1.816 (95% CI: 1.506, 2.189). The slope variance was 0.016 (95% CI: 0.013, 0.020).

TABLE 7.5. Longitudinal model estimates

Longitudinal model	Mean change		95% CI
RCS time 1, days	0.500	0.485,	0.514
RCS time 2, days	0.0003	0.0003,	0.0003
Centred BMI at 25kg/m^2	-0.034	-0.049,	-0.019
Centred cycle length at 28 days	0.004	-0.022,	0.031
Constant	-2.453	-2.619,	-2.288
Random effect parameter	**Estimate**	**95% Confidence Interval**	
$\sigma^2_{u_1}$	0.016	0.013	0.020
$\sigma^2_{u_0}$	1.816	1.506	2.189
$cov(time, intercept)$	-0.139	-0.170	-0.108
σ^2_e	0.243	0.236	0.251

$\sigma^2_{u_1}$: slope variance
$\sigma^2_{u_0}$: intercept variance
$cov(time, intercept)$: covariance between slope and intercept
σ^2_e: measurement error variance
CI: Confidence interval

7.5.4 Survival modelling

The Kaplan-Meier survival probability for time-to-miscarriage is shown in Figure 7.8. Overall survival probability for the pregnancies at the end of follow-up was approximately 70%. There is a lag in events up to day 10 due to the use of time since conception as the timeline.

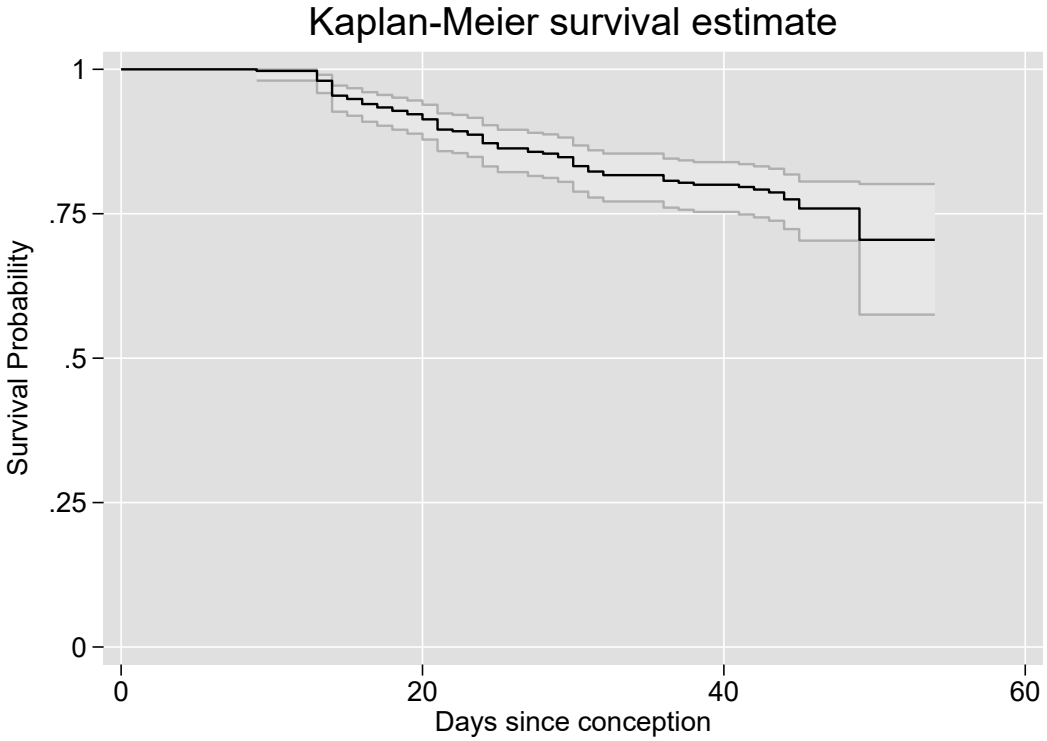


FIGURE 7.8. Kaplan-Meier Survival Estimates

The shape of the baseline hazard was explored. Figure 7.9 demonstrates that the most appropriate way to model the underlying baseline hazard was two or three spline terms. These functional forms of the baseline hazard were compared to the exponential, Weibull and Gompertz distributions. Estimates of fit can be found in Table 7.6.

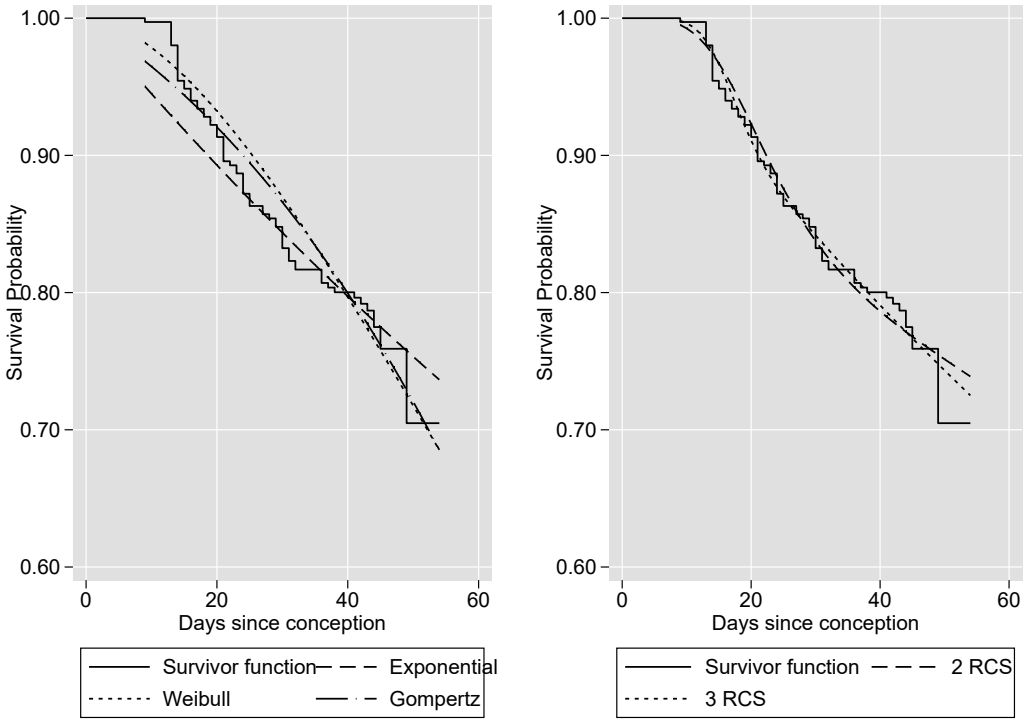


FIGURE 7.9. Survival probability curves for Exponential, Weibull, Gompertz and 2 and 3 restricted cubic spline baseline hazard models

TABLE 7.6. Model fit estimates for various functional forms of the baseline hazard

Model	Log likelihood	df	AIC	BIC
Exponential	-229.27	1	460.5401	464.4262
Weibull	-219.3006	2	442.6012	450.3734
Gompertz	-224.9454	2	453.8908	461.663
Two RCS	-208.0396	3	422.0792	433.7375
Three RCS	-205.2554	4	418.5109	434.0553
df: degrees of freedom				

The flexible parametric models fit best, with little difference in AIC between the two-RCS and three-RCS models. Of the distributions the Weibull model gave the most appropriate fit. Though modelling the baseline hazard flexibly was

attempted, the Weibull was ultimately chosen as it is the most stable when fitting a joint model. This was seen in the analysis presented in Chapter 6. Considering the number of variables intended to be included in the survival submodel, the simpler Weibull distribution was preferred over a more flexible baseline hazard.

Estimates for a Weibull survival model including the time-varying covariate for log hCG are shown in Table 7.7. Using the most naive method (see section 5.3.2) a unit increase in log hCG suggested a 52.5% decrease (HR: 0.475, 95% CI: 0.307, 0.734) in the rate of miscarriage.

TABLE 7.7. Weibull survival model for time-varying log hCG

Variable	Hazard Ratio	95% CI
Log hCG	0.475	0.307, 0.734
CI: Confidence Interval		

The additional variables suggested for inclusion in the model for their recognised association with early miscarriage were maternal age, BMI, smoking status, number of previous miscarriages, along with P3G and/or FSH3. Non-linear effects for age and BMI were explored and the AIC and BIC estimates for each addition to the model are shown in Tables 7.8 and 7.9. The addition of quadratic centred age was significant at the 5% level (p=0.002) and was included in the model. No non-linear BMI terms were included in the model.

TABLE 7.8. Survival model fit estimates for different functional forms of centred age

Model	Log Likelihood	df	AIC	BIC
Linear age	-183.800	9	385.600	420.550
Quadratic age	-179.043	10	378.087	416.920
RCS 2df	-179.405	10	378.811	417.644
RCS 3df	-177.864	11	377.728	420.445
df: degrees of freedom				

TABLE 7.9. Survival model fit estimates for different functional forms of centred BMI

Model	Log likelihood	df	AIC	BIC
Linear BMI	-183.800	9	385.600	420.550
Quadratic BMI	-182.905	10	385.810	424.644
Two RCS	-182.959	10	385.919	424.752
df: degrees of freedom				

The levels of P3G and FSH3 on the day of implantation were included in the model as a single observation per individual. Table 7.10 shows that the level of P3G on the day of implantation was significantly associated with time-to-miscarriage. A one μg/ml increase in P3G on the day of implantation reduced the rate of miscarriage by 2.4% (HR: 0.976, 95% CI: 0.956, 0.996).

TABLE 7.10. Weibull survival model including P3G on the day of implantation

Variable	Hazard Ratio	95% CI	
Log hCG	0.566	0.357,	0.897
Centred age, years	1.013	0.973,	1.055
Quadratic centred age, years	1.009	1.004,	1.014
Centred BMI, kg/m^2	0.971	0.930,	1.014
Smoking status			
Never	Reference		
Previous/current	0.795	0.487,	1.297
Previous number of miscarriages	0.972	0.774,	1.221
P3G on day of implantation, μg/ml	0.976	0.956,	0.996
CI: Confidence Interval			

When FSH3 on the day of implantation was included independently of P3G, there was a significant association with miscarriage. The results of this model are shown in Table 7.11. A one mIU/ml increase in the level of FSH3 on the day of implantation resulted in a 14.2% increase (HR: 1.237, 95% CI: 1.142, 1.339) in the rate of miscarriage.

TABLE 7.11. Weibull survival model including FSH3 on the day of implantation

Variable	Hazard Ratio	95% CI
Log hCG	0.520	0.335, 0.805
Centered age, years	1.009	0.969, 1.051
Quadratic centred age, years	1.008	1.003, 1.014
Centred BMI, kg/m^2	0.994	0.953, 1.037
Smoking status		
Never	Reference	
Previous/current	0.884	0.542, 1.440
Previous number of miscarriages	0.976	0.780, 1.225
FSH3 on day of implantation, mIU/ml	1.237	1.142, 1.339
CI: Confidence interval		

FSH3 was a statistically significant addition to the model when compared using a likelihood ratio test to the P3G only model ($p<0.0001$). The final survival submodel included both P3G and FSH3 levels on the day of implantation. Results are shown in Table 7.12.

TABLE 7.12. Survival submodel estimates with both P3G and FSH3 on the day of implantation

Variable	Hazard Ratio	95% CI
Log hCG	0.574	0.366, 0.900
Centered age, years	1.003	0.963, 1.044
Quadratic centred age, years	1.009	1.004, 1.014
Centred BMI, kg/m^2	0.982	0.941, 1.026
Smoking status		
Never	Reference	
Previous/current	0.864	0.527, 1.415
Previous number of miscarriages	0.996	0.791, 1.255
P3G on day of implantation, μg/ml	0.970	0.950, 0.991
FSH3 on day of implantation, mIU/ml	1.250	1.157, 1.350
Parameter	Estimate	95% CI
$\ln(\lambda)$	-4.184	-6.525, -1.844
γ	1.017	0.605, 1.709
CI: Confidence Interval		

The overall hazard ratio for centred age suggests that for women aged 31 and over there was an increase in the rate of miscarriage, whilst for women below 30 years of age there was a decrease in the rate of miscarriage, when keeping all other covariates constant. There were small decreases in the rate of miscarriage for a 1kg/m^2 increase in BMI from a BMI of 25kg/m^2 (HR: 0.982, 95% CI: 0.941, 1.026). Previous or current smokers had a 2.9% (HR: 0.864, 95%: 0.527, 1.415) lower rate of miscarriage than women who had never smoked. A unit increase in the number of previous miscarriages suggested a 0.04% (HR:0.996, 95% CI: 0.791, 1.255) decrease in the rate of miscarriage. None of these effects were statistically significant at the 5% level. A one μg/ml increase in the level of P3G on the day of implantation indicated a statistically significant 3.0% (HR:0.970, 95%: 0.950, 0.991) decrease in the rate of miscarriage. Whilst a one mIU/ml increase in the level of FSH3 on the day of implantation corresponded to a 25.0% (HR: 1.250; 95% CI: 1.157, 1.350) increase in the rate of miscarriage. A one-unit increase in log hCG, here modelled as a time-varying covariate, suggested a 42.6% decrease in the rate of miscarriage.

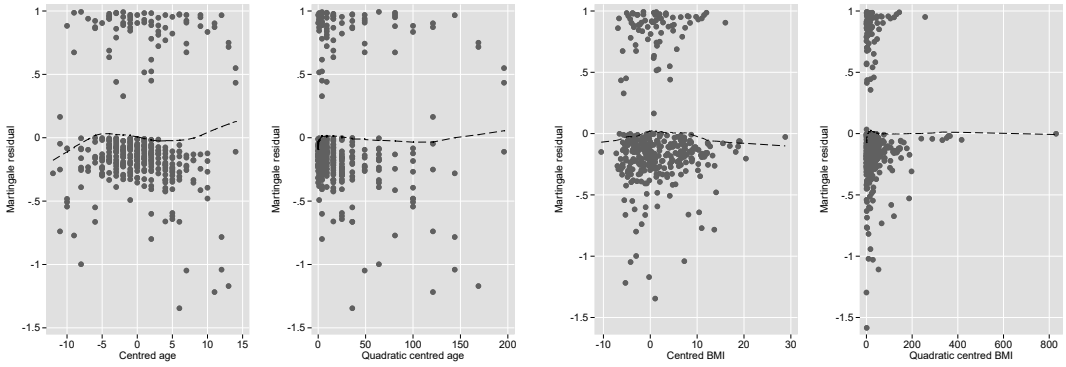
As FSH3 can be an indicator of reproductive age, interactions with maternal age were investigated. The model fit estimates for the added interactions are given in Table 7.13. The interactions with age did not significantly improve the fit of the model and were consequently not included.

TABLE 7.13. Model fit estimates for survival submodel FSH3 interactions with age

Model	Log likelihood	df	AIC	BIC
P3G only model	-190.448	9	366.547	433.846
FSH3 only model	-183.488	9	345.004	419.925
P3G and FSH3 model	-179.043	10	336.878	416.920
FSH3 interaction with linear age	-179.043	11	378.087	422.576
df: degrees of freedom				

Further interactions investigated included an interaction between centred age and centred BMI (p=0.6841), centred age and the number of previous miscarriages (p=0.7415), centred BMI and the number of previous miscarriages (p=0.2562), smoking status and the number of previous miscarriages (p=0.8875), log hCG and centred age (p=0.3909) and log hCG and centred BMI (p=0.8873). There was no evidence of an interaction between each of these pairs of variables.

Predicted Martingale residuals plotted against linear and quadratic centred age are shown in Figure 7.10a. The addition of the quadratic term improved the modelling of age, and confirms non-linearity. The martingale residual plots for BMI are shown in Figure 7.10b. There is indication that modelling BMI linearly is not sufficient. Despite evidence to the contrary at the modelling stage (see Table 7.9), inclusion of a quadratic centred BMI term addressed the non-linearity.



(A) Centred age (B) Centred BMI

FIGURE 7.10. Martingale residual plots

Plots for all other variables are shown in Figure 7.11. Linear terms were sufficient for the number of previous miscarriages, P3G on the day of implantation and log hCG. There was some curvature for FSH3 on the day of implantation towards the end of follow-up, most likely due to fewer observations.

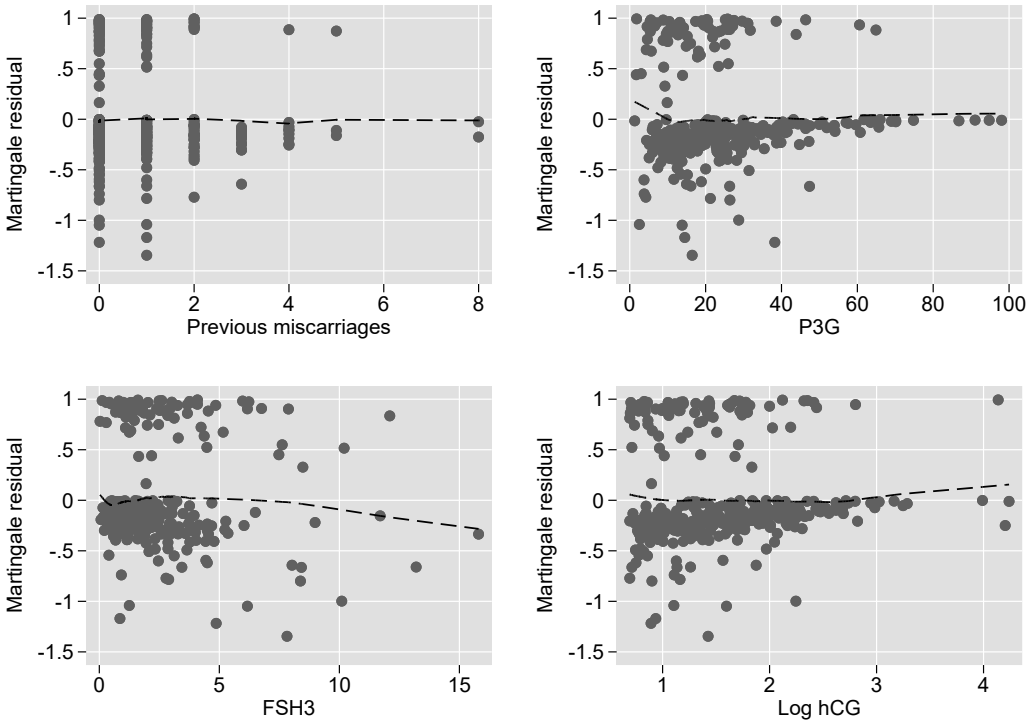


FIGURE 7.11. Martingale residual plots for previous miscarriages, P3G and FSH3 on the day of implantation and log hCG

The proportional hazards assumption was assessed for categorical variables by plotting the natural log analysis time against the -log[-log(survival probability)]. Other than at the beginning and end of follow-up Figure 7.12 shows fairly parallel lines for the never smoker and previous/current smoker groups.

Figure 7.13 shows the time-dependent log hazard ratios for each continuous variable included in the model. Non-proportional hazards, i.e. interactions with time, were not considered for log hCG, though it will be modelled as a time-varying covariate via the joint model due to the repeated observations. There was no evidence of a violation of the proportional hazard assumption for age, BMI, P3G or FSH3. The time-dependent log hazard ratio for previous number

of miscarriages showed some decrease over time, however there was no evidence that the addition of the time-dependent variable improved fit (p=0.413).

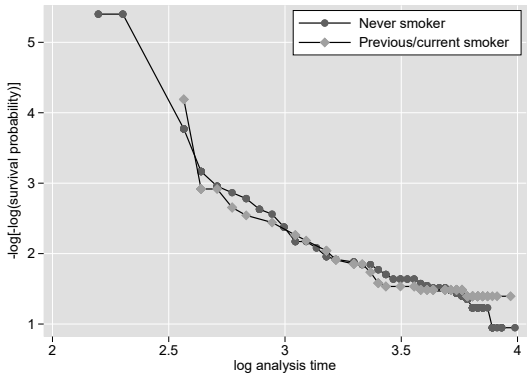


FIGURE 7.12. Plot of -log[-ln(survival probability)] against analysis time for smoking status

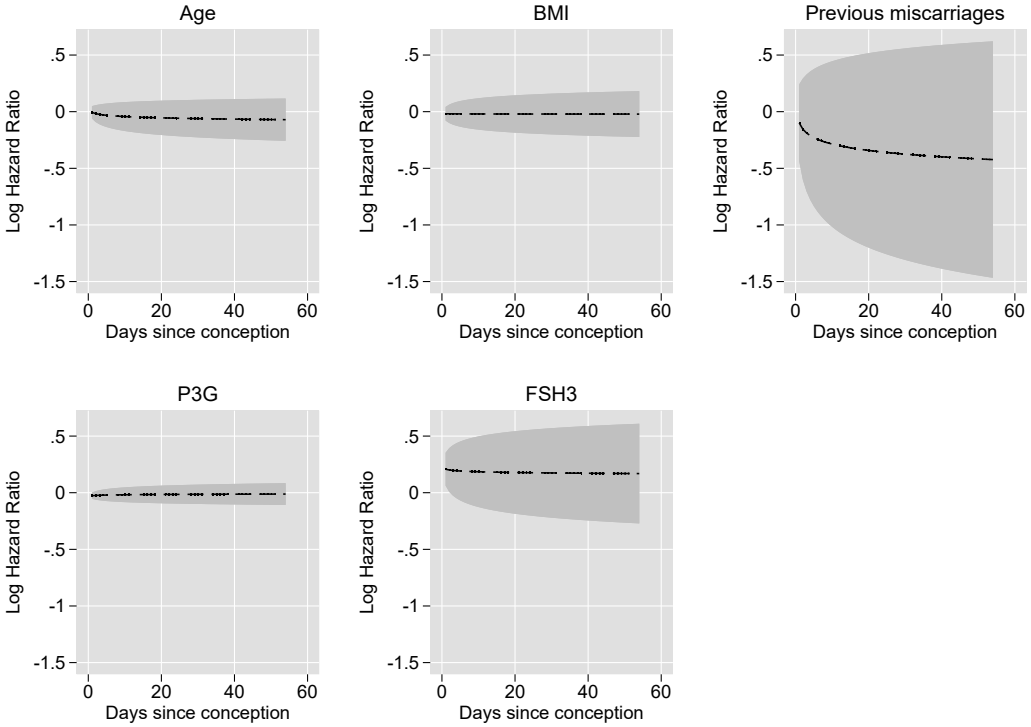


FIGURE 7.13. Time-dependent log hazard ratio plots for centred age and BMI, previous number of miscarriages, and P3G and FSH3 on day of implantation

Deviance residuals for the fitted model are shown in Figure 7.14. The majority of women with a deviance residual greater than two experienced miscarriages early on during follow-up. Sensitivity analyses were conducted by removing different subsets of these women in turn - eight women whose BMI was greater than 30kg/m^2, six women who were over the age of 35, three women who reported experiencing two previous losses and ten women who were previous or current smokers. Model estimates did not alter significantly after the removal of women who were older or had previously experienced two losses. However, when women with BMIs over 30kg/m^2 and women who were current or previous smokers were removed the effect of an increase in BMI (from 25 kg/m^2) or being a previous/current smoker had a protective effect which was significant at the 5% level.

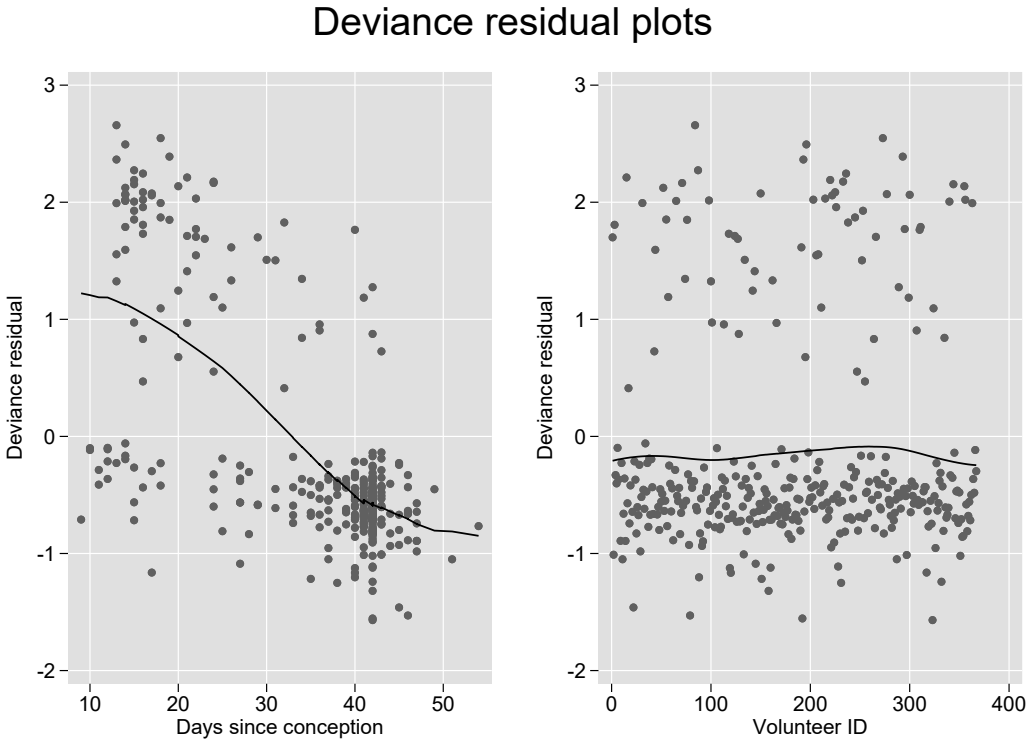


FIGURE 7.14. Deviance residuals for the fitted survival submodel

7.5.5 Joint longitudinal-survival model

The estimates from the joint longitudinal-survival model with current value association structure are presented in Table 7.14. As the intention was to obtain subject-specific predictions, the joint models were fitted in R. This capability is not yet available in Stata for longitudinal trajectories modelled using spline terms. The parallel command to `stjm` and `merlin` in R is JMfit, which was not able to fit the models successfully. Instead the Bayesian equivalent JMBayes was utilised.

TABLE 7.14. Model estimates for the fitted joint longitudinal-survival model with current value association structure

Survival submodel	Hazard Ratio	95% CrI	
Current value of log hCG, mIU/ml	0.469	0.407,	0.532
Centered Age, years	1.005	0.953,	1.051
Quadratic centred age, years	1.003	0.999,	1.010
Centered BMI, kg/m2	0.986	0.934,	1.038
Smoking status			
Never	Reference		
Previous/current	1.518	0.828,	2.726
Number of previous miscarriages	1.116	0.866,	1.409
P3G on day of implantation, μg/ml	1.002	0.980,	1.023
FSH3 on day of implantation, mIU/ml	1.222	1.121,	1.315
Longitudinal submodel	Mean change	95% CrI	
Restricted cubic spline time term 1	19.749	19.368,	20.177
Restricted cubic spline time term 2	5.319	4.926,	5.621
Centered BMI, kg/m^2	-0.051	-0.059,	-0.042
Centered cycle length, days	-0.0094	-0.019,	0.013
Intercept	-2.106	-2.193,	-2.018
Random effect parameter	Estimate	95% CrI	
$\sigma^2_{u_1}$	1.061	0.922	1.232
$\sigma^2_{u_0}$	1.348	1.113,	1.615
$cov(time, intercept)$	-0.032	-0.153,	0.097
σ^2_e	0.4811	0.474,	0.488
$\sigma^2_{u_1}$ slope variance			
$\sigma^2_{u_0}$ intercept variance			
$cov(time, intercept)$ covariance between the slope and intercept			
σ^2_e residual measurement error variance			
CrI: Credible Interval			

After accounting for possible confounding variables a strong association between log hCG and miscarriage was observed. Specifically a one-unit increase in

the current absolute value of log hCG resulted in a 53.1% (HR:0.469, 95% CrI: 0.407, 0.532) decrease in the rate of miscarriage at time t. An increase in P3G on the day of the day of implantation no longer provided a protective effect. An increase in FSH3 on the day of implantation continued to infer a significantly greater rate of miscarriage. A one mIU/ml increase in FSH3 indicated a 22.2% increase (HR:1.222, 95% CrI: 1.121, 1.315) in the rate of miscarriage.

TABLE 7.15. Model estimates for the fitted joint longitudinal-survival model with first derivative association structure

Survival submodel	Hazard Ratio	95% CrI	
Current value of log hCG	0.605	0.505,	0.712
Slope value of log hCG	0.007	0.0004,	0.138
Centered Age, years	1.006	0.958,	1.061
Quadratic centered age, years	1.003	0.997,	1.008
Centered BMI, kg/m2	0.987	0.934,	1.042
Smoking status			
Never	Reference		
Previous/current	1.692	954,	3.110
Number of previous miscarriages	1.136	0.859,	1.445
P3G on day of implantation, μg/ml	0.998	0.978,	1.019
FSH3 on day of implantation, mIU/ml	1.202	1.111,	1.300
Longitudinal submodel	**Mean change**	**95% CrI**	
Restricted cubic spline time term 1	20.146	19.703,	20.703
Restricted cubic spline time term 2	5.595	5.282,	6.037
Centered BMI, kg/m2	-0.040	-0.049,	-0.028
Centered cycle length, days	0.031	0.012,	0.057
Intercept	-2.182	-2.268,	-2.097
Random effect parameters	**Estimate**	**95% CrI**	
$\sigma^2_{u_1}$	1.062	0.912,	1.231
$\sigma^2_{u_0}$	1.376	1.135,	1.652
cov(time, intercept)	-0.034	-0.169,	0.107
σ^2_e	0.481	0.474,	0.489

$\sigma^2_{u_1}$ slope variance
$\sigma^2_{u_0}$ intercept variance
cov(time, intercept) covariance between the slope and intercept
σ^2_e residual measurement error variance

CrI: Credible Interval

The results for the first derivative association structure are given in Table 7.15. The hazard ratio for a unit increase in the current linear slope of log hCG was 0.007 95% CrI(0.0004, 0.138). This suggests an increase in the log hCG

gradient results in a large reduction in the rate of miscarriage at time t. Such a large effect must be interpreted with caution due to the limited sample size.

Examination of trace plots showed no evidence of non-convergence. Plots demonstrated high autocorrelation for a number of successive lags before reducing to acceptable levels, across a number of parameters. A longer chain length and greater thinning were used to address this issue. The density plots indicated normality.

7.5.6 Subject-specific survival predictions

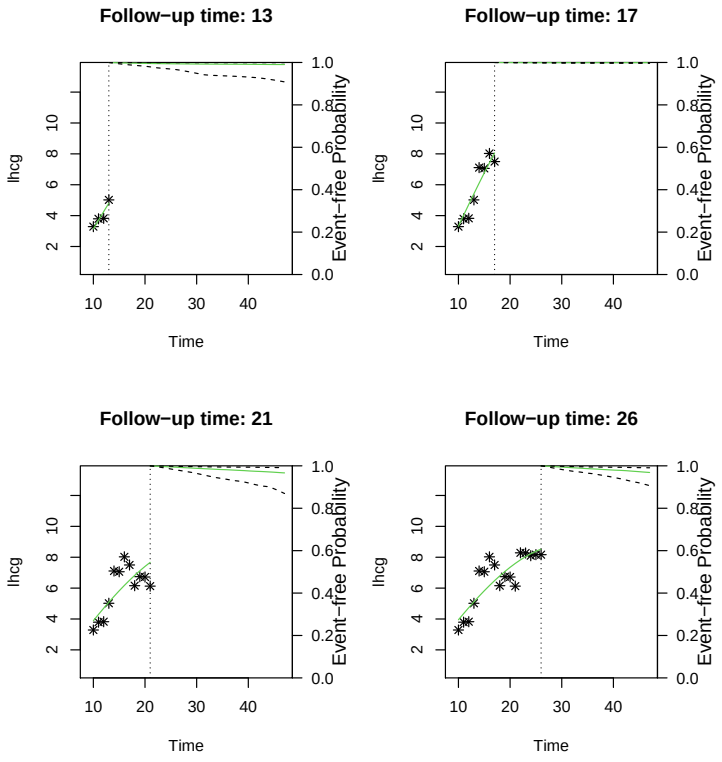


FIGURE 7.15. Dynamically updated conditional event-free (survival) probability predictions for a healthy pregnancy

Individualised predictions were obtained from the final first derivative association model. To illustrate the intended use of these survival predictions, they were plotted for three participants who experienced a viable pregnancy, early and later loss. Figure 7.15 presents dynamically updated conditional survival curves as longitudinal measurements are observed. Participant 58 was experiencing an ongoing viable pregnancy when censored at the end of follow up. They contributed a total of 18 measurements and were a non-smoker, aged 24 with an average cycle length of 26 days. The participant had experienced one miscarriage previously and was categorised as obese with a BMI of 30.17kg/m^2.

TABLE 7.16. Subject-specific survival probabilities for a viable pregnancy

log hCG observations up to	2-day Probability (95% CrI)			5-day Probability (95% CrI)			10-day Probability (95% CrI)		
Day 13	0.999	(0.995,	0.9998)	0.997	(0.987,	0.9995)	0.995	(0.975,	0.999)
Day 17	0.9998	(0.999,	0.9999)	0.9995	(0.998,	0.9999)	0.9993	(0.997,	0.9999)
Day 21	0.998	(0.995,	0.9993)	0.995	(0.987,	0.998)	0.987	(0.957,	0.996)
Day 26	0.997	(0.992,	0.999)	0.992	(0.977,	0.997)	0.989	(0.963,	0.995)

The survival predictions, updated after consecutive sets of 'new' observations, are presented for two, five and ten-day prediction intervals in Table 7.16. After the initial four log hCG observations up to day 13, a strong survival probability was indicated for the pregnancy even up to day 47 of 99.0% (0.990; 95% CrI: 0.908, 0.999). As time elapsed with no further measurements there was greater uncertainty indicated by the wider credible intervals beyond day 30. Updating the predictions by adding observations, both decreased the uncertainty in the survival probability curve, and also maintained the high survival probability owing to the large increases in the current value and slope of log hCG. Predictions updated again with observations up to day 22 and then again at day 27 reflect the less stable increase in log hCG. However, the survival probability again did not drop

below 90% in the 10 or 20 day prediction interval after the final measurement (0.986; 95% CrI: 0.964, 0.995 and 0.971; 95% CrI: 0.914, 0.991 respectively).

Predictions for an early loss pregnancy are given in Figure 7.16. Participant 262 was a 24 year old previous/current smoker with a BMI of 26.29kg/m^2, a cycle length of 29 days and having experienced one previous miscarriage. Overall log hCG values stayed consistently low for this pregnancy in comparison to participant 58's healthy pregnancy.

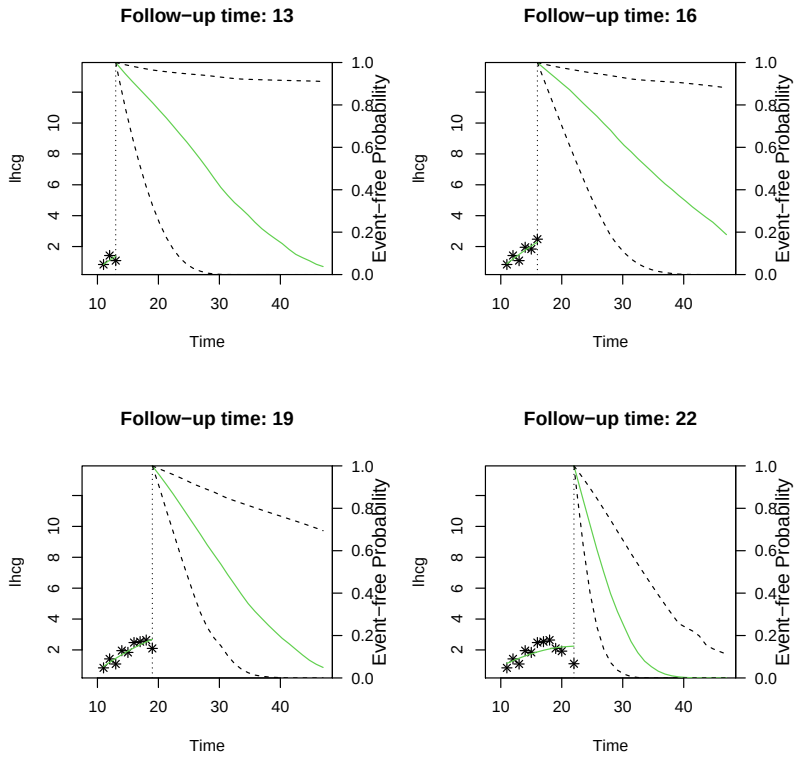


FIGURE 7.16. Dynamically updated conditional event-free (survival) probability predictions for an early loss ($\leq$ 6 weeks

Subject-specific conditional survival probabilities are presented in Table 7.17. With the contribution of the initial three log hCG measurements up to day 13, the survival probability was 84.5% (0.845; 95% CrI: 0.420, 0.970) at 18 days since conception, reducing to 66.3% (0.663; 95% CrI: 0.079, 0.950) by day 23. The

addition of several new measurements up to day 16 caused the survival prediction to decrease to 71.3% (0.713; 95% CrI: 0.244, 0.943) for a 10 day prediction interval. When all longitudinal observations up to day 22 were included, the five-day survival prediction was 44.9% (0.449; 95% CrI: 0.136, 0.796). Credible intervals were wide as the prediction window widened. This is in line with the decreasing log hCG observations, which eventually plateaued and decreased to below baseline levels.

TABLE 7.17. Subject-specific survival probabilities for an early loss

Log hCG observations up to	2-day	5-day	10-day
	Probability (95% CrI)	Probability (95% CrI)	Probability (95% CrI)
Day 13	0.946 (0.784, 0.989)	0.845 (0.420, 0.970)	0.663 (0.079, 0.950)
Day 16	0.952 (0.854, 0.987)	0.871 (0.608, 0.969)	0.713 (0.244, 0.943)
Day 19	0.913 (0.791, 0.973)	0.820 (0.587, 0.945)	0.582 (0.189, 0.876)
Day 22	0.838 (0.621, 0.941)	0.499 (0.136, 0.796)	0.119 (0.002, 0.537)

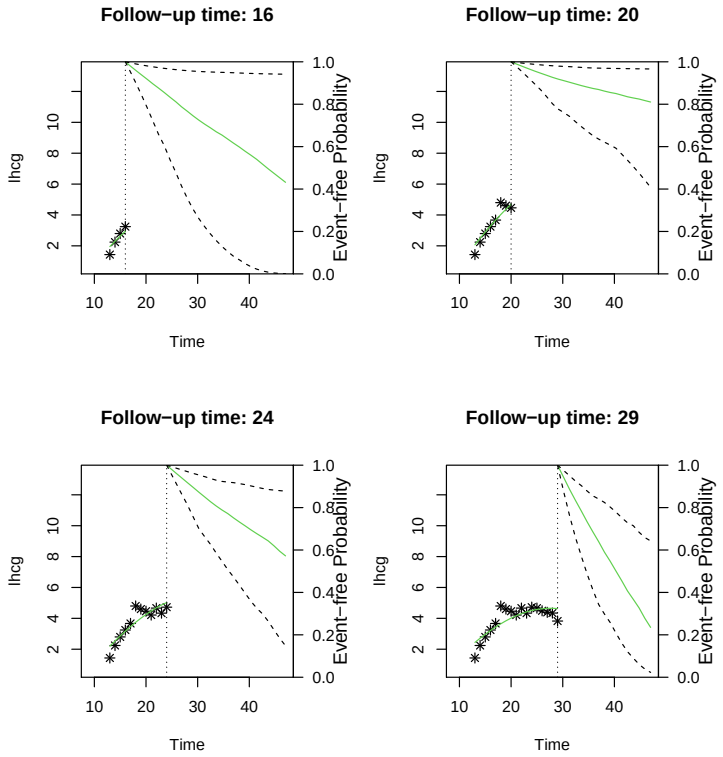


FIGURE 7.17. Dynamically updated conditional event-free (survival) probability predictions for a late loss (> 6 weeks)

As with the early loss, the log hCG observations did not increase as steeply or as rapidly for the later loss presented in Figure 7.17, when compared to the healthy pregnancy. Participant 1 was a 31 year old previous smoker with a BMI of 25.74 kg/m^2 and average cycle length of 29 days. They had also experienced one previous miscarriage. The individual contributed 17 log hCG observations in total.

Conditional survival predictions are shown in Table 7.18. The survival prediction accounting for measurements up to day 16 indicated a 89.7% (0.897; 95% CrI: 0.723, 0.977) survival probability 5 days after the last observed log hCG measurement and 81.2% (0.812; 95% CrI: 0.478, 0.963) for a 10 day prediction

window. Updating the predictions at day 20 post conception, the survival predictions improve due to the increases in log hCG. For a five-day prediction window the survival probability for the pregnancy was 95.6% (0.956; 95%: 0.894, 0.988), which decreased to 91.1% (0.911; 95% CrI: 0.760, 0.978) at 10 days post observed log hCG. However after updating at day 24 and day 29, log hCG plateaued and began to decrease. Pregnancy survival dependent on log hCG observations up to 24 days was 89.5% (0.895; 95% CrI: 0.757, 0.962) for a 5-day window, and 80.8% (0.808; 95% CrI: 0.586, 0.929) for a ten-day prediction window. Once survival predictions were updated with the final log hCG observations up to day 29, survival for a five and ten-day prediction window was 78.4% (0.784; 95% CrI: 0.571, 0.904) and 54.7% (0.547; 95% CrI: 0.258, 0.803).

TABLE 7.18. Subject-specific survival probabilities for a later loss

Log hCG observations up to	2-day	5-day	10-day
	Probability (95% CrI)	Probability (95% CrI)	Probability (95% CrI)
Day 16	0.961 (0.905, 0.990)	0.897 (0.723, 0.977)	0.812 (0.478, 0.963)
Day 20	0.976 (0.946, 0.993)	0.956 (0.894, 0.988)	0.911 (0.760, 0.978)
Day 24	0.963 (0.917, 0.987)	0.895 (0.757, 0.962)	0.808 (0.586, 0.929)
Day 29	0.884 (0.758, 0.949)	0.784 (0.571, 0.904)	0.547 (0.258, 0.803)

7.5.6.1 Prediction intervals

Subject-specific survival probabilities were predicted after arbitrary time intervals. However, there may be an optimal prediction interval particularly in the early stages of pregnancy when drastic changes in log hCG can occur quite suddenly.

Probabilities for a later loss were updated daily, every two, three, four and five days. The daily updated probabilities are shown in Figure 7.18. They exemplify the fact that the greatest changes in risk are observed early during the first five measurements. Once the trajectory is established as increasing, changes in survival probability are less pronounced and every added observation serves to

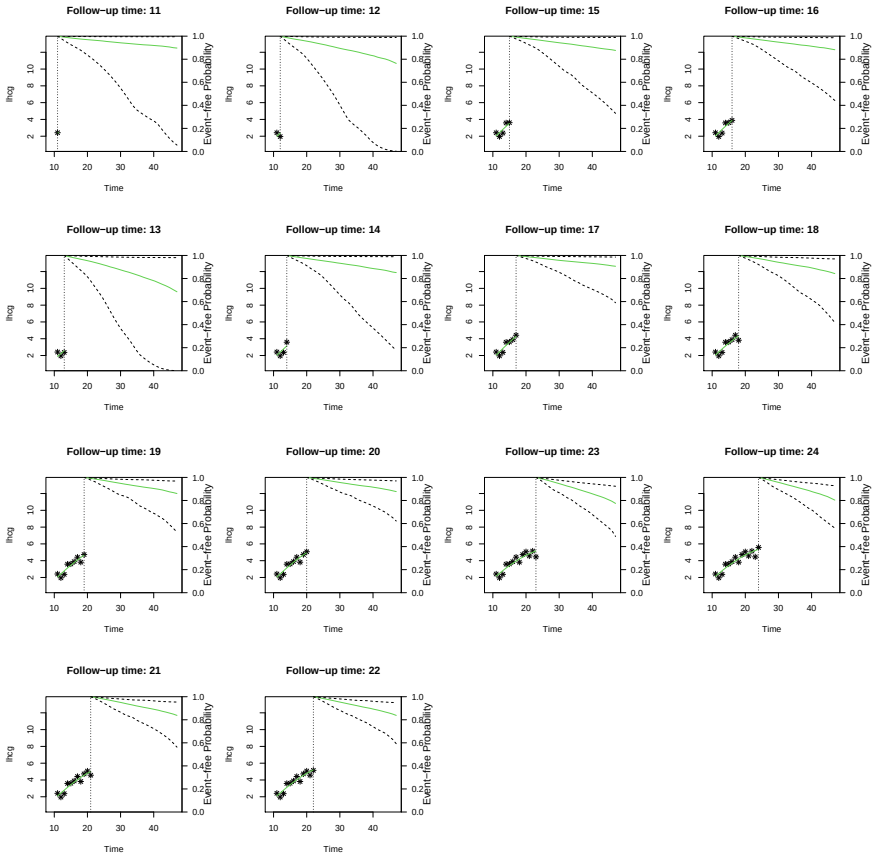


FIGURE 7.18. Subject-specific survival probabilities for a participant who experienced a later loss, updated daily

shore up estimates of uncertainty. However the reduction in log hCG at day 23 brings the survival curve down again. This is even with just a single added measurement.

Figure 7.19 shows probabilities which were updated every two days. Updating trajectories every two days seems to capture changes in pregnancy survival much more succinctly than the daily updates. Each update describes a significant change in survival probability, particularly at days 14 and 16, and also captures the decrease with the lower log hCG observation at day 22.

Figure 7.20 shows subject-specific conditional survival probabilities updated every three, four and five days. Predicting for an interval of 3 days continues to

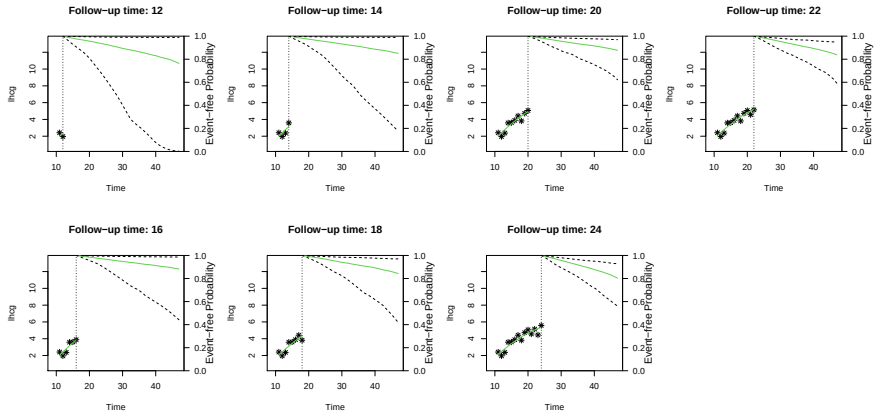


FIGURE 7.19. Subject-specific survival probabilities for a participant who experienced a later loss, updated every two days

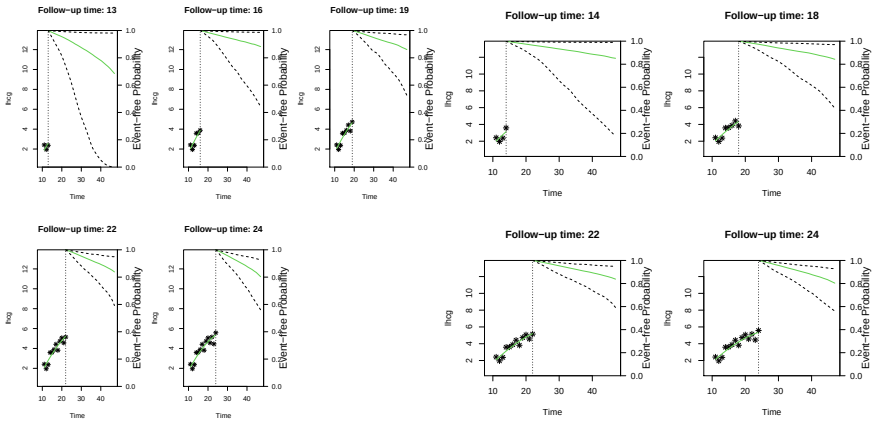
capture key changes in survival probability, however when the interval is widened to four days the survival probability remained constant at close to 0.7. Further widening the window to five days gave similar results, with less fluctuation in survival probability for this individual pregnancy.

7.5.7 Longitudinal predictions

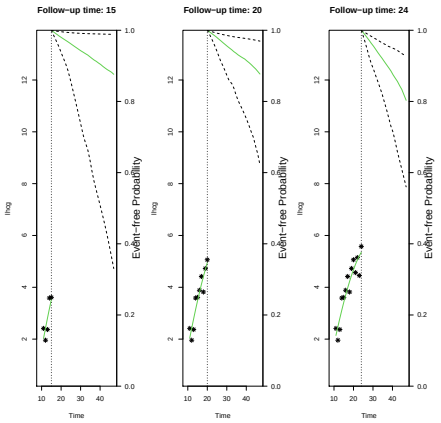
Longitudinal predictions, updated after each observation, were calculated and plotted for participants who experienced healthy pregnancies, early and later losses.

Figure 7.21 illustrates daily updated longitudinal predictions for a later loss. As the trajectory was updated, the uncertainty decreased. The final prediction of the log hCG trajectory from day 29 up to day 50 indicated projected decrease in log hCG to below detection limit levels.

Predictions for the healthy pregnancy are shown in Figure 7.22. Predictions across all follow-up points tended to plateau before decreasing as predictions were extrapolated further.



(A) Updated every three days (B) Updated every four days



(C) Updated every five days

FIGURE 7.20. Subject-specific survival probabilities for a participant who experienced a later loss

Figure 7.23 shows the longitudinal predictions by follow-up point for a participant who experienced an early loss. Log hCG trajectories decreased over time as they were updated, similar to the later loss. However, over time trajectories decreased much further than the later loss and at a steeper rate.

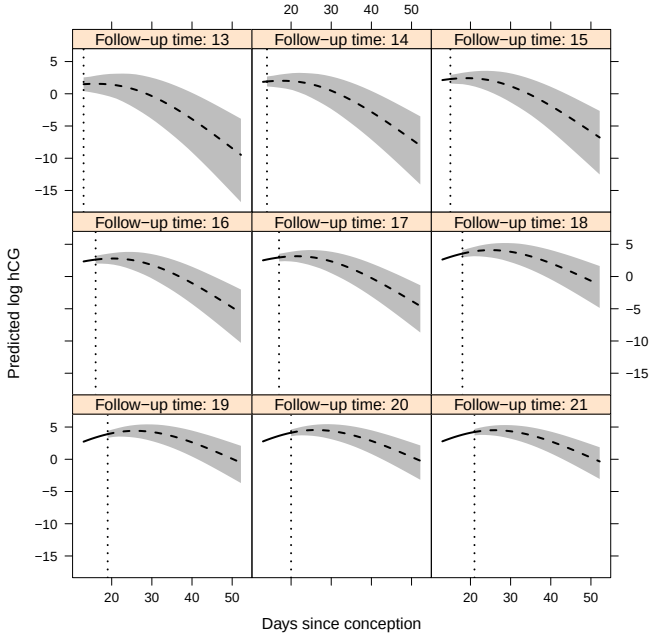


FIGURE 7.21. Daily longitudinal predictions for participant who experienced a later loss

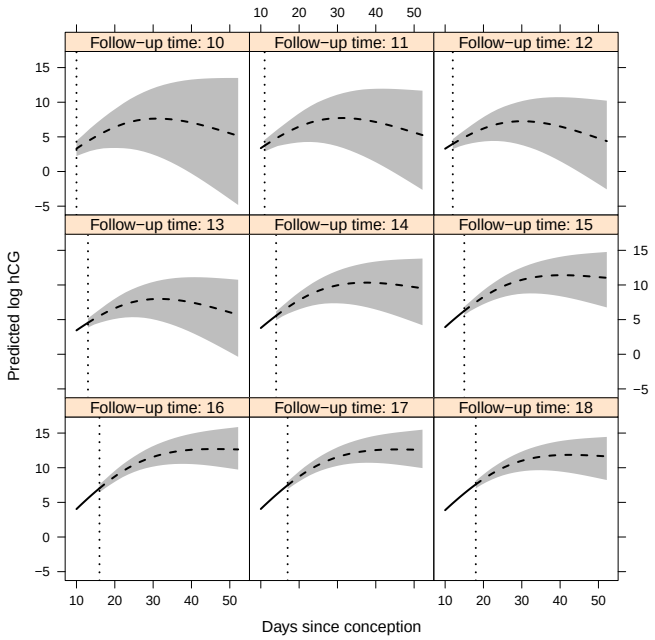


FIGURE 7.22. Daily longitudinal predictions for participant who experienced a healthy pregnancy

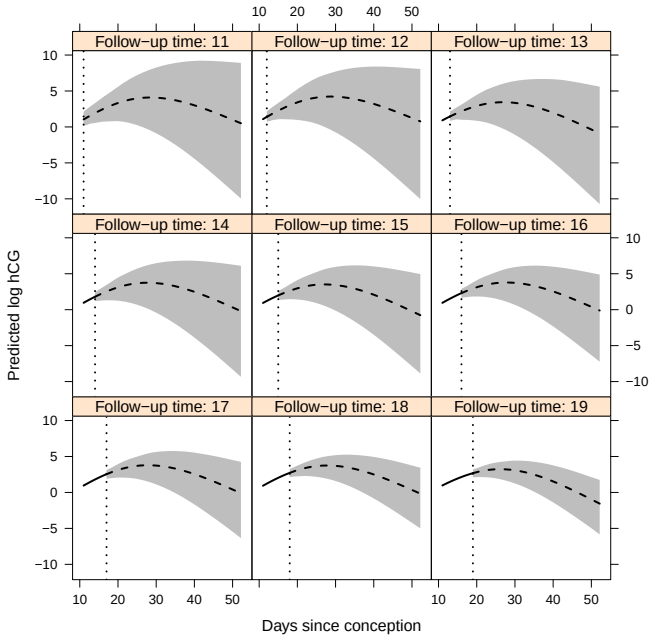


FIGURE 7.23. Daily longitudinal predictions for participant who experienced an early loss

7.5.8 Discrimination

The sensitivity and specificity of the fitted first derivative association model was assessed at different time-points during follow-up. Discrimination statistics for five and ten-day prediction windows are presented in Table 7.19.

ROC curves were plotted for various prediction windows and are presented in Figure 7.24. Twenty-nine women experienced a miscarriage within the prediction window between 10 and 20 days post conception. The optimal threshold was selected based on both the F score and Youden index. The optimal threshold based on the F Score was a subject-specific survival probability of 0.79, which corresponded to a sensitivity of 0.488 and a specificity of 0.970. When selecting the threshold based on Youden's index the optimal cut-off was a subject-specific survival probability of 0.95, which gave a sensitivity of 0.990 and a specificity

of 0.637. The overall ROC AUC was 0.895, so that the probability of a randomly selected person who experienced a miscarriage between 10 and 20 days since conception, had a lower survival probability than a participant who did not experience an event in the time period was 89.5%.

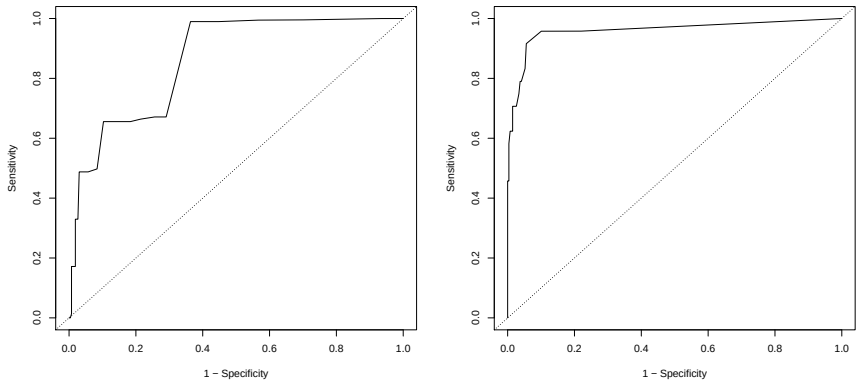
The ROC curve for the prediction window of 20 to 30 days post conception is presented in Figure 7.24b. The number of individuals at risk 20 days after conception was 252, with 27 events during this window. The overall ROC AUC was 0.983. The optimal threshold for the subject-specific survival probability, as indicated by the F score, was 0.76. The proportion of correctly identified miscarriages was 0.706 and the proportion of correctly specified ongoing pregnancies in this time window was 0.996. The Youden index, gave the optimal threshold of 0.96. The sensitivity for this cut-off was 0.957 and specificity was 0.938, an improvement from the previous ten-day window.

At day 30 there were 257 individuals still at risk with seven miscarriages observed between day 30 and 40. The ROC curve for this window can be viewed in Figure 7.24c. The optimal cut-off for the subject-specific survival probability as estimated by the F score and Youden index was lower for this prediction window at 0.79. The sensitivity for this cut-off was 0.731 and the specificity 0.992. Unfortunately the AUC could not be calculated for this prediction window due to a software error. However the figure shows a ROC AUC of very close to one.

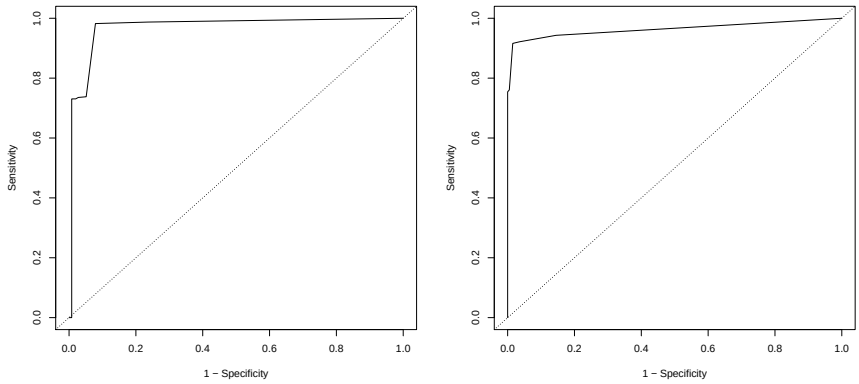
In the ten-day prediction window beginning at 40 days post conception, there were 198 individuals still at risk. Six individuals experienced an event during this period. The ROC curve for this window can be viewed in Figure 7.24d. The optimal threshold for the subject-specific survival probability corresponding to the maximum F score increased to 0.95 from the last ten-day window. The sensitivity for this cut-off was 0.754 and the specificity 1.000. The optimal cut-off

TABLE 7.19. Discrimination estimates for the first derivative association model

Start of prediction window, day	End of prediction window, days	At risk	Optimal cut-off		Sensitivity/Specificity		AUC
			F score	Youden index	F score	Youden index	
10	15	267	0.89	0.97	0.486/0.951	0.989/0.615	0.895
10	20	267	0.79	0.95	0.488/0.970	0.990/0.637	0.895
15	20	311	0.75	0.95	0.499/0.987	0.749/0.933	0.949
20	25	296	0.83	0.98	0.687/0.982	0.937/0.907	-
20	30	292	0.76	0.96	0.706/0.996	0.957/0.938	0.983
25	30	275	0.89	0.89	0.998/0.993	0.998/0.993	0.997
30	35	257	0.89	0.99	0.497/0.980	0.994/0.894	-
30	40	257	0.79	0.79	0.731/0.992	0.731/0.992	-
35	40	242	0.65	0.65	0.977/0.992	0.977/0.992	-
40	45	198	0.98	0.99	0.819/0.995	0.982/0.974	-
40	50	198	0.95	0.97	0.754/1.000	0.916/0.985	-
45	50	67	0.98	0.99	0.088/0.985	0.366/0.897	-



(A) 10 to 20 days since conception (B) 20 to 30 days since conception



(C) 30 to 40 days since conception (D) 40 to 50 days since conception

FIGURE 7.24. Receiver operating characteristic (ROC) curves for various prediction windows

maximising Youden's index was 0.97, with a sensitivity of 0.916 and specificity of 0.985. Again the AUC could not be calculated for this prediction window.

Overall sensitivity estimates were more variable than specificity. The F score tended to signal a lower threshold for survival probability and as a result indicated a lower sensitivity, whereas the Youden index gave higher thresholds for a larger sensitivity. The sensitivity was generally lower for the larger ten-day windows than the five-days predictions windows.

7.5.9 Calibration

Estimates for the prediction error both for a square and absolute loss function are given in Table 7.20. The difference in the observed and expected rates in the data were all below 0.07. The highest estimates for the error were given for 5 and 10 ten day windows from day 10 to day 30. The lowest error was for the 35 to 40 day interval with a prediction error of 0.0073 for the square and 0.0141 for the absolute loss function estimates. The low scores across the board indicated accurate predictions of event rates.

TABLE 7.20. Prediction error estimates

Start of prediction window, days	End of prediction window, days	Prediction error	
		Square	Absolute
10	20	0.0205	0.0603
15	20	0.0287	0.0499
20	25	0.0353	0.0531
20	30	0.0345	0.0596
25	30	0.0130	0.0257
30	35	0.0095	0.0178
30	40	0.0147	0.0260
35	40	0.0073	0.0141
40	45	0.0188	0.0229
40	50	0.0186	0.0265

7.5.10 Multivariate joint longitudinal-survival model

Progesterone has also been named as a biomarker of interest for the prediction of pregnancy outcomes [276]. Including just one observation per individual, as the P3G level on the day of the implantation variable did, wasted valuable information. The viability of utilising log P3G as an additional biomarker in a multivariate model was investigated. When added to a survival model as a time-varying covariate, a unit increase in log P3G resulted in a 42% decrease (HR: 0.581, 95% CI: 0.426, 0.794) in the rate of miscarriage. This indicated evidence of an association between time-to-miscarriage and log P3G. Even when included alongside log hCG in the time-varying covariate model, a unit increase in log P3G still indicated a 35% decrease (HR: 0.652, 95% CI: 0.471, 0.903) in the rate of miscarriage.

As progesterone is detectable prior to conception, using the time since conception timeline caused the loss of observations prior to conception. Women who had viable pregnancies contributed 18.1 (SD: 3.25) log P3G measurements on average, whilst for those who miscarried a mean of 16.4 (SD: 3.50) measurements

were observed. Moving to the time since conception timeline resulted in the loss of 222 P3G observations overall. As interest lies in observations during pregnancy this does not pose a problem. After exclusion, on average women in the viable pregnancy group had 17.4 (SD: 3.17) P3G records and women in the failing pregnancy group had 15.8 (SD: 3.27).

7.5.10.1 Longitudinal model

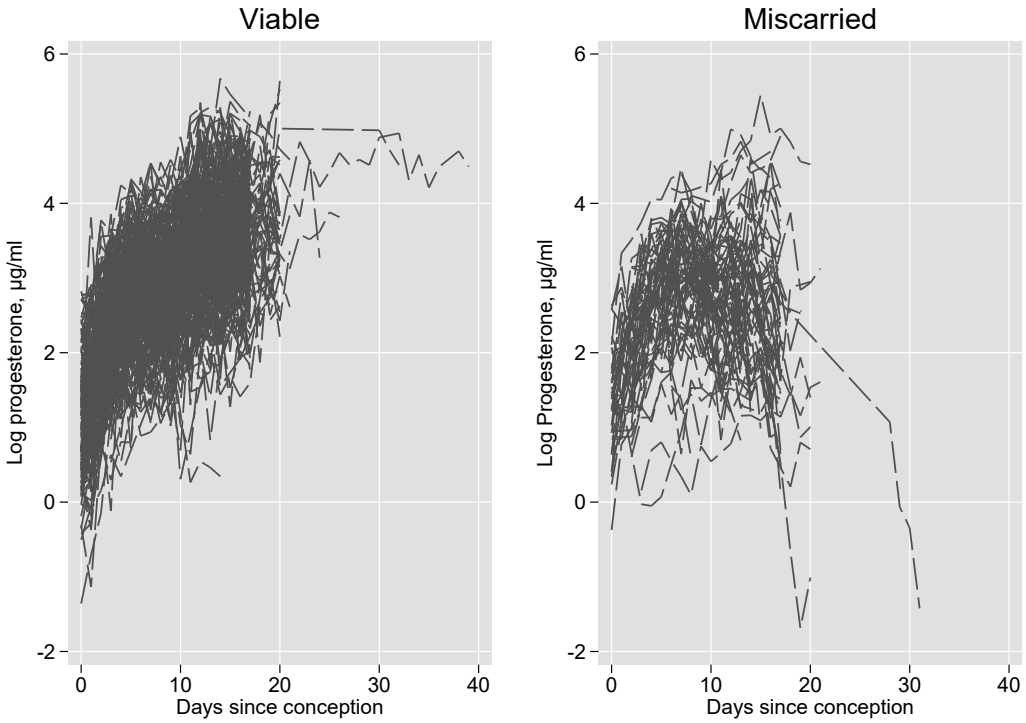


FIGURE 7.25. Log progesterone trajectories by pregnancy viability group

Log P3G trajectories by viability group can be seen in Figure 7.25. There was a general increasing trend across trajectories for women who experienced viable pregnancies. For women who miscarried there was an initial increase at much the same rate as the viable pregnancies. However, most women did not reach the

height of log P3G experienced by the women with healthy pregnancies, before reducing down close to levels at time zero.

The shape of log P3G was modelled using both polynomials and restricted cubic splines. Model fit statistics for each model are given in Table 7.21. The best fitting model with an AIC of 14019.54 utilised three splines with internal knots at 0, 6, 12 and 30 days post conception.

TABLE 7.21. Model fit estimates for the fixed longitudinal log P3G trajectory

Model	Log Likelihood	df	AIC	BIC
Linear model	-7263.107	3	14532.21	14552.28
Quadratic model	-7642.616	3	15291.23	15311.3
Cubic model	-7930.389	3	15866.78	15886.85
2 RCS model	-7087.392	4	14182.78	14209.54
3 RCS model	-7004.77	5	14019.54	14052.99
df: degrees of freedom				

The modelled trajectory can be seen in Figure 7.26. The final model included both a random intercept and a non linear random slope modelled with two splines. This provided the best fit according to the AIC. All random effects explored are shown in Table 7.22.

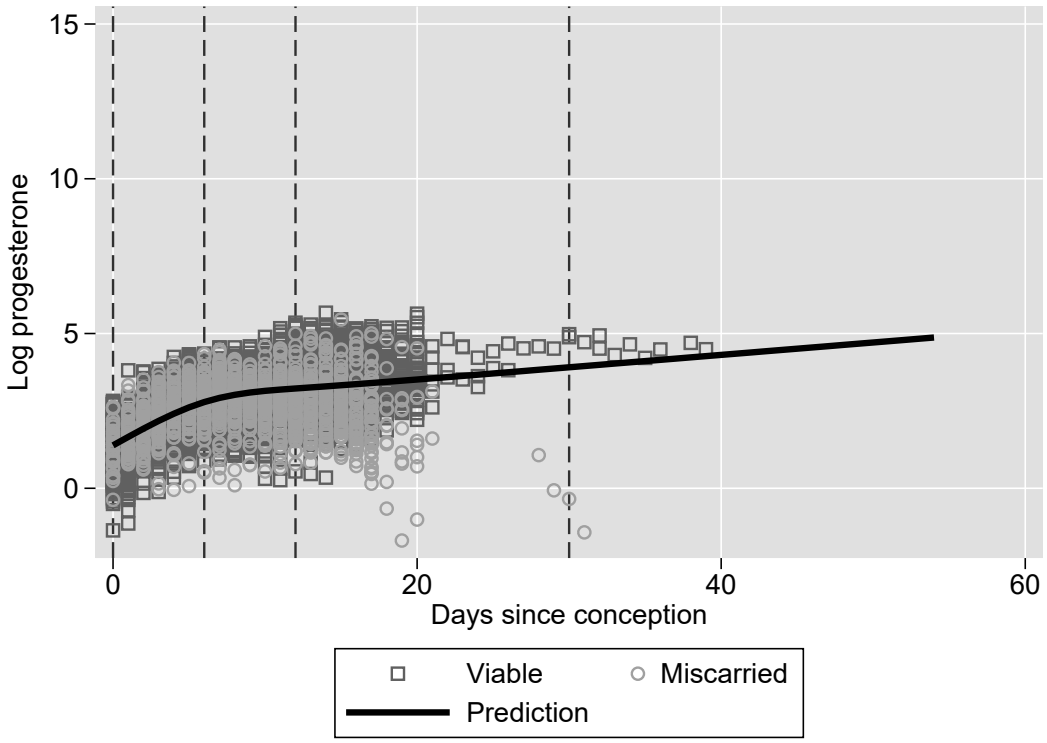


FIGURE 7.26. Log progesterone trajectory modelled using three restricted cubic splines

TABLE 7.22. Model fit statistics for the random effects of the longitudinal P3G trajectory

Model	Log likelihood	df	AIC	BIC
Random intercept	-5040.658	6	10093.32	10133.45
Random linear slope	-4489.088	8	8994.175	9047.688
Random two RCS slope	-4438.216	8	8892.433	8945.946
Random three RCS slope	-4459.224	8	8934.447	8987.96
df: degrees of freedom				

The final model also included a BMI term as evidence in the literature has linked low P3G levels in pregnancy to obesity [277]. An LMM with an added grouping variable demonstrated an association between viability and log P3G (see Table 7.23). On average a woman who experienced an early loss before 6

weeks had a -0.247 (95% CI: -0.401, -0.094) lower log P3G level than a woman who experienced a viable pregnancy. On the other hand, for women experiencing a later loss an increase of log P3G of 0.093 (95% CI: -0.187, 0.373) was observed on average, when compared with the viable pregnancies. This was not statistically significant.

TABLE 7.23. Linear mixed effect model estimates for log P3G

Variable	Mean Change	95% CI
RCS 1 time, days	0.274	0.260, 0.289
RCS 2 time, days	0.002	0.002, 0.003
RCS 3 time, days	-0.001	-0.001, -0.001
Centred BMI, kg/m^2	-0.015	-0.025, -0.005
Viability group		
Viable	Reference	
Early loss (<= 6 weeks)	-0.247	-0.401, -0.094
Later loss (>6 weeks)	0.093	-0.187, 0.373
Intercept	1.463134	1.376, 1.551
CI: Confidence Interval		

The estimates for the longitudinal trajectory model for log P3G (without grouping variable) are given in Table 7.24. An association between BMI and P3G was observed. A one kg/m^2 increase in BMI from 25kg/m^2 resulted in a statistically significant corresponding -0.016 (95% CI: -0.026, -0.006) decrease in log progesterone.

TABLE 7.24. Linear mixed effects model estimates for log P3G

Variable	Mean change		95% CI
RCS 1 time, days	0.274	0.260,	0.288
RCS 2 time, days	0.002	0.002,	0.003
RC3 3 time, days	-0.001	-0.001,	-0.001
Centred BMI, kg/m^2	-0.016	-0.026,	-0.006
Intercept	1.426	1.344,	1.507
Random effect parameters	Estimate		95% CI
$\sigma^2_{u_2}$	0.0000002	0.0000002	0.0000003
$\sigma^2_{u_1}$	0.006	0.004	0.008
$\sigma^2_{u_0}$	0.410	0.338	0.497
cov(linear time, non-linear time)	0.00003	0.00002	0.00003
cov(linear time, intercept)	-0.024	-0.033	-0.014
cov(non-linear time, intercept)	-0.0001	-0.0001	-0.00003
σ^2_e	0.171	0.165	0.178

$\sigma^2_{u_1}$ $\sigma^2_{u_2}$: slope variances
$\sigma^2_{u_0}$: intercept variance
cov(,): covariances between pairs of variances
σ^2_e: measurement error variance
CI: Confidence Interval

The survival submodel remained the same, apart from the omission of the singular P3G variable which was previously included in the log hCG survival submodel.

7.5.10.2 P3G joint longitudinal-survival model

A joint longitudinal-survival model was fitted for log P3G with both a current value and first derivative association structure. The results for each model are presented in Tables 7.25 and 7.26. A unit increase in the absolute value of log P3G corresponded to a 56.9% (HR: 0.431; 95% CrI:0.326, 0.569) decrease in the rate of miscarriage at time t. A unit increase in FSH3 on the day of implantation resulted in a 23.8% (HR: 1.238 95% CrI: 1.145, 1.336) increase in the rate of miscarriage at time t. The log P3G slope was not significantly associated with time to miscarriage with a hazard ratio of 0.092 (95% CrI: 0.006, 1.073). The

expected current value of the slope continued to be associated with miscarriage. A unit increase in the absolute value of log P3G corresponded to a 51.6% (HR: 0.484; 95% CrI: 0.347, 0.642) reduction in miscarriage rate, indicating the absolute increases in log P3G are more important than increases in the slope in reducing the rate of miscarriage.

TABLE 7.25. Joint longitudinal-survival model estimates for log P3G with current value association structure

Survival submodel	Hazard ratio	95% CrI	
Current value of log P3G	0.431	0.326,	0.569
Centered age, years	1.013	0.946,	1.062
Quadratic centered age, years	1.007	1.001,	1.013
Centered BMI, kg/m^2	0.987	0.936,	1.040
Smoking			
Never	Reference		
Current/previous	0.710	0.635,	2.125
Number of previous miscarriages	1.090	0.838,	1.408
FSH3 on day of implantation, mIU/ml	1.238	1.145,	1.336
Longitudinal submodel	**Mean Change**		**95% CrI**
RCS 1 time, days	1.552	1.370,	1.717
RCS 2 time, days	3.757	2.530	5.185
RCS 3 time, days	0.199	-1.473,	2.063
Centred BMI, kg/m2	-0.018	-0.031,	-0.008
Intercept	1.1098	0.722,	1.475
Random effect parameters	**Estimate**		**95% CrI**
$\sigma^2_{u_2}$	36.652	25.816,	49.856
$\sigma^2_{u_1}$	7.297	5.446,	9.415
$\sigma^2_{u_0}$	1.036	0.609,	1.490
cov(linear time, intercept)	-0.165	-0.887,	0.544
cov(non-linear time, intercept)	3.209	1.447,	5.229
cov(non-linear time, linear time)	11.258	7.924,	15.332
σ_e	0.372	0.362,	0.384

$\sigma^2_{u_1}$ $\sigma^2_{u_2}$: slope variances
$\sigma^2_{u_0}$: intercept variance
cov(,): covariance between pairs of variance components
σ_e: measurement error SD
CrI: Credible Interval

TABLE 7.26. Joint longitudinal-survival model estimates for log P3G with first derivative slope association structure

Survival submodel	Log hazard ratio	95% CrI
Current value of log P3G value	0.484	0.347, 0.642
Slope of log P3G slope	0.092	0.006, 1.073
Centred age, years	0.937	0.942, 1.057
Quadratic centred age, years	1.007	1.001, 1.014
Centred BMI, kg/m^2	0.919	0.934, 1.042
Smoking status		
Never	Reference	
Previous/current	0.922	0.690, 2.548
Number of previous miscarriages	1.027	0.882, 1.459
FSH3 on day of implantation, mIU/ml	1.228	1.125, 1.336
Longitudinal submodel	**Mean Change**	**95% CrI**
RCS 1 time, days	1.6160	1.3866, 1.8440
RCS 2 time, days	3.9403	2.6177, 5.5342
RCS 3 time, days	0.558	-1.1384, 2.4309
Centered BMI, kg/m2	-0.019	-0.031, -0.006
Intercept	1.134	0.581, 1.577
Random effect parameters	**Estimate**	**95% CrI**
$\sigma^2_{u_2}$	33.314	17.856, 46.581
$\sigma^2_{u_1}$	7.884	5.594, 10.688
$\sigma^2_{u_0}$	0.861	0.408, 1.321
cov(linear time, intercept)	-0.188	-1.246, 0.746
cov(non-linear time, intercept)	2.200	-0.740, 4.393
cov(non-linear time, linear time)	12.312	9.202, 16.128
σ_e	0.374	0.363, 0.386

$\sigma^2_{u_1}$ $\sigma^2_{u_2}$: slope variances
$\sigma^2_{u_0}$: intercept variances
cov(,): covariance between pairs of variance components
σ_e: measurement error SD
CrI: Credible Interval

7.5.10.3 Multivariate joint longitudinal-survival model

The multivariate joint longitudinal-survival model was fitted with a longitudinal submodel each for log hCG and log P3G. Due to the restrictions of the package, the multivariate model could only be fitted for time points at which both hCG and P3G were observed. As hCG was measured for longer than P3G, this resulted in the loss of longer-term hCG follow-up data. Additionally as hCG is detectable only once implantation has occurred, P3G data was lost during the time at which hCG was still under 2mIu/ml. The dataset as a result after conclusions contained 258 viable and 52 miscarried pregnancies. Women across both groups had on

average similar numbers of hCG and P3G observations (8.97 ± 2.26 and 7.08 ± 2.06 for viable and failing groups respectively). Figure 7.27 shows the overlaid hCG and P3G curves by viability group.

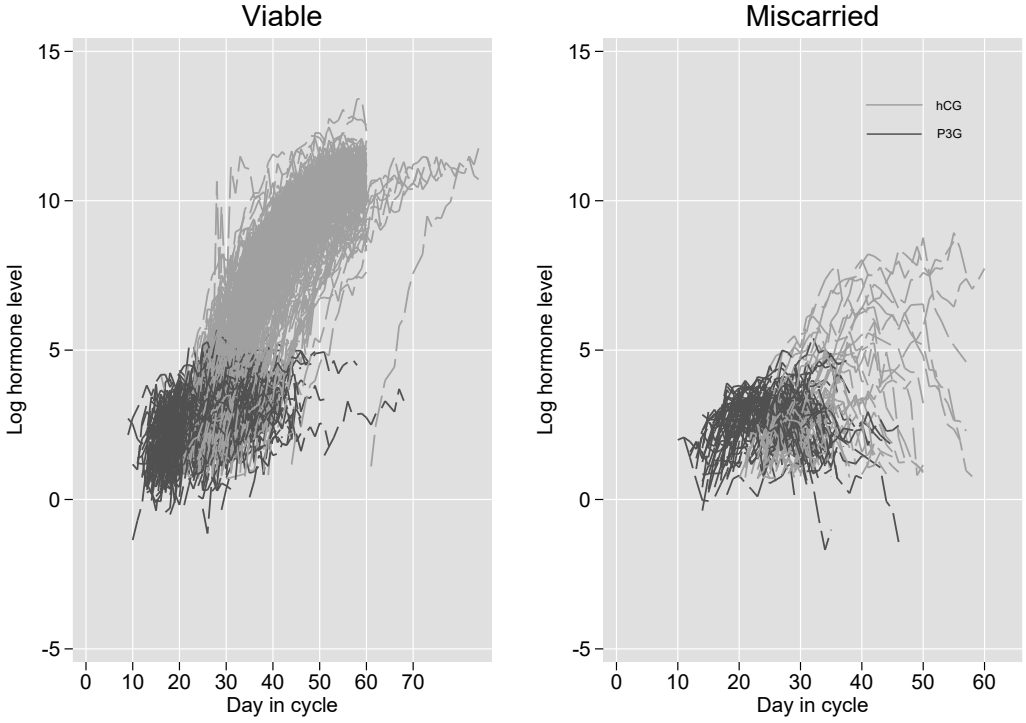


FIGURE 7.27. Log P3G and log hCG observations pregnancy viability group

The final model estimates for the multivariate joint model are presented in Table 7.27. When included as an additional longitudinal submodel, the association between change in the current value of log P3G and the rate of miscarriage changed direction, with increases suggesting a harmful effect. In particular a unit increase in the absolute value of log P3G resulted in a 2.24 times (HR: 2.236; 95% CrI: 1.445, 3.610) increase in the rate of miscarriage at time t. However a unit increase in the absolute value of log hCG resulted in a statistically significant reduction in the rate of miscarriage of 67.4% (HR: 0.326, 95% CrI: 0.232, 0.444).

Increases in FSH3 levels on the day of implantation continued to be significantly associated with miscarriage. A one mIU/ml increase resulted in a 25.9% (HR: 1.259; 95% CrI: 1.159, 1.365) increase in the rate of miscarriage.

TABLE 7.27. Multivariate joint longitudinal-survival model estimates for the current value association of log hCG and log P3G

Survival submodel	Hazard Ratio	95% CrIl	
Current value of log hCG	0.326	0.232,	0.444
Current value of log P3G	2.236	1.445,	3.610
Centred age, years	0.989	0.938,	1.044
Quadratic centred age, years	1.005	0.998,	1.011
Centred BMI, kg/m^2	0.961	0.908,	1.014
Smoking status			
Never	Reference		
Previous/current	0.970	0.520,	1.739
Number of previous miscarriages	1.167	0.896,	1.484
FSH3 on day of implantation, mIU/ml	1.259	1.159,	1.365
Longitudinal submodel for log hCG	Mean change	95% CrI	
RCS 1 time, days	19.925	19.240,	20.629
RCS 2 time. days	7.965	7.015,	8.892
Centred cycle length, days	-0.008	-0.032,	0.016
Centred BMI, kg/m^2	-0.020	-0.036,	-0.004
Intercept	-3.951	-4.237,	-3.674
Longitudinal submodel for log P3G	Mean change	95% CrI	
RCS 1 time, days	1.640	1.448,	1.849
RCS 2 time, days	4.061	2.813,	5.188
RCS 3 time, days	0.705	-0.729,	2.175
Centred BMI, kg/m2	-0.011	-0.023,	-0.001
Intercept	1.091	0.704,	1.494
CrI: Credible Interval			

7.6 Discussion

This analysis builds on the previous application of the joint longitudinal-survival model to the pregnancy setting in Chapter 6, with the availability of a larger dataset and extended follow-up. Both the change in current value and slope of log hCG were important in relation to the rate of miscarriage. Subject-specific

survival probabilities were shown to be sensitive to the change in slope of log hCG and demonstrated the importance of updating predictions with current biomarker observations. Furthermore, discrimination and calibration measures emphasized the model's capability of differentiating between viable and non-viable pregnancies.

7.6.1 Main findings

The joint modelling analysis reiterated that changes in the current value of longitudinal hCG are highly associated with time-to-miscarriage. Specifically, a unit increase in log hCG was shown to reduce the rate of miscarriage by 53%. Addition of the slope association demonstrated the slope of log hCG is just as important in predicting the rate of miscarriage as the absolute log hCG value. The hazard ratio for the slope association was very close to zero, suggesting that when holding the current value constant, a unit increase in the linear slope of log hCG reduced the rate of miscarriage by almost 100%. This is a huge effect and so must be interpreted with caution, particularly with the relatively small sample of data. In real terms this suggests that for two women with the same current value of log hCG, an increase in hCG accompanied by an increase in the slope of log hCG for one woman will increase the probability of pregnancy survival, compared to a woman who experiences an increase in log hCG, but no corresponding increase or even a decrease in slope. This finding was generally reflected in visualisations of log hCG trajectories. Miscarried profiles were characterised by increases in log hCG, at a slower rate and a failure to achieve the height of the log hCG trajectory of a viable pregnancy as shown in 7.3. This is reflected in the literature.[14; 278]

P3G was no longer significantly associated with miscarriage at the 5% level when included in the joint model, however the detrimental effect of rising FSH3 was maintained in the joint model setting. This corresponds with the uncertainty

around progesterone as a predictor for miscarriage. P3G varies between women and is influenced by maternal factors such as age and BMI [277]. Utilising a single observation of P3G in the model does not reflect progesterone changes throughout early pregnancy, and the choice of implantation as the time of inclusion may be too early to witness significant changes in P3G. A meta-analysis investigating progesterone as a diagnostic marker for early pregnancy loss found that a single serum progesterone value was able to distinguish between a viable and non-viable pregnancy for women presenting with pain or bleeding and an inconclusive ultrasound result. [279] Crucially, the omission of an ultrasound scan result reduced the marker's ability to delineate between failing and healthy pregnancies.

FSH3 is the hormone which triggers ovulation and is the highest just prior to ovulation. If conception is successful, FSH3 should reduce to close to a zero level. It has been hypothesised that elevated levels of FSH3 could be linked to reproductive ageing, and certainly an indication of menopause is an abnormal increase in FSH3 levels. The average age of women in the data was 30 years old and only 53 (14.4%) of women were aged over 35 years. Interactions between age and FSH3 were not found to be statistically significant. An alternative hypothesis is that an elevation in FSH3 may represent an early indication of impending loss, hence the body readying itself for the next cycle and the next egg release. Research into FSH3 focuses on fertility treatment rather than its role in early miscarriage. Indeed, the early losses exhibited the highest levels of FSH3 on the day of implantation with a mean of 3.96 (SD: 3.49), however later losses presented with similar FSH3 values to the viable pregnancies (1.85 ± 1.83 and 2.27 ± 1.36 mIU/ml respectively). Again, only a single value of FSH3 was utilised in the model at a specific time-point, although it must be noted that inclusion of FSH3 on the day of the missed period garnered similar results, as with P3G. This would

indicate that elevated FSH3 throughout the first week after implantation may be indicative of an impending loss.

Prediction of subject-specific survival probabilities exemplified the potential uses of the joint longitudinal-survival model in the pregnancy setting. Trajectories and probabilities between the viable, early and later loss were distinct particularly at the 5-day prediction window. Rates of miscarriage were predicted well as evidenced by the low prediction error across time intervals. However, the prediction of the true positives (sensitivity) was more variable, though the model tended to stably minimise the number of false positives.

The focus of this analysis was log hCG trajectories and their association with the time-to-miscarriage. However, longitudinal P3G was collected as part of the study although for variable lengths of time and for the most part for a shorter time than hCG. As a secondary analysis a joint model was fitted to log P3G and time-to-miscarriage. The longitudinal submodel included BMI as a predictor of log P3G, and as with log hCG higher values of BMI were associated with a reduction in log P3G. The results of the joint model demonstrated the importance of the increase in absolute values of log P3G to reduce the rate of miscarriage. A unit increase of log P3G indicated a 57% decrease in the rate of miscarriage at time t, however addition of the slope suggested that the slope of P3G was less important in relation to miscarriage. Progesterone analysis has mostly focused on single observations [48; 279]. Although increasingly random effects modelling of repeated measurements of P3G are finding that non-viable pregnancies are associated with decreases in progesterone as indicated by the joint model analysis[278].

Current guidelines do not advocate the use of progesterone observations with log hCG in a diagnostic capacity for early miscarriage, particular in the case of

ectopic pregnancies [17]. However, there has been increasing interest in utilising pregnancy markers together, and suggestions that there is value in utilising progesterone and hCG in combination [47; 276; 280]. The results from the multivariate joint model including log hCG and log P3G as longitudinally measured biomarkers, emphasized the association between changes in absolute values of log hCG and miscarriage. A unit increase in log hCG resulted in a 67% reduction in the rate of miscarriage. Whilst increases in P3G indicated a 2.2 times increase in miscarriage rate. This may be a symptom of fitting the multivariate joint model to a truncated dataset in order to allow complete data for all women. Certainly examination of slopes of log P3G at times where log hCG was also observed showed muted increases for P3G in comparison to hCG. The accepted variability of P3G between women may also be a factor in why changes indicated an increase in the miscarriage rate [281]. The modelling results suggest that P3G does not add any value compared with using hCG alone. A study comparing the ratio of hCG measurements 48 hours apart and a progesterone observation demonstrated the superiority of hCG in detecting a loss (sensitivity 75.6% for an 11% increase in hCG vs. 20.0% for a P3G level of 6.2ng/ml) [48]. This has also been seen in studies looking at the benefits of progesterone therapy for women who experience recurrent miscarriage. Although the therapy did do harm, there was not conclusive evidence of a benefit for all women, rather a select few women [22].

7.6.2 Strengths and limitations

The extensive collection of longitudinal hCG observations in this data allowed the exploration of the relationship of log hCG with early pregnancy loss up to the 9^{th} week of pregnancy. Commonly, focus of such studies hinges on the predictive power of serum biomarker measurements, in the majority of cases hCG

and progesterone [11]. This data is unique in both the serial nature of observations of hCG and the collection of a urinary biomarker as opposed to serum. Although hCG is not detectable post conception in urine as soon as in serum, evidence shows patterns of hCG, particularly intact hCG, are similar across both mediums[25; 282]. The urinary hCG biomarker then lends itself well to cost-effective predictive modelling of early miscarriage.

This data however was not without its idiosyncrasies. Inevitably women recruited could have been more likely to join a study with such an intensive sample collection protocol if indeed they had been trying to conceive unsuccessfully for a time. As recruitment was managed through online advertising, women would have had to be actively looking for an aid to conception. Obviously this was provided via Clearblue ovulation and fertility monitoring products. This suggests a possibly biased sample of women with suspected fertility problems. However, on examining the reason for withdrawal for women who did not continue on with the study, only 113 women in total cited irregular cycle, contravention of eligibility criteria (possible fertility issues) or intention to pursue fertility treatment as a reason. Furthermore though the prospect of daily sample collection may be considered daunting, only 27 women cited study stress or related concerns for their desire to stop taking part. Indeed, it has been shown that utilising digital ovulation tests as in this study, do not significantly increase stress levels when compared to women who did not use the aid to time intercourse. Rather testing increased women's awareness of their cycle and provided reassurance that intercourse timing would increase the likelihood of conception [256]. Most surprisingly of all the majority of women who were excluded (n=575) was a consequence of a pre-trial pregnancy, which negates concerns of a population struggling to conceive.

Joint longitudinal-survival models theoretically provide a natural framework for such modelling, however software capabilities particularly where the aim is to predict individual can prove to be barriers. It is due to limitations of commands in Stata, which can not yet provide predictions for complex non-linear longitudinal trajectories, that the analysis was moved to R. And it was due to the insurmountable obstacles encountered in fitting a joint model utilising maximum likelihood estimation, that ultimately the more flexible Bayesian approach was embraced. Although joint modelling tutorials have increasingly adopted the use of non-linear longitudinal trajectories for demonstrative purposes, delayed entry has been largely ignored. The complexities of a second layer of integration due to the addition of left truncation increases the possibility that the model will not converge, and when this is extended to the inclusion of multiple confounding variables and more complex association structures, adaptive quadrature can fail in the maximum likelihood approach [52]. Bayesian joint models therefore lend themselves to these situations of computational complexity. However, they need to be used carefully with appropriate assessment for evidence of non-convergence and sensitivity to model specification and estimation.

The appropriate specification of the shape of the longitudinal log hCG trajectory was a key element for this analysis. The submodel was carefully curated to best describe how log hCG behaves over time, and it is clear from the prediction of the fitted values in Figure 7.7 that the spline terms and the linear slope are a marked improvement on the quadratic trajectory model utilised in Chapter6 (see Figure 6.10). For this analysis variables related to hCG were also included in the longitudinal submodel. Increases in BMI were associated with a small but statistically significant reduction in log hCG. This is in line with evidence which has

found hCG is negatively associated with BMI [283]. This was especially relevant for this sample of women who collectively had a mean BMI just shy of 27kg/m^2.

Proportional hazards were assumed for the biomarker, which maintains that the hazard ratio for the biomarker effect is constant across follow-up time. Assessing whether this assumption holds in relation to the biomarker effect has not been addressed in the literature, likely due to the lack of software availability until now. The Stata program `merlin` can allow for time-biomarker interactions for non-proportional hazards. In the case of hCG, there may be points in time, such as around the time of implantation, that the hazard of miscarriage is much higher than once the pregnancy is established. The potential for non-proportional hazards should be considered in future analyses, particularly as assuming proportionality can lead to incorrect inferences of the data.

Markers were not standardised to allow a uniform interpretation across the board. This is something to consider to allow meaningful interpretation at perhaps the standard deviation change level rather than (relatively small) unit changes in P3G, FSH3 and hCG. In the case of hCG and similar reproductive markers the word standardisation is usually invoked when discussing the various assays which are utilised to measure the marker [284]. However, where hCG has been standardised, it is in relation to gestational age and expressed as a multiple of the median rather than the mean, i.e. how much an individual diagnostic test result deviates from the median [285; 286; 287].

The timing of observations was not investigated in this analysis. Log hCG observations were collected daily and so updates of the survival probability could be obtained after each daily observation. Of interest, however, would be whether measurements could be obtained less frequently, particularly if the survival probability indicated a low risk of imminent miscarriage. This extension would need

to pinpoint the acceptable cut-off for an increased lag between measurements and conversely when observations should once again be taken more frequently.

Sensitivity and specificity estimates were obtained without corresponding credible or confidence intervals. Two means of obtaining a measure of uncertainty involve bootstrapping and MCMC. Each gives a different interpretation of uncertainty and one is perhaps more relevant here than the other. Bootstrapping involves sampling from the data with replacement many hundreds or thousands of times to obtain n estimates of the parameter, here sensitivity and specificity. No underlying probability distribution is assumed. The standard error can be obtained by considering the distribution of the bootstrap estimates. This would give the uncertainty around the sampling of the data, i.e. how much does the observed data reflect the true population from which it was sampled [288]. The MCMC approach simulates draws from a joint posterior distribution of all parameters. The posterior probabilities are estimated based on defining the prior uncertainty around the parameters and updating this using the likelihood of the data, conditional on the hypothesis under investigation being true. The uncertainty of the parameter estimates is drawn from the simulations [289]. In this context it would be sensible to follow the MCMC simulation scheme approach detailed by Rizopoulos, which was also used to obtain standard error estimates for the subject-specific predictions [30]. The proposed scheme simulates parameter estimates assuming the sample is large enough to be approximated by the normal distribution, accounting for variability in the maximum likelihood estimates. Then plausible longitudinal histories up to time t are drawn. The final step simulates random effect estimates conditional on the longitudinal history. This allows quantification of parameter uncertainty which is more immediately relevant in ascertaining confidence in sensitivity and specificity estimates. Sample

uncertainty can be addressed at the validation stage. Bootstrapping however can be much simpler to implement and has been utilised by Kolamunnage-Dona and Kamarudin [290] in their package offering discrimination measures for joint models, particularly dynamic ROC curves. This package is currently only available on request, however.

The choice of cut-off was based on two different indices, the F score and the Youden index. The F score is calculated as double the product of precision and recall over the sum of precision and recall. Precision is equivalent to the sensitivity. Whilst the recall is the positive predictive value. This measure focuses on the sensitivity, almost ignoring specificity altogether [291]. In the case of miscarriage, it is important to correctly classify true viable pregnancies, as well as true failing pregnancies, which precludes the use of the F score. The Youden index opts for the cut-off which maximises the difference between the sensitivity and 1-specificity [292]. This is an intuitive index which aims to select the cut-point which gives the highest sensitivity and lowest 1-specificity. This is universally accepted as the best measure for cut-point selection, as it accounts for misclassification[293]. Another procedure is to take the maximum of the product of the sensitivity and specificity. This gives the probability of a randomly chosen viable pregnancy being below a cut-point and a failing pregnancy being above a cut-point, again accounting for misclassification [294].

The major limitation of this analysis is that the predictions for individuals were obtained from a model based on the very same data. This means predictions are not necessarily generalisable to another dataset. Validation of the joint model predictions is required by testing on a different dataset. In the absence of an external dataset suggested internal validation techniques include bootstrapping or cross-validation. Cross validation involves randomly splitting the dataset into

k number of smaller datasets. In turn one of the k datasets is held as a test dataset on which to evaluate the model on and the other $k-1$ groups of data are used to fit the chosen model. Generally $k = 10$ is the default number of data groups, therefore utilising 90% of the data for 'training' [295]. However the joint model was developed on a small sample of 346 women with complete data and this included only 64 miscarriages. Given the number of confounding variables included in the survival model, attempting to evenly distribute characteristics across ten groups may be problematic. Instead a five-fold cross validation may be more appropriate, utilising 80% of the data for model fitting. The joint model is then fitted to the training dataset and the predictive error and AUC calculated for each test dataset. The average of the estimates can be compared to the AUC and predictive error estimates fitted to the original dataset [202]. An alternative is to create a bootstrap sample and compare this with estimates of the AUC and predictive error from the original sample [296]. These methods have been implemented in tandem to validate a prediction model for pregnancy viability at the end of the first trimester utilising ultrasound results [297]. This particular study utilised two-thirds of the data for training leaving a third for testing.

The association between urinary log P3G and time-to-miscarriage was modelled utilising the joint model framework. Though increases in the marker suggested a decrease in miscarriage rate, data was inconsistently observed across individuals. This was emphasized when attempts were made to model the biomarker alongside log hCG, and increases in P3G indicated a detrimental effect on pregnancy outcome. This may be a result of the quality of the assay used to measure P3G which is more variable that that for hCG. Further investigation is required to understand whether these biomarkers can be monitored in in parallel, with fuller observation of the progesterone profiles as well as the hCG profiles.

7.6.3 Conclusion

This analysis provides a good grounding for jointly modelling a longitudinal biomarker and time-to-event outcome in the early pregnancy setting. Improvements were made in modelling both the longitudinal and survival submodel, thinking carefully about the shape of log hCG and potential confounders. Estimation of subject-specific predictions demonstrated the potential for dynamic monitoring in early pregnancy although more refining is required when thinking about the ideal timing of observations, as well as defining appropriate cut-offs for identifying failing and viable pregnancies.

The analysis confirmed hypotheses that the dependency between log hCG and miscarriage is defined by the slope of hCG. Yet, several new questions were raised in the undertaking of this analysis, one of which concerns FSH3 and its role in early pregnancy. The variable nature of P3G associations with miscarriage was also an interesting result. These findings provide potential avenues for future research, and serve as a primer for fitting joint models to real life pregnancy data.

Chapter 8

The timing of intercourse and miscarriage

8.1 Chapter overview

In this chapter intercourse diaries collected as part of the General Cycle Collection study will be analysed to assess the association between miscarriage and timing of intercourse. Several time-windows will be defined and analysed (i) to quantify the effect of intercourse on pregnancy outcome in key phases of the menstrual cycle and (ii) to assess whether utilising acts of intercourse in the fertile window as a proxy for sperm quality can be linked to pregnancy loss. In each case Cox proportional hazard models will be fitted, including the intercourse covariate and those variables associated with miscarriage as identified in Chapter 7. Model assumptions and fit were assessed and sensitivity analyses were conducted.

8.2 Introduction

Conception takes place if intercourse is timed within the fertile window. This is widely agreed to span the day of ovulation and the five days preceding, however wider windows up to seven days before and two days after ovulation have also been suggested [63; 298; 299; 300]. Various studies have attempted to quantify the probability of conceiving dependent on intercourse occurring on a day relative to

ovulation. Figure 8.1 presents single-day probabilities of conception for a selection of key studies [63; 299; 300; 301; 301; 302; 303; 304].

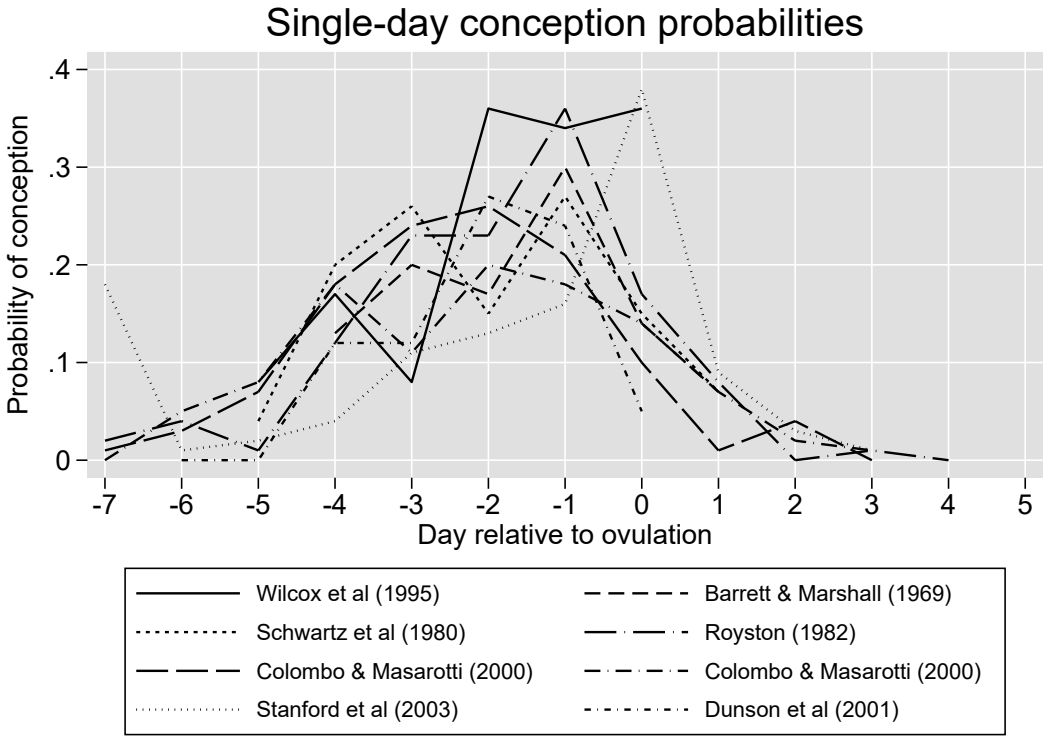


FIGURE 8.1. Single-day conception probabilities by study based on probabilities reported by [4] (Table 1)[4]

The greatest probability of conception has been attributed to the day prior to ovulation ranging from probabilities of 0.21 to 0.37 when based on basal body temperature, and 0.16 to 0.18 based on cervical secretions [4]. Slightly lower probabilities have been reported for the second and third day preceding ovulation [4]. More recently the advent of fertility aids, which allow monitoring of hormone levels to pinpoint the best time for intercourse, have allowed couples to target intercourse more effectively [305; 306]. Research emphasis remains on establishing the ideal window for conception, particularly as couples seek the most up to date information to maximise their chances of pregnancy. However, less has been

said about the consequences of intercourse during each phase of the cycle prior and during the period in which the pregnancy is established; and whether the timing of intercourse can affect the viability of a pregnancy, assuming conception is successful.

8.2.1 The luteal phase, peri-implantation and intercourse

The implantation window sits within the luteal phase. This spans the second half of the cycle from post-ovulation up to the day before the next cycle or bleed. During the luteal phase progesterone surges and the lining of the womb matures ready for implantation of the fertilised ovum [307]. This is the point in the cycle when pregnancy is established, so it is natural that women may be uncertain about having intercourse during this time. Current guidelines, however, do not advise against intercourse during pregnancy or implantation. This is due to lack of evidence of a detrimental relationship between timing of intercourse and miscarriage [308].

Steiner *et al.* [55] has suggested that intercourse during the peri-implantation phase (five to nine days post ovulation) reduces the chances of conception. It is suggested that implantation may be disrupted by contractions induced by intercourse and exacerbated by female orgasm [309]. A further supposition is that seminal fluid can illicit an inflammatory response from the female reproductive system, again interfering with implantation [310]. However, this raises the question of whether these are incidences of failed conception or rather early losses. If intercourse during implantation can prevent conception, can it also interrupt or interfere with implantation to the extent that a successful conception can end in miscarriage? The findings of the study have not been replicated since, with one study going as far as to refute claims that intercourse during the peri-implantation

window reduces the probability of conception [311]. It must be noted that the couples in the latter study, however were not all trying to conceive.

Guilt and emptiness have been shown to characterise a woman's emotions post-miscarriage [312]. Feelings of guilt and self-blame are usually borne from the idea that the loss could have been prevented, however unrealistic that belief may be [89]. It is conceivable that a woman may attribute a miscarriage to a seemingly ill-timed act of intercourse. Investigating the association between miscarriage and timing of intercourse during the peri-implantation window is an important analysis to shed light on whether intercourse should be avoided when the pregnancy is being established.

8.2.2 Age and quality of sperm

The timing of intercourse may provide information on the age and quality of the sperm. It is suggested that pregnancies resulting from sperm which has survived inside the female reproductive tract for longer before conception are more likely to lead to miscarriage [313]. Pregnancies resulting from intercourse two or more days prior to or post ovulation have been shown to be more likely to end in loss than those pregnancies arising from intercourse occurring the day before or after ovulation [313]. Sperm can live for up to three days inside the female reproductive tract, but delay in fertilisation is thought to result in sperm damage leading to the greater likelihood of a conceptus which will miscarry [314]. A further study however, which assumed the most recent act in the fertile window led to conception, found no association between the age of sperm and pregnancy viability[315]. This theory has received little attention in the research community, with more emphasis placed on sperm DNA fragmentation for ageing or sub-fertile males and its relation to recurrent miscarriage [316; 317]. Intercourse data may be used to investigate the association between the quality of sperm

and pregnancy viability. In particular the number of acts of intercourse may represent a conduit for the quality of the sperm. A higher number of acts in the fertile window operates as a natural method of sperm selection, whereby the 'best' sperm succeeds in fertilisation. Conversely fewer acts could provide fewer quality sperm for possible conception, and these pregnancies may be more likely to end in a loss.

8.2.3 Aims

This analysis aims to answer three specific questions about the association between intercourse and miscarriage, utilising the diary data collected as part of the General Cycle Collection study.

- Does intercourse in the luteal phase or peri-implantation window increase the rate of miscarriage?
- Are pregnancies conceived of acts of intercourse in advance of ovulation more likely to end in miscarriage due to ageing sperm?
- Do more acts in the fertile window reduce the rate of miscarriage by increasing sperm quality?

It is not expected that an association will be found, however the analysis will not only remedy the dearth of information in this area, but a null finding may provide reassurance and assuage guilt for mothers who experience losses.

8.3 Methods

8.4 Data and timelines

The data collected as part of the General Cycle Collection study introduced in section 7.3 of Chapter 7 will be used for this analysis. A reminder that this study followed up women, aged 18-45, who were attempting to conceive. Women

collected daily early morning urine samples from day one of their cycle (first day of period) to seven days of the next cycle if they did not conceive and up to day 60 of the current cycle if they did. In addition to contributing urine samples and maternal history data, volunteers also self-reported the frequency of intercourse on each day of the cycle in a daily diary.

The data consists of 288 viable pregnancies and 82 losses. Of the 82 losses, 65 were classified as losses before 6 weeks and 17 as after 6 weeks. Three twin pregnancies were excluded from analysis to leave 285 viable singleton pregnancies. The baseline variables by pregnancy group can be viewed in Table 7.1.

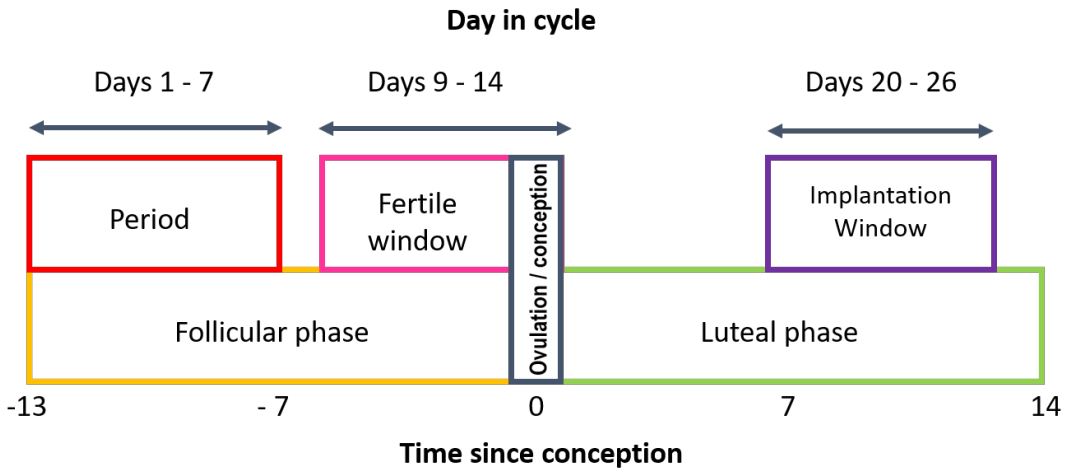


FIGURE 8.2. Key phases for a 28-day menstrual cycle

Figure 8.2 shows the key phases of the menstrual cycle which determine the timeline for the analysis, assuming a 28-day cycle. Miscarriage was modelled from the time of conception as women must first conceive to be at risk of loss. Conception was assumed to occur on the day of ovulation. Women entered the study at first detection of an hCG value of at least 2 mIU/ml, indicating successful implantation and confirming pregnancy at the sensitivity of the assay. Implantation

occurs approximately a week after conception, resulting in delayed entry into the study.

8.4.1 Statistical methods

For each aim the analysis followed a similar pattern. Where the aim was to determine whether an act in a time-window was associated with miscarriage, a binary intercourse variable was defined, assuming that one act carried the same risk as multiple acts on a given day. For questions around sperm age and quality continuous variables were defined to quantify the number of; (i) days before ovulation the last act in the fertile window took place and (ii) acts in the fertile window. Women only entered into the study if they contributed an hCG observation above 2mIU/ml. For women who had only one hCG observation above the detection limit, survival times were imputed by adding one day to the entry time. Delayed entry Cox proportional-hazards models were fitted to model time-to-miscarriage. No assumptions about the baseline hazard were made. Univariable and multivariable models were fitted. The multivariable model also included the covariates known to be associated with miscarriage, as identified in the analysis conducted in 7, with the exception of P3G and FSH3 on the day of implantation. This included the covariates maternal age centred at 30 years, BMI centred at 25kg/m^2, smoking status and the previous number of miscarriages. P3G and FSH3 were not measured on the day of implantation for all individuals in the study, resulting in a significant reduction in sample size when included in the model. An additional quadratic term for age was included in the model. The appropriateness of the functional form of continuous variables, BMI, age and previous number of miscarriages was assessed by plotting Martingale residuals against each covariate. The proportional hazards assumption was assessed by plotting Schoenfeld residuals against time. Outliers and influential observations were identified using deviance

residuals and delta-beta estimates. Where appropriate sensitivity analyses were conducted excluding any identified outlying observations.

8.5 Results

For seven women hCG levels above 2mIU/ml were not observed. Subsequently an entry time into the study and censoring time could not be confirmed. These women were excluded from the analysis. After exclusion of hCG values less than 2mIU/ml an additional three individuals experienced an event on the same day they entered the study. To allow these individuals to be included in the analysis survival times were imputed by adding one day to the entry time. All subsequent models were fitted to a maximum of 285 viable and 75 failing pregnancies.

8.5.1 Acts during the luteal phase

For the purposes of this analysis, the luteal phase was defined as two to ten days post-ovulation. A binary variable of whether an act occurred in this time-window or not was defined.

Acts in the luteal phase are summarised in Table 8.1 with the percentage of acts presented graphically in Figure 8.3. A slightly higher percentage of women in the failing pregnancy group reported having intercourse in the luteal phase than women who experienced a viable pregnancy. However in both groups the majority of women did have intercourse post ovulation, 73.3% of women who experienced viable pregnancies and 79.3% of women who experienced losses. The median number of acts for both groups was 2, with an interquartile range of 3 for the viable pregnancy group and 2 for the failing pregnancy group.

TABLE 8.1. Intercourse in the luteal phase (two to ten days post ovulation) by viability group

Act	Viable (n=285)	Miscarriage(n=82)
No	76 (26.7)	17(20.7)
Yes	209 (73.3)	65 (79.3)
All values are n (%)		

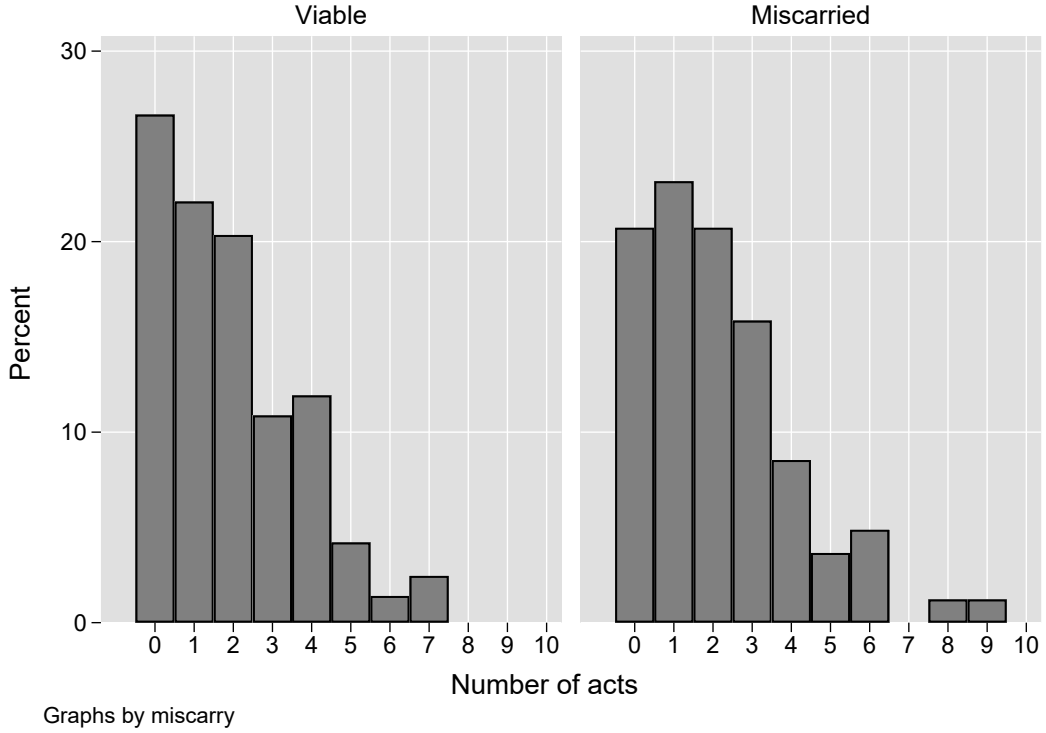


FIGURE 8.3. Number of acts in luteal phase by viability group

The univariable model estimates are given in Table 8.2. There was an increased rate of miscarriage of 42% (HR: 1.422; 95% CI: 0.807, 2.506) if an act occurred in the two to ten days post ovulation. This was not statistically significant. There was no evidence of a time-dependent effect for acts in the luteal

phase, when an interaction with log time was included in the model. The resulting likelihood ratio test gave a p-value of 0.938. Figure 8.4 shows the plotted time-dependent log hazard ratio for luteal acts.

TABLE 8.2. Univariable Cox model estimates for time-to-miscarriage and acts in the luteal phase

Variable	Hazard Ratio	95% CI
Act in luteal phase		
Yes	1.422	0.807, 2.506
CI: Confidence Interval		

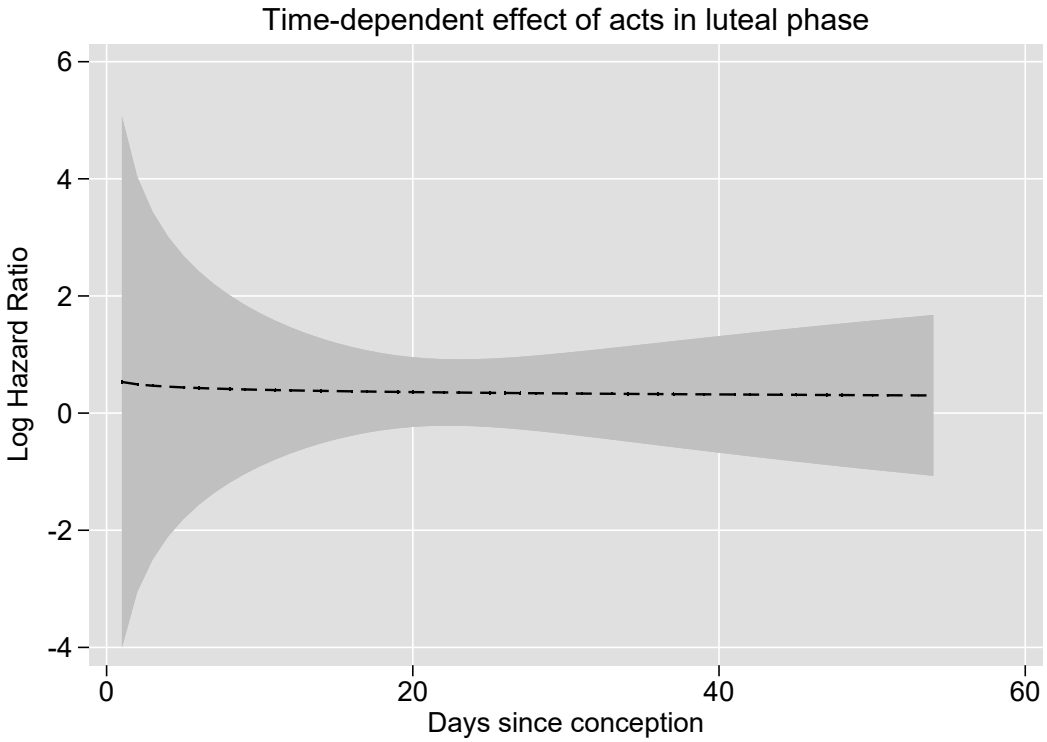


FIGURE 8.4. Time-dependent effect for acts in luteal phase

Multivariable model estimates are given in Table 8.3. When maternal age, BMI, smoking status and the number of previous miscarriages were included in the model an act of intercourse corresponded to a 58.0% increase (HR: 1.580; 95%

CI: 0.886, 2.817) in the rate of miscarriage, however this again was not significant at the 5% level.

TABLE 8.3. Multivariable Cox model estimates for time-to-miscarriage and acts in the luteal phase

Variable	Hazard Ratio	95% CI
Act in luteal phase		
Yes	1.580	0.886, 2.817
Centred age, years	1.019	0.979, 1.062
Centred quadratic age, years	1.009	1.004, 1.015
Centred BMI, kg/m^2	0.982	0.942, 1.023
Smoking status		
Never	Reference	
Current/Previous	0.844	0.521, 1.368
Number of previous miscarriages	0.917	0.739, 1.138

CI: Confidence Interval

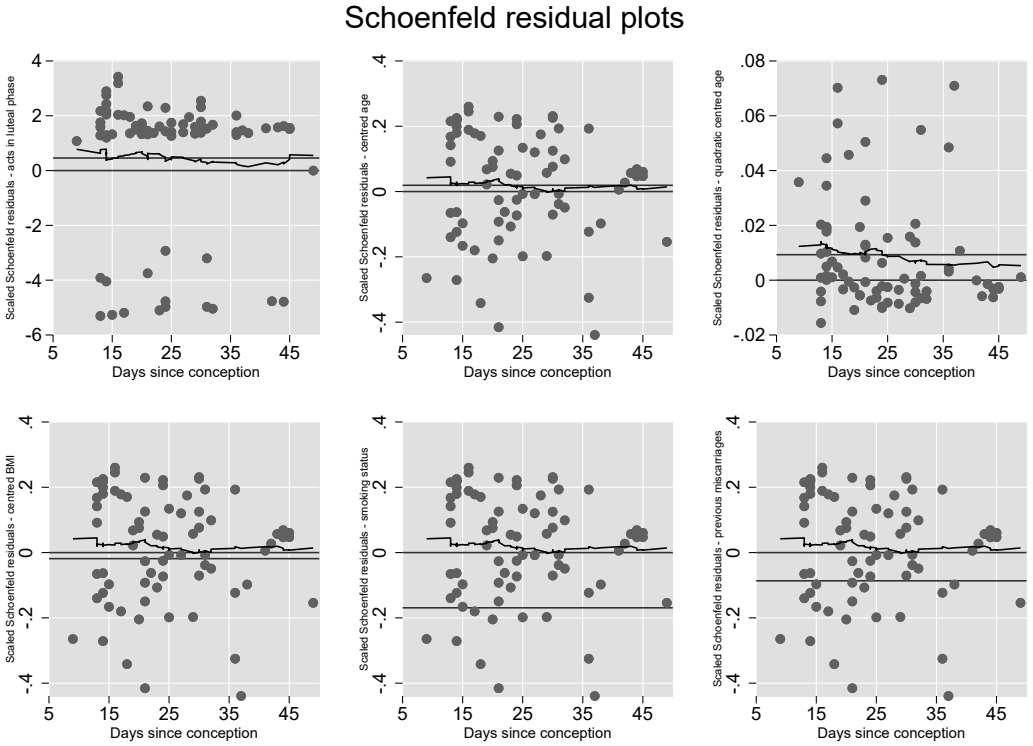


FIGURE 8.5. Schoenfeld residual plots to assess the proportional hazards assumption for the multivariable model for acts in the luteal phase

Schoenfeld residual plots for each variable included in the model are shown in Figure 8.5. Smoothed lines for smoking status and previous number of miscarriages did not follow reference lines for the log hazard ratio, although corresponding p-values (p=0.621 and p=0.358 respectively) indicated no violation of the proportional hazards assumption.

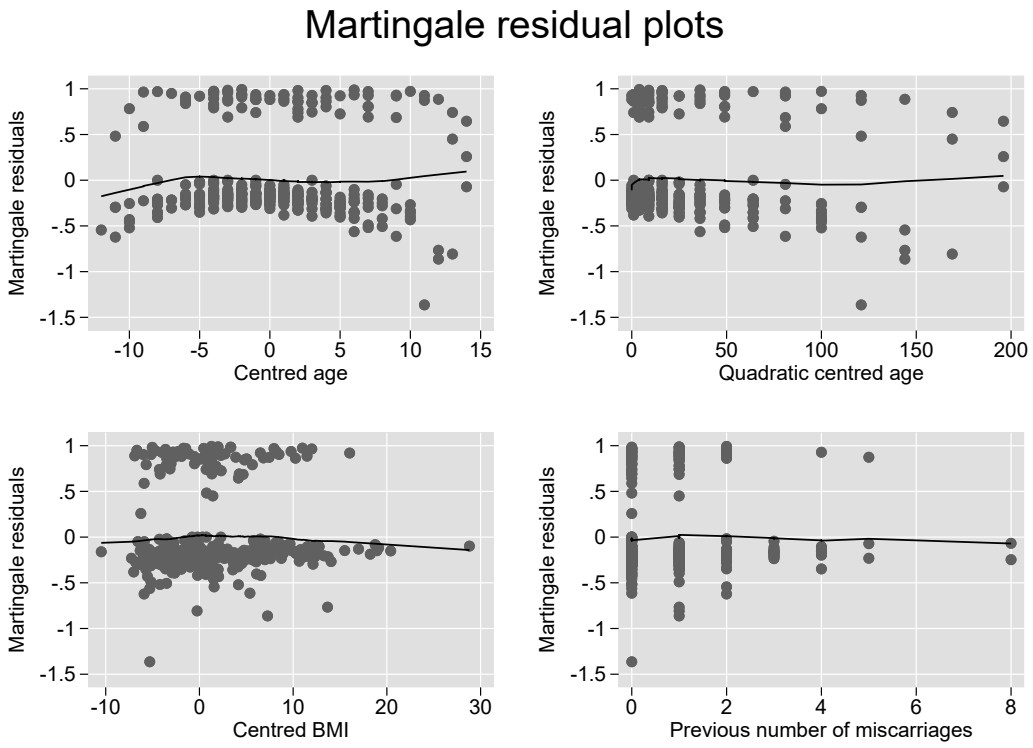


FIGURE 8.6. Martingale residual plots to assess the functional forms of variables included in the multivariable model for acts in the luteal phase

Martingale residuals were plotted to assess the functional form of continuous variables included in the model and are shown in Figure 8.6. The addition of the quadratic term to model age addressed non-linearity at the younger and older ages. Linear modelling of BMI in this instance appears adequate, with some curvature at the higher extreme. This could be influenced by a very large BMI

value of 53.81kg/m^2 (centred BMI: 28.81kg/m^2). Removal of this BMI value did not significantly alter the log hazard ratio estimate for BMI or other variables in the model.

A deviance residual plot to identify outliers is shown in Figure 8.7. Residuals above 1.96 correspond to early losses occurring before six weeks of pregnancy. By definition these losses will occur very early on in the follow-up period, and so do not represent true outlying observations.

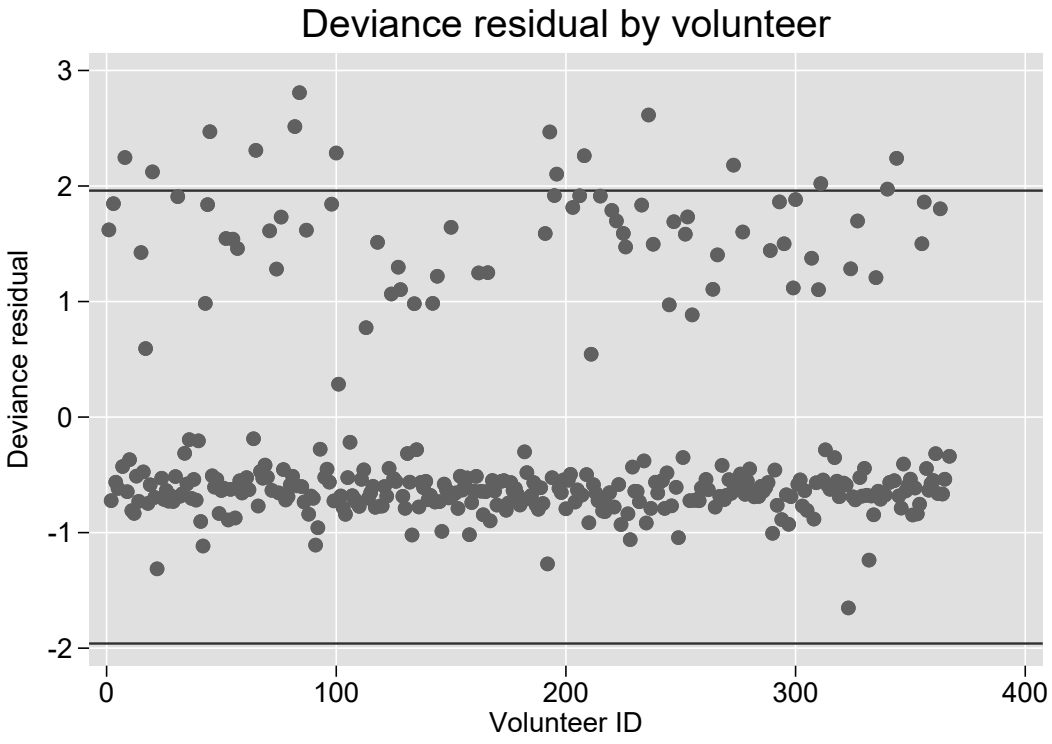


FIGURE 8.7. Deviance residual plot for multivariable model for acts in the luteal phase

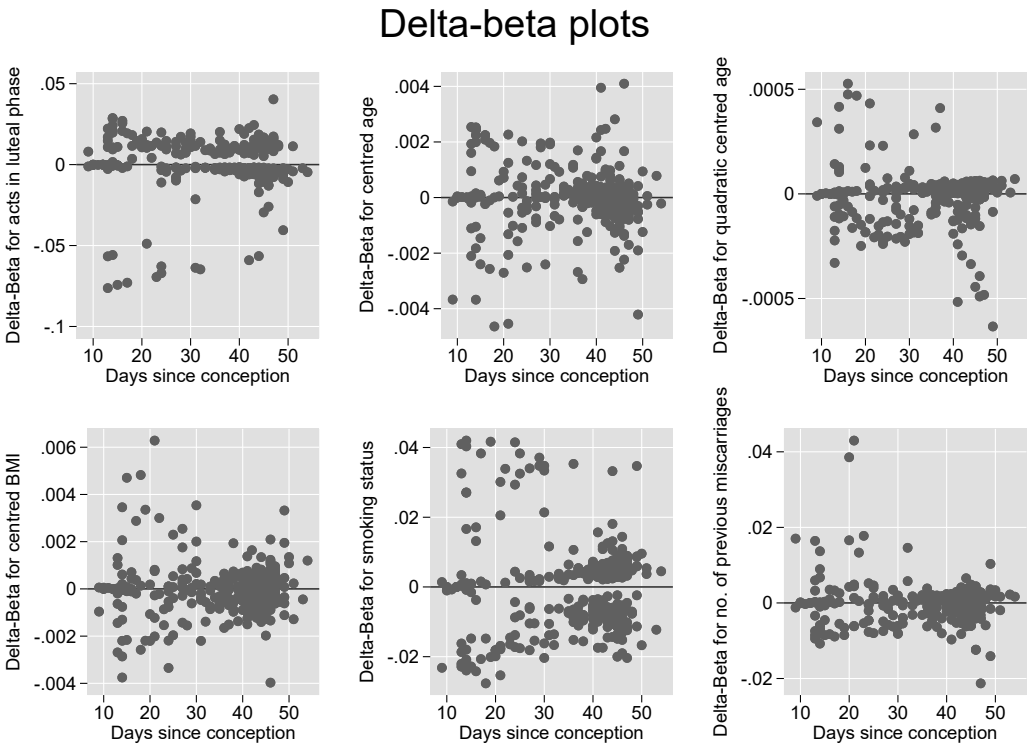


FIGURE 8.8. Delta-beta plot for multivariable model for acts in the luteal phase

Delta-beta plots for identifying influential observations are presented in Figure 8.8. On the whole delta-betas for each covariate were close to zero signalling little impact of systematically removing individuals on coefficient estimates. For acts in the luteal phase, delta-betas were ≤ -0.05 for twelve individuals who did not report having intercourse in the luteal phase but did miscarry. When these individuals were removed and the model refitted, having intercourse in the luteal phase became statistically significant (HR:8.04 95% CI:2.48, 26.04). Removal of these individuals, however, also reduced the number of events from 75 to 63. For smoking status five women who were current or previous smokers had a delta-beta estimate of ≥ 0.04, and all miscarried. The rate of miscarriage for current/previous smokers decreased (HR: 0.67 95% CI: 0.40, 1.12) on removal of

these individuals, however the effect as before was not statistically significant. Removal of one individual who experienced eight previous miscarriage did not alter the estimate of this coefficient greatly (HR:0.95 95% CI: 0.75, 1.18).

8.5.2 Acts during implantation

The peri-implantation window was defined as five to nine days post ovulation [55]. Three-day implantation windows were also identified. This encompassed the first day a minimum of 2 mIU/ml hCG measurement was observed, the day before and the day after. Binary variables for whether an act took place either in the peri-implantation window or narrower three-day implantation window were defined.

Table 8.4 shows the number of women who did and did not have intercourse in the peri-implantation window by pregnancy viability group. Of the women who experienced viable pregnancies 59.3% had intercourse compared with 69% of women who experienced losses. Figure 8.9 displays the percentage of intercourse by pregnancy group. A greater percentage of women in the viable pregnancy group (n(%): 116(40.7%)) did not have intercourse in the implantation window compared with women who miscarried (n(%): 26(31.7%)). For women who miscarried 37.8% reported one instance of intercourse in this window, compared with 27.7% of women who experienced viable pregnancies. The median number of acts was 1 (Q_1, Q_3: 0,2) for both viable and miscarried groups.

TABLE 8.4. Acts in the peri-implantation window (five to nine days post ovulation) by viability group

Acts	Viable (n=285)	Miscarriage(n=82)
No	116 (40.7)	26 (31.7)
Yes	169 (59.3)	56 (68.3)
All values are n (%)		

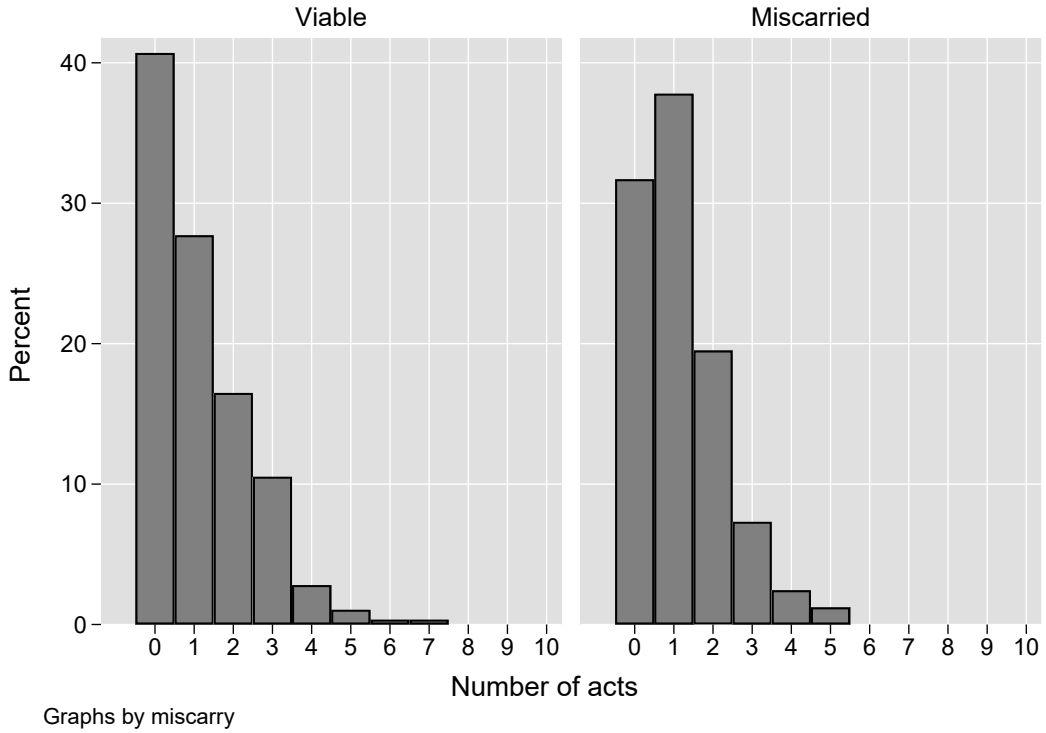


FIGURE 8.9. Number of acts in the peri-implantation window (five to nine days post ovulation) by viability group

Table 8.5 shows the univariable estimates for acts in the peri-implantation window. There was a 44.7% (HR: 1.447; 95% CI: 0.891, 2.352) increase in the rate of miscarriage at time t if an act occurred in the specified window.

TABLE 8.5. Univariable Cox model estimates for time-to-miscarriage and acts in the peri-implantation window

Variable	Hazard Ratio	95% CI
Act in peri-implantation window		
Yes	1.447	0.891, 2.352
CI: Confidence Interval		

There was no evidence of a time-dependent effect for acts in the implantation window when an interaction with log time was included in the model ($p = 0.741$)

Figure 8.10 shows a plot of the time-dependent log hazard ratio for acts in the peri-implantation window.

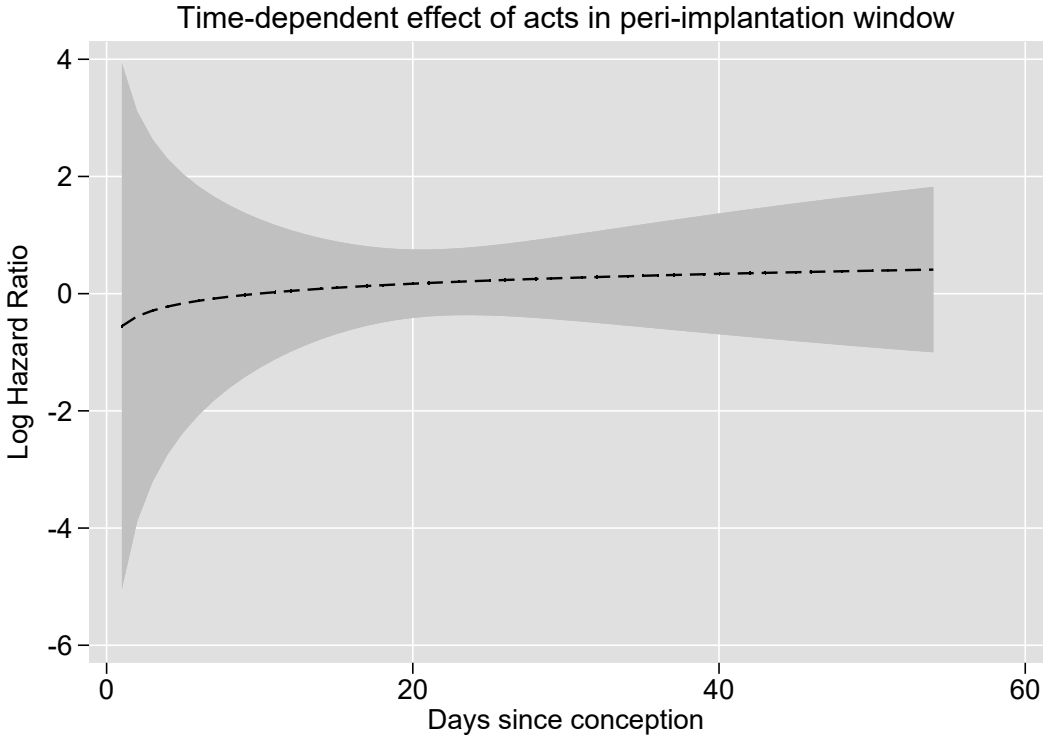


FIGURE 8.10. Time-dependent effect for acts in the implantation window

On inclusion of additional variables in the model, the rate of miscarriage increased to 63.6% (HR:1.636; 95%: 0.993, 2.693) as shown in Table 8.6. As with the univariable model, this effect was not statistically significant.

TABLE 8.6. Multivariable Cox model estimates for time-to-miscarriage and acts in the peri-implantation window

Variable	Hazard Ratio	95% CI
Act in peri-implantation window		
Yes	1.636	0.993, 2.693
Centred age, years	1.018	0.978, 1.060
Centred quadratic age, years	1.010	1.005, 1.015
Centred BMI, kg/m^2	0.982	0.941, 1.024
Smoking status		
Never	Reference	
Current/Previous	0.852	0.526, 1.379
Number of previous miscarriages	0.922	0.745, 1.142
CI: Confidence Interval		

Model checks were carried out in a similar manner to that presented in section 8.5.1. Schoenfeld residuals did not show evidence of a violation of the proportional hazards assumption. Martingale residuals showed continuous variable had been modelled adequately. A similar set of outlying observations were identified, and removal did not affect estimates. See Appendix B.1 for plots.

8.5.2.1 Acts during three-day implantation window

The previous peri-implantation window is generalised, assuming that the day of implantation is not known. As implantation is the point at which hCG begins to be produced, the actual day of implantation can be more closely pinpointed. The window was narrowed to the day hCG of at least 2mIU/ml was detected, the day before and day after.

Table 8.7 shows the number of women who experienced an act in this three-day implantation window. More women who experienced viable pregnancies (n(%): 124(43.5%)) had intercourse in the three-day implantation window than women who miscarried (n(%): 31(37.8)). The percentage of acts by viability group are shown in Figure 8.11. The median number of acts for healthy and miscarried pregnancy groups was 0 (Q_1, Q_3: 0, 1). Of the women who did report having

intercourse, 70% and 61% of the healthy and failing pregnancy groups respectively recorded a single act.

TABLE 8.7. Acts in the narrowed implantation window relative to first hCG observation $\geq$ 2 mIU/ml, by viability group

Acts	Viable (n=285)	Miscarriage(n=82)
No	161 (56.5)	51 (62.2)
Yes	124 (43.5)	31 (37.8)
All values are n (%)		

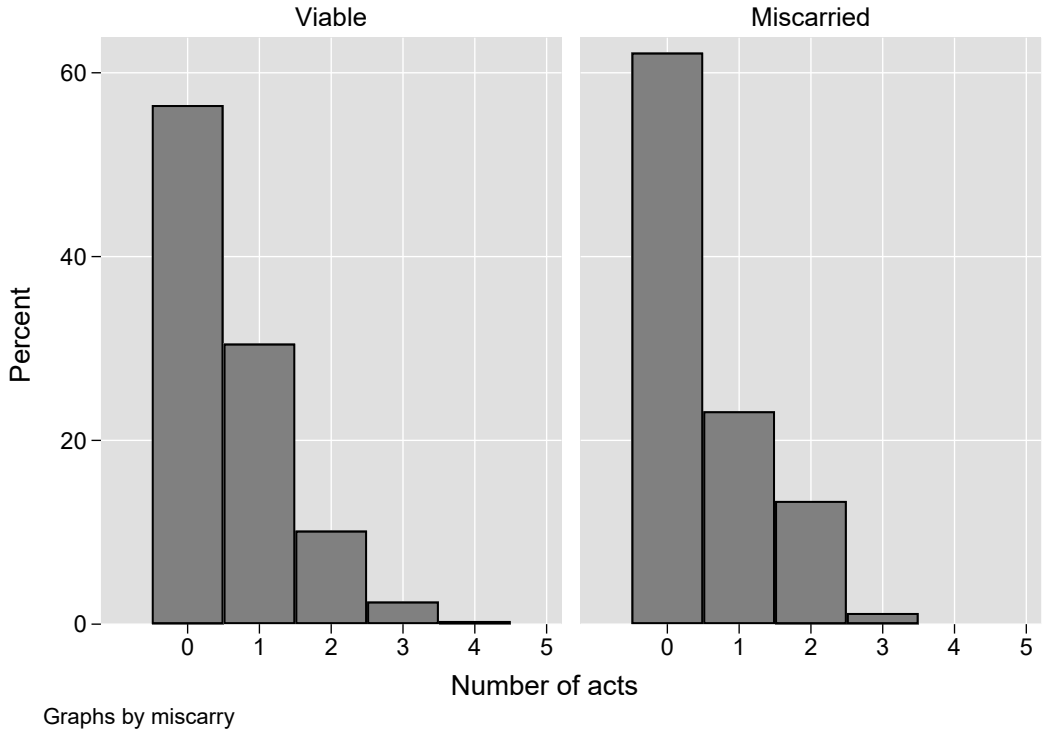


FIGURE 8.11. Number of acts in the narrowed implantation window relative to first hCG observation $\geq$ 2 mIU/ml, by viability group

Results from the univariable model are shown in Table 8.8. An act in the window inferred a 7.3% (HR:0.927 95% CI: 0.582, 1.473) decrease in the rate

of miscarriage. This was not significant at the 5% level. Addition of a log-time interaction failed to indicate evidence of a time-dependent effect (p=0.803). Figure 8.12 shows the time-dependent log hazard ratio for an act in the three-day implantation window.

TABLE 8.8. Unadjusted Cox model estimates for time-to-miscarriage and acts in individual implantation window

Variable	Hazard Ratio	95% CI
Act in 3-day implantation window		
Yes	0.928	0.584, 1.473
CI: Confidence Interval		

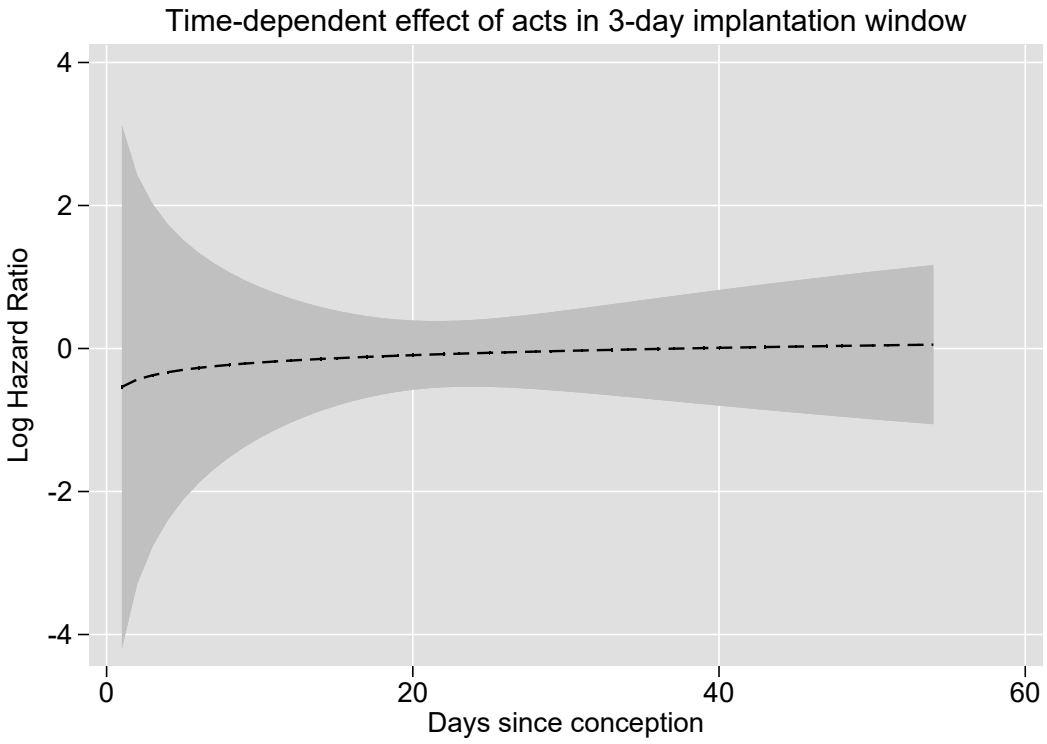


FIGURE 8.12. Time-dependent effect for acts in the three-day implantation window

The multivariable model estimate, shown in Table 8.9, demonstrated an act in the window corresponded to a smaller 3.7% (HR: 0.963; 95% CI: 0.605, 1.533) decrease in the rate of miscarriage. Again this was not significant at the 5% level.

TABLE 8.9. Multivariable Cox model estimates for time-to-miscarriage and acts in individual implantation window

Variable	Hazard Ratio	95% CI
Act in 3-day implantation window		
Yes	0.963	0.605, 1.533
Centred age, years	1.014	0.973, 1.056
Centred quadratic age, years	1.009	1.004, 1.014
Centred BMI, kg/m^2	0.982	0.942, 1.023
Smoking status		
Never	Reference	
Current/Previous	0.847	0.523, 1.370
Number of previous miscarriages	0.920	0.741, 1.142

CI: Confidence Interval

Model checks showed no violation of the proportional hazards assumption. The functional forms of variables were adequately modelled. Removal of outlying observations did not affect model estimates. For plots see Appendix B.2.

8.5.3 Acts in the fertile window

Acts in the fertile window were used as surrogates to answer questions about sperm age and quality and their association with miscarriage. The fertile window was defined as the day of ovulation and the five days preceding ovulation. The last act of intercourse in the fertile window was identified and used as a marker for the age of the sperm which contributed to conception.

Table 8.10 shows the day prior to ovulation on which women experienced their last act of intercourse, by pregnancy viability. For the majority of women the last act in the fertile window most commonly fell on the day of ovulation or the day before. This represents 89.1% of women who experienced healthy pregnancies and 90.1% of women who miscarried.

TABLE 8.10. Last act in the fertile window (the day of ovulation and the five days prior) relative to the day of ovulation

Number of days prior to ovulation	Viable (n=285)	Miscarriage (n=82)
0	157 (55.3)	40 (49.4)
1	96 (33.8)	33 (40.7)
2	19 (6.7)	7 (8.6)
3	7 (2.5)	0 (0.0)
4	2 (0.7)	1 (1.2)
5	3 (1.1)	0 (0.0)
All values are n(%)		

Univariable model estimates are presented in Table 8.11. For each additional day before ovulation that the last act of intercourse took place there was a 3.4% (HR: 0.966; 95% CI: 0.743, 1.257) decrease in the rate of miscarriage. Addition of an interaction with log time did not indicate evidence of a time-dependent effect (p=0.325). Figure 8.13 shows the time-dependent log hazard ratio for the last act in the fertile window.

TABLE 8.11. Univariable Cox model estimates for time-to-miscarriage and last act in fertile window

Variable	Hazard Ratio	95% CI
Day of last act prior to ovulation	0.966	0.743, 1.257
CI: Confidence Interval		

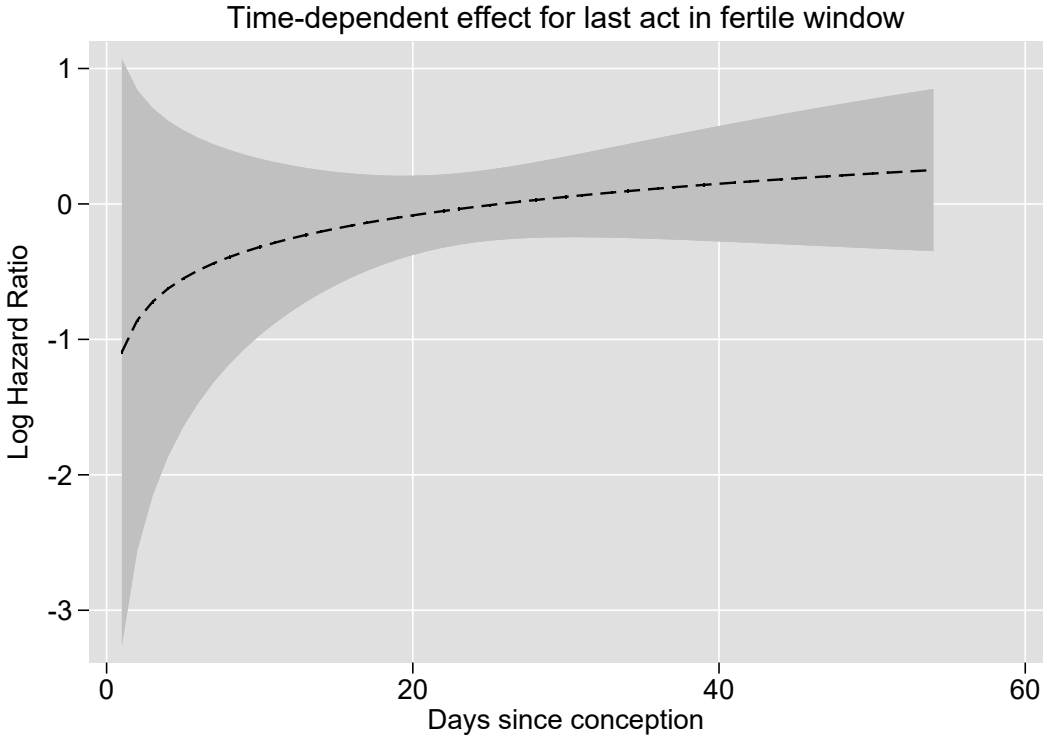


FIGURE 8.13. Time-dependent effect for the last act in the fertile window

TABLE 8.12. Multivariable Cox model estimates for time-to-miscarriage and last act in fertile window

Variable	Hazard Ratio	95% CI
Day of last act relative to ovulation	0.962	0.737, 1.256
Centred age, years	1.013	0.972, 1.055
Centred quadratic age, years	1.009	1.004, 1.014
Centred BMI, kg/m^2	0.984	0.944, 1.026
Smoking status		
Never	Reference	
Current/Previous	0.855	0.527, 1.368
Number of previous miscarriages	0.923	0.744, 1.145

CI: Confidence Interval

Multivariable model estimates are shown in Table 8.12. A similar 3.8% (HR: 0.962; 95% CI: 0.737, 1.256) decrease in miscarriage rate for each day prior to ovulation the last act occurred was found. This again was not statistically significant at the 5% level.

There was no evidence that the proportional hazards assumptions was violated. The functional form of variables included in the model were appropriately modelled. Removing influential observations did not appreciably affect estimates. See Appendix B.3 for plots.

8.5.3.1 Number of acts in the fertile window

The number of acts in the fertile window by viability group are presented in 8.13. Most women in both the healthy and failing pregnancy groups reported having intercourse three or four times in the fertile window, 53.1% and 56.1% for healthy and failing pregnancy groups respectively. One woman who conceived and subsequently miscarried did not report having intercourse in the fertile window.

TABLE 8.13. Number of acts in the fertile window, the day of ovulation and five days preceding, by viability group

Number of acts	Viable (n=285)	Miscarriage (n=82)
0	0 (0.0)	1 (1.2)
1	18 (6.3)	1 (1.2)
2	44 (15.5)	12 (14.6)
3	81 (28.5)	24 (29.3)
4	70 (24.6)	22 (26.8)
5	42 (14.8)	13 (15.9)
6	18 (6.3)	5 (6.1)
7	5 (1.8)	2 (2.4)
8	2 (0.7)	1 (1.2)
9	3 (1.1)	0 (0.0)
10	0 (0.0)	0 (0.0)
11	1 (0.4)	1 (1.2)
All values are n(%)		

Univariable model estimates are presented in Table 8.14. For every additional act in the fertile window there was a 9.4% (HR:1.094; 95% CI: 0.952, 1.257) increase in the rate of miscarriage at time t. Inclusion of an interaction with log time did not indicate a time-dependent effect(p=0.189). Figure 8.14 shows the time-dependent log hazard ratio for the number of acts in the fertile window.

TABLE 8.14. Univariable Cox model estimates for time-to-miscarriage and number of acts in the fertile window

Variable	Hazard Ratio	95% CI
Number of acts	1.094	0.952, 1.257
CI: Confidence Interval		

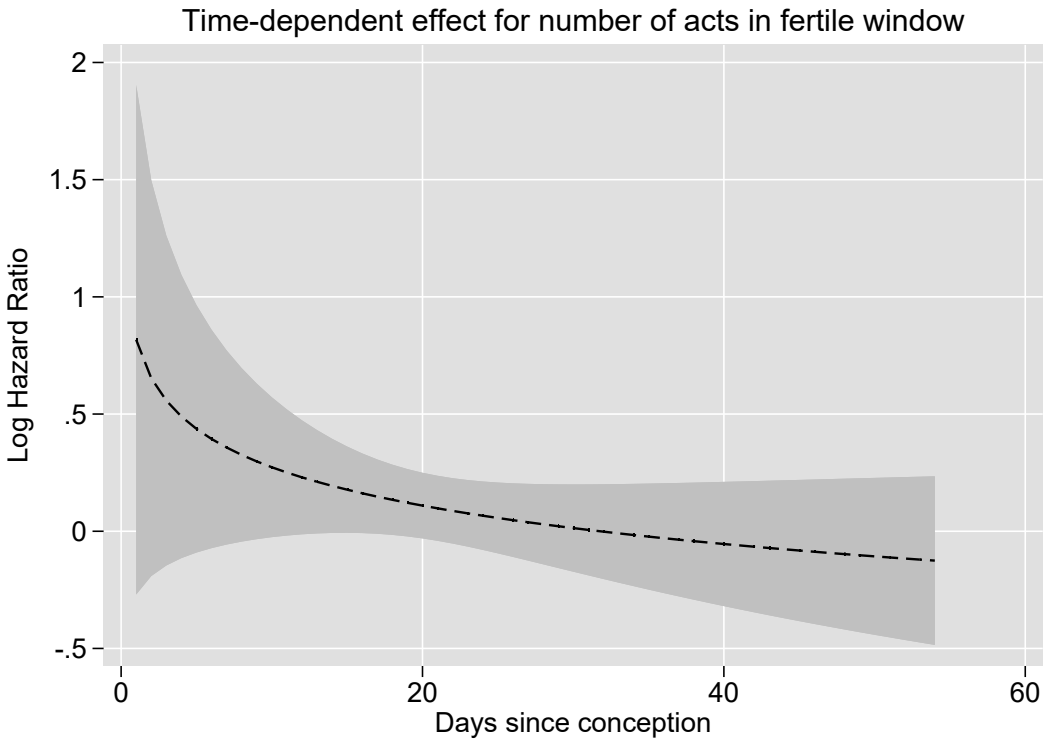


FIGURE 8.14. Time-dependent effect for the number of acts in the fertile window

The multivariable model estimates are presented in Table 8.15 suggest that each additional act of intercourse in the fertile window corresponded to a 9.1%

(HR: 1.091; 95% CI: 0.945, 1.260) increase in the rate of miscarriage. This effect was not statistically significant.

TABLE 8.15. Multivariable Cox model estimates for time-to-miscarriage and number of acts in the fertile window

Variable	Hazard Ratio	95% Confidence Interval	
Number of acts	1.091	0.945,	1.260
Centred age, years	1.014	0.973,	1.056
Centred quadratic age, years	1.009	1.004,	1.014
Centred BMI, kg/m^2	0.983	0.943,	1.026
Smoking status			
Never	Reference		Reference
Current/Previous	0.824	0.508,	1.337
Number of previous miscarriages	0.905	0.729,	1.123

CI: Confidence Interval

There was no evidence to suggest that the proportional hazards assumptions was violated. Linear and non-linear covariates were modelled adequately. Sensitivity analyses removing potentially influential observations did not substantially change model estimates. See Appendix B.4 for additional plots.

8.6 Discussion

The timing of intercourse at key points of the cycle and the ramifications for pregnancy outcomes were explored in this analysis. Intercourse in the luteal phase resulted in a 58% increase in the rate of miscarriage compared with those individuals who did not report having intercourse in this window. Similar results were seen for the an act of intercourse in the peri-implantation window, which sits within the luteal phase of the menstrual cycle. Specifically women who reported at least one act in the five to nine days post ovulation had a 63.6% increase in the rate of miscarriage when compared with women who did not experience an act in the window. When this implantation window was narrowed down to the day of implantation (hcg $> 2mIU/ml$) and the day preceding and following

implantation, acts of intercourse had a small protective effect, reducing the rate of miscarriage by 3.7% when compared with women who did not have intercourse in this three-day window.

Acts in the fertile window were used as a marker for the age of the sperm which resulted in fertilisation and for sperm quality. For each day before ovulation the last act of intercourse in the fertile window took place, the rate of miscarriage at time t decreased by 3.8%. This suggests that pregnancies occurring from sperm from older ejaculate, which has remained inside the woman longer before ovulation, are less likely to end in loss. For each additional act of intercourse in the fertile window there was a corresponding 9.1% increase in the rate of miscarriage at time t. If greater number acts signifies higher sperm quality, this suggests a detrimental effect for each additional act.

For each of these analyses the effect was not significant at the 5% level, though this is not unexpected given the exploratory nature of the analyses and the available sample size. The direction of effect for intercourse in the three-day implantation window indicated a non-significant protective effect, which goes against the idea that intercourse at implantation may interfere with pregnancy establishment. Findings for acts in the fertile window used as a proxy for sperm age and quality were also unexpected. A greater lag between ovulation and the last act of intercourse decreased the rate of miscarriage, suggesting acts close to ovulation do not necessarily mean a greater likelihood of a healthy pregnancy. Equally, targeting the fertile window with greater number of acts, and thereby increasing the quality of the sperm available for fertilisation, did not necessarily correspond to a better outcome.

8.6.1 Current evidence

As previously noted the research into whether intercourse can harm a pregnancy is sparse. In the words of Moscrop [308] we do not know and when mentioned in the literature it is to emphasise the absence of evidence [308]. Advice from healthcare professionals is varied with some advising restraint if experiencing bleeding, but intercourse as usual if not [318]. Yet the risks of intercourse in early pregnancy remain a concern for many couples [319]. A Canadian study on sexual activity during pregnancy, found 49% of women worried that intercourse would harm the pregnancy [320]. A similar Turkish study on sexual behaviour during pregnancy, found that of the women who avoided intercourse 49.1% held the belief that it would harm the baby [321]. It is conceivable that having intercourse during early pregnancy may be a convenient excuse for self-recriminations later if the pregnancy is not brought to term.

Guilt, self-blame and what-ifs have been established as very common emotions post loss[312]. This analysis did not find an association between the timing of intercourse and pregnancy loss, and in this case no news may be good news. This can provide reassurance to women who experienced miscarriages that abstaining would not have prevented the loss.

Those who miscarried tended to experience implantation later, at a median of 11 days ($Q_1, Q_3 : 10, 12$) post conception than women who experienced healthy pregnancies at a median of 9 days ($Q_1, Q_3 : 9, 10$) post conception. This delayed implantation is thought to be an indication of late fertilisation. Essentially there is a longer waiting time between ovulation and the sperm fertilising the egg [322]. This is more likely to occur when intercourse takes place on the day of ovulation rather than the five days preceding[315]. This was not the case for this dataset, in which a similar percentage of women who experienced viable pregnancies (55.3%),

and women who miscarried (49.4%) recorded the act thought to be responsible for conception on the day of ovulation.

Assessment of acts in the luteal phase served as a wider window for the period of corpus luteum development and subsequent implantation [323]. Although not analysed, a shorter luteal phase of less than ten days has been shown to be associated with difficulty in conceiving, and if pregnancy is achieved then it is more likely to end in an early loss [307]. A short luteal phase stunts the development of the womb lining which requires prolonged exposure to progesterone to mature, and this can in turn interfere with implantation. Progesterone treatment to ensure appropriates levels for maturation have been suggested for women who experience a short luteal phase, and therefore recurrent miscarriages [324]. As with research into intercourse and pregnancy in general, it is not known whether intercourse, particularly for women with short luteal phases, exacerbates matters.

Acts in the fertile window were assessed as a surrogate for sperm age and quality. It has been hypothesized that pregnancies from sperm retained for longer in the female reproductive tract are more likely to end in miscarriage, however there is little research available on the matter. It is acknowledged every ejaculate will contain inferior sperm, which have DNA fragmentation, and this is likely to be more of an issue with older males[316; 325]. A recent meta-analysis of 15 studies found that for cases of recurrent miscarriage the mean level of sperm DNA fragmentation was higher than for women who did not have a history of loss [317]. The GCC data is characterised by women who have suffered a loss previously. Approximately 44% of women who had healthy pregnancies and 45% of women who miscarried had suffered at least one loss previously. However only a smaller 8% and 2% of women who experienced healthy and failing pregnancies

respectively reported suffering recurrent losses ($\geq$ 3), suggesting sperm DNA fragmentation is not a factor here.

The number of acts in the fertile window were not significantly associated with time-to-miscarriage. Increased intercourse, however, is very well known to improve sperm quality. Evidence has shown that for sub-fertile men each subsequent ejaculation improved sperm motility from the first [326]. Conversely, long periods of abstinence between acts can lower sperm motility, which is why frequent intercourse during the fertile window is recommended [327].

8.6.2 Strengths and limitations

One of the main weaknesses of this study was the use of self-reported data on acts of intercourse. This introduces a level of unreliability, particularly as it was assumed that no observation was equivalent to no acts occurring on that day. It is likely that at least some of these blank entries are missing data, when an individual may have forgotten to fill out the diary though intercourse did take place. As binary variables were used for analyses of acts in the luteal and implantation windows, as long as one act was recorded this would have made no difference to estimates. However, for analyses in the fertile window, missing records could have affected the act which was assumed to lead to conception, as well as the frequency of acts and the association with miscarriage.

Use of binary variables for selected analyses means information about the number of acts was lost, however due to the wide range of categories and low numbers in the higher categories this seemed the best option. This however, assumes that one act of intercourse is as hazardous as multiple acts, which may not be the case. If we are to belive that intercourse physically affects pregnancy from being securely established than, multiple physical acts are likely to be more harmful than a solitary physical act. To appropriately model this however, more

data for multiple acts would be required to allow for intercourse to be modelled continuously.

Models were fitted to a total of 360 individuals, including 75 events. Considering the number of variables included in the model, the number of events is likely too few to produce unbiased parameter estimates [328]. The much-debated general rule of thumb recommends at least ten events per variable included in the model [329]. In some cases, for example binary variables with low prevalence, up to 20 events are advised, however this does not apply to the binary intercourse variables used in the luteal and implantation window models [328].Vittinghoff and McCulloch [330] points out however that reduced coverage, increased Type I error rate and increase relative bias are infrequent even when the events per variable range from 5 to 9, though they are still seen when events per variable exceed 10 up to 16 [330]. Regardless, considering the low number of events estimates must be interpreted with caution, particularly as these analyses were not powered to detect a significant effect size. Post-hoc power analyses show that a sample size of 360, with an event probability of 21% had at best 57% power to detect a hazard ratio of 1.636 for acts in the implantation window. Whilst for the last act in the fertile window, the power to detect a log hazard ratio of 0.087 was only 18%. The sample size was clearly a barrier for this study.

Time since conception was used as the timeline for modelling time-to-miscarriage. The rationale is that women cannot be at risk of miscarriage until they conceive, and ovulation serves as the closest landmark to implantation, when pregnancy can be confirmed. Ovulation was estimated as the day after the LH surge, which is generally accepted as a reasonable marker for conception [331]. However, realistically it cannot be known that this day truly signals conception, particularly as ovulation can take place up to 36 hours after the surge [243]. As this analysis

focusses on intercourse in phases of the cycle, using an estimate may mean definitions of each window may not be accurate. However, using this approach also allows for a comparable timeline across women, which makes it the most useful. In a population who may have been struggling to conceive, generalised definitions of luteal and implantation windows (see section 8.2.1) are not likely to represent the women in the study. For example luteal phases have been shown to vary in length from 7 to 19 days, so using a 2 to 10 day post conception window may not capture the entire length of the phase [332]. With regards to the defined implantation window, for 83% of women who miscarried, implantation did not actually occur five to nine days post ovulation, but rather ranged from 10 to 15 days post ovulation. This could mean analyses investigating acts in the luteal and peri-implantation phase targeted the incorrect time-windows.

Women in this study were provided with conception aids in the form of ovulation testing. This suggests intercourse behaviour for this sample of women may not match that of women not using fertility aids. Certainly looking at the number of acts in the fertile window in Table 8.13, it is clear that women tended to target intercourse to best maximise conception. However, the data also suggests that women combined this with advice to have intercourse at least two to three times a week, with specific targeting in the fertile window [327]. It would be helpful however to investigate the relationship between timing of acts and miscarriage in a sample of women exhibiting 'natural' intercourse behaviour.

Acts of intercourse in the fertile window were assumed a reasonable proxy for sperm age and quality. However, it is not possible from the data available to firmly attribute conception to a particular day of intercourse. As seen in this very analysis, it is possible to conceive even when the act does not occur in the most commonly defined fertile window of ovulation and the five preceding days.

It is conceivable the sperm of acts post ovulation may have been responsible for conception. In this instance sperm age is not as relevant a question as the age of the egg, which could be the instigator of miscarriage [314]. Though the assumption of the last act has been utilised in the literature, pregnancy cannot necessarily be attributed to that particular act [315]. DNA fragmentation of sperm is known to be a factor in cases of recurrent miscarriage. However in this study sperm count was not directly measured, and of course a surrogate is not as helpful for analysis if it is possible to directly collect and measure sperm to assess quality [333].

The three-day implantation window was very much dependent on the detection limits of the assay used to measure hCG. It was assumed that implantation occurred on the day hCG first reached at least 2mIU/ml, however it is likely that implantation occurred prior to this, except the assay was not sensitive enough to detect this. However even if implantation was not correctly pinpointed, the estimation is sufficient as the day before and after the purported day were included for analysis.

8.6.3 Conclusion

No evidence of an association between intercourse in the luteal phase or peri-implantation window was found. This was reflected in the analysis looking at the three-day implantation window. There was no evidence that pregnancies as a result of acts further away from ovulation (in the fertile window) were more likely to end in miscarriage. Finally there was no evidence that a greater number of acts in the fertile window reduced the rate of miscarriage. The results are presented with the caveat that the sample size did not allow for adequate power to detect statistically significant effects at the 5% level. Further studies are required to add

to the evidence base in order to effectively advise women on whether intercourse in early pregnancy is unsafe.

Chapter 9

The effect of misspecifying a non-linear association structure in a joint model - a simulation study

9.1 Chapter overview

Joint models fitted in Chapters 6 and 7 assumed a linear association between the longitudinally modelled biomarker and risk of miscarriage, however it is unlikely that this assumption holds in many cases. In this chapter the effect of incorrectly assuming a linear association structure on subsequent survival predictions will be investigated via a simulation study. The focus here will be a quadratic association between the trajectory function and hazard. For each model the details of the data generating mechanism for the simulated datasets will be presented, along with the measures intended to be used to evaluate performance of the misspecified models. The corresponding estimates of bias and comparison of model and empirical standard errors for misspecified models will be presented. Finally the simulation will be evaluated for its strengths and limitations.

9.2 Introduction

It is common when developing a prediction model to consider non-linear effects of covariates. For example age is often modelled non-linearly to appropriately describe the greater effect or risk associated with certain age brackets [334; 335]. Another notable example is body mass index (BMI), which has been shown to be non-linearly associated with mortality and even healthcare costs [336; 337]. BMI values that fall outside of the healthy range(18.5kg/m^2 - 24.9kg/m^2) correspond to worse outcomes. This results in a U or J shaped hazard, with event rates higher for those with low and high BMIs. When developing prediction models it is instinctive to investigate whether covariate behaviour is linear, yet this is not yet a parallel consideration when selecting the appropriate association structure between the biomarker and time-to-event outcome for a joint model. In a joint model the biomarker, though modelled via a longitudinal model, is effectively a covariate in the survival submodel. Therefore joint model development should also consider non-linear effects of the biomarker on the time-to-event outcome. If these potentially non-linear effects are ignored then this will have a knock-on effect on model coefficients and predictions, introducing bias into the estimation.

Simulation studies have confirmed that the joint longitudinal-survival model improves upon simpler approaches discussed in section 5.3, both by utilising full longitudinal biomarker information, addressing measurement error and appropriately estimating uncertainty around estimates [28; 338; 339]. Misspecification of the hazard function of standard proportional hazards survival models has been studied extensively [246; 340; 341]. Where joint models are concerned, simulation studies have focussed on misspecification of the longitudinal and survival submodels. Crowther *et al.* [52] conducted a simulation study investigating the

impact of misspecifying the longitudinal submodel for various current value and first derivative association models [52]. Results suggested that when a simple linear trajectory was assumed over the true trajectory, which included quadratic and cubic polynomials, the parameter estimate for the time-dependent slope association (changes in which are an important indicator of miscarriage as shown in Chapter 7), was found to be substantially biased [52]. Yet the current value model estimates remained robust to misspecification. A simulation study conducted by Arisido *et al.* [53] found that, like in the standard survival model context, misspecifying the baseline hazard function in a joint model resulted in biased model estimates. And when a non-monotonic baseline hazard was misspecified the estimate for the association parameter was biased and had poor coverage [53]. Yet, leaving the baseline hazard unmodelled as per the approach by Song *et al.* [342] can lead to difficulty in obtaining reliable standard errors [54]. Notably these simulations have been carried out assuming that the relationship between the biomarker trajectory function and the hazard is linear in form. Though various associations between the trajectory and hazard function have been explored (see section 5.5), there has been no, to the author's knowledge, investigation of the consequences of misspecification of a non-linear association structure. Furthermore limited methodological and software development in this aspect mean that tools have not been available until recently to address the issue of misspecifying a non-linear association structure.

9.3 Simulation study

A simulation study was conducted to assess (i) the effect of misspecifying a *quadratic* association structure in a joint longitudinal-survival model on predicted

survival; (ii) the effect of misspecification on predicted survival as the strength of the quadratic association increases.

9.3.1 Methods

The following specification of the the joint model was utilised for simulation. A simple linear longitudinal trajectory, with random intercept and random slope was assumed so that,

$$m_i(t) = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i})t \tag{9.1}$$

where

$$\begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_0} & \rho\sigma_{u_{01}} \\ \rho\sigma_{u_{10}} & \sigma^2_{u_1} \end{bmatrix}\right) \tag{9.2}$$

To evaluate misspecification of the non-linear association structure, data was simulated for a quadratic association. The true survival submodel included a Weibull baseline hazard function, a time-independent treatment variable and association parameters which correspond to the relevant non-linear trajectory-hazard relationship. The true survival submodel is shown in Equation 9.3.

$$h_i(t) = \lambda\gamma t^{\gamma-1} \exp\left[\beta trt + \alpha_1 m_i(t) + \alpha_2 m_i(t)^2\right] \tag{9.3}$$

where,

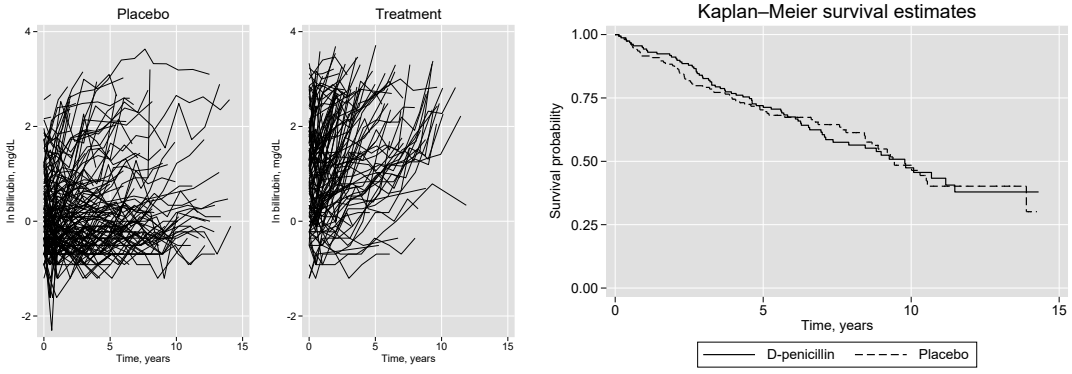
$$m_i(t) = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i})t$$

9.3.1.1 Data and parameter values

Simulated datasets were based on a commonly utilised joint model training dataset [30; 343]. The data was collected by the Mayo Clinic between 1974 and 1984 [344].

This trial followed up 312 patients diagnosed with the liver disease primary biliary cirrhosis (PBC). The patients were either randomised to placebo ($n = 154$) or the drug D-penicillamine ($n = 158$). PBC is characterised by an excessive build-up of bile in the liver due to damaged bile ducts. This eventually leads to cirrhosis and when left untreated, to death. The longitudinal biomarker of interest was billirubin, which due to skew was modelled on the natural logarithm scale. The time-to-event outcome was death.

Figure 9.1a shows the longitudinal log billirubin profiles for each treatment group. Log billirubin profiles for the treatment group are steeper than for the placebo group. The corresponding survival probability curves, by treatment group are presented in Figure 9.1b. Survival at the end of follow-up was approximately 37% for the treatment group and 30% for the placebo group.



(A) Longitudinal profiles of log serum billirubin, by treatment group

(B) Survival probability curves for mortality, by treatment group

FIGURE 9.1. Pulmonary billiary cirrhosis data graphs

To simulate data from a joint longitudinal-survival model setting the following distributions were used for all scenarios.

- $trt \sim U(0, 1) > 0.5$
- $u_0 \sim N(0, \sigma_{u_0})$

- $u_1 \sim N(0, \sigma_{u_1})$
- For $s \in \mathbb{Z} \in [1, 4]$ time $= |N \sim (s, 0.5)|$
- $y \sim N(m_i(t), \sigma_e)$

The fixed effect parameters for the intercept and slope were $\beta_0 = 0.5$ and $\beta_1 = 0.16$ respectively. The measurement error or residual error standard deviation was assumed to be $\sigma_e = 0.35$. The standard deviation for the random intercept u_{0i} was $\sigma_{u_0} = 1.02$, with the corresponding SD for the random slope u_{1i} equal to $\sigma_{u_1} = 0.17$. The covariance $\sigma_{u_{01}} = 0$. The scale and shape parameter values for the Weibull baseline hazard were chosen as $\lambda = 0.06$ and $\gamma = 1.58$. The corresponding log hazard ratio for the treatment variable was assumed to be -0.5.

For each true model the quadratic association parameter was varied as follows.

- Scenario 1: Fixed $\alpha_1 = 0.4$ and $\alpha_2 = (0.05, 0.1, 0.2)$
- Scenario 2: Fixed $\alpha_1 = -0.6$ and $\alpha_2 = (-0.05, -0.2, -0.3)$

Scenario 1 describes an association where an increase in biomarker implies greater risk, as with the PBC dataset. Conversely Scenario 2 presents an association where a decrease in biomarker indicates higher risk such as the GCC data (see Chapter 7).

Figure 9.2 shows the increase or decrease of the overall hazard ratio for the biomarker, as modelled by the trajectory function, as values increase for each model in Scenarios 1 and 2 respectively. For each subsequent model in a given scenario the J or reflected J shape, typical of a quadratic association, is more pronounced.

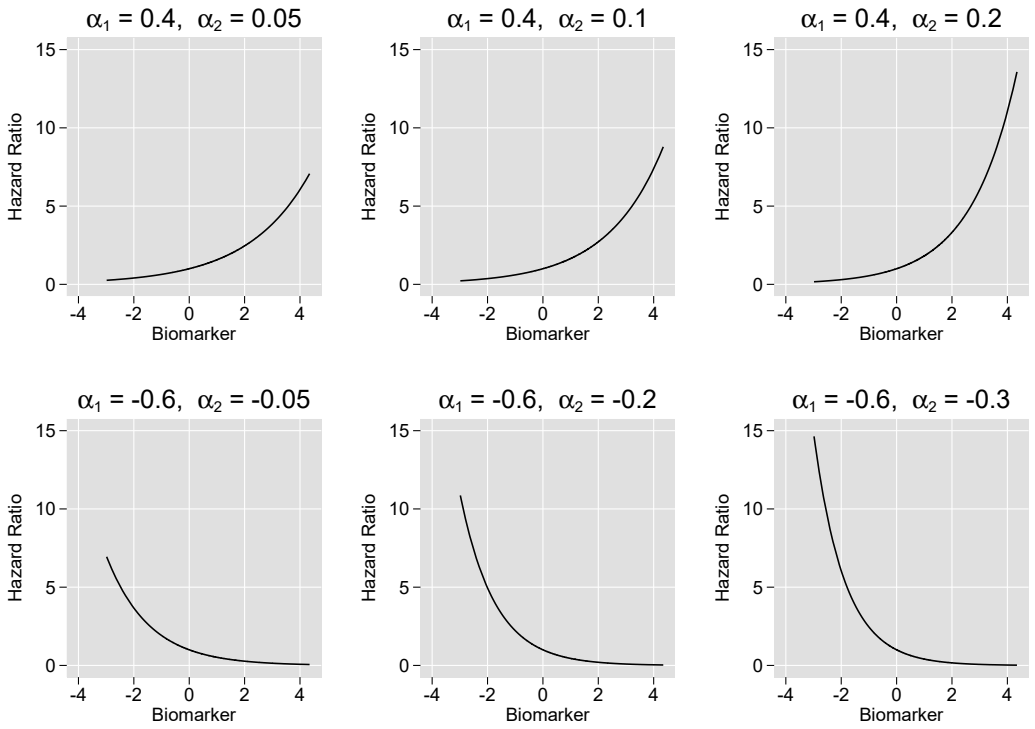


FIGURE 9.2. Association between the hazard function and biomarker trajectory for each model

The assumed true survival functions across the two scenarios are presented by treatment group in Figure 9.3. Survival is dependent on the fixed mean biomarker response (setting the random effects to zero) estimated by the fitted longitudinal trajectory function. As treatment is the only added covariate and only appears in the survival submodel, differences in true survival between treatment groups are based on the estimated time-varying fixed biomarker response, subject to the specification of the hazard-biomarker association, here quadratic for the true model.

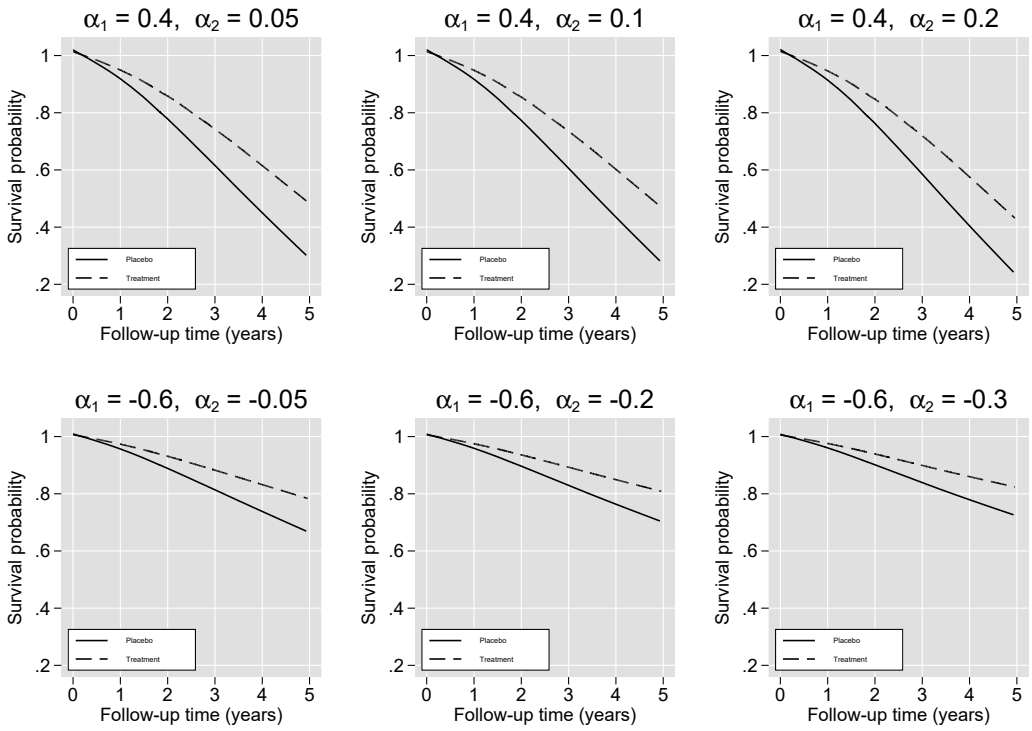


FIGURE 9.3. Assumed survival functions (setting random effects to 0) for each model from Scenarios 1 and 2

Survival times and events were simulated for 500 individuals, with a follow-up time of five years. Each line of data was then expanded to create a maximum of five repeated measurements for each individual. The initial observation was assumed to have been observed at baseline $t = 0$. To enable simulation of biomarker observation times which were spaced roughly one year apart, the times for observations 2 to 5 were simulated using a normal distribution with mean $\mu = 1, \cdots, 4$ where $\mu \in \mathbf{Z}$ is an integer and SD $\sigma = 0.5$. Any simulated times after the simulated event time were dropped. A joint model assuming only a linear association between biomarker trajectory and the hazard function was fitted to each simulated dataset. A maximum of thirty iterations was allowed to achieve convergence. This was based on investigating the number of iterations required for

model convergence for a small set of simulated datasets. Survival probabilities, independent of random effects, were then estimated for each treatment group and at baseline and each year thereafter.

9.3.1.2 Simulating survival times

The survival times were simulated using the Stata package `survsim`.[345; 346] This implements and extends methods by Bender *et al.* [347] which presents techniques to simulate survival times for a Cox proportional hazards model, utilising one of the exponential, Weibull or the Gompertz distributions.

Assuming proportional hazards, the survival function, $S(t)$ given a vector of baseline covariates $\boldsymbol{x_i}$ can be written in terms of the cumulative baseline hazard function $H_0(t)$ as shown by Equation 9.4.

$$S(t \mid x_i) = \exp[-H_0(t)\exp(\boldsymbol{\beta x_i})] \tag{9.4}$$

For a Weibull distribution the baseline cumulative hazard function is equivalent to $H_0(t) = \lambda t^{\gamma}$. Following from Equation 9.4, the probability of death is given by,

$$J = F(t \mid \boldsymbol{x_i}) = 1 - S(t \mid \boldsymbol{x_i}) \tag{9.5}$$

$$1 - J = 1 - F(t \mid \boldsymbol{x}_i) = \exp[-H_0(T)e^{\boldsymbol{\beta x_i}}] = \exp[-\lambda t^{\gamma} e^{\boldsymbol{\beta x_i}} \sim U[0,1] \tag{9.6}$$

So if $J \sim U[0,1]$ then it follows that $1 - J \sim U[0,1]$. So provided the baseline hazard function $h_0(t) > 0$ for all t and that $H_0(t)$ is invertible, survival times T can be calculated after drawing from a uniform distribution. More specifically we solve Equation 9.6 for T so that it is written as a function of U, the relevant

parameters for the baseline cumulative hazard function (λ and γ), and vector of baseline covariates $\boldsymbol{x_i}$ with associated parameter estimates $\boldsymbol{\beta}$[347][345].

$$T = H_0^{-1}[-ln(J)\exp(-\boldsymbol{\beta x_i})] = \Big(- \frac{ln(J)}{\lambda \exp(\boldsymbol{\beta x_i})} \Big)^{\frac{1}{\gamma}} \tag{9.7}$$

9.3.1.3 Estimands

Coefficients for linear and non-linear association joint models could not be directly compared to assess misspecification. Instead, baseline survival probabilities (setting random effects parameters to 0) predicted from the misspecified model were compared to true model survival probabilities at yearly follow-up points in each treatment group. The estimand of interest then, denoted θ, is the joint model survival function evaluated at $t = 0, 1, 2, 3, 4$ and 5 for treatment groups $trt = 0$ and 1 and is shown in Equation 9.8. Parameter values are referenced in section 9.3.1.1.

$$\theta = \exp\Big(- \int_0^t h(w)dw \Big) \tag{9.8}$$

where for the true value of the estimand,

$$h_i(w) = \lambda\gamma w^{\gamma-1} \exp\Big[\beta trt + \alpha_1(\beta_0 + \beta_1 w) + \alpha_2(\beta_0 + \beta_1 w)^2\Big]$$

9.3.1.4 Performance measures

The bias of the survival probabilities predicted from the simulated data and denoted $\hat{\theta}$ is defined as $Bias = E[\hat{\theta}] - \theta$. This was assessed using Equation 9.9. The empirical standard error, $\sqrt{Var\hat{\theta}}$, was estimated using Equation 9.10. Here $\bar{\theta}$ represents the mean of $\hat{\theta}$. The term n_{sim} refers to the number of simulations.

$$\text{Bias} = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{\theta}_i - \theta \tag{9.9}$$

$$\text{Emp SE} = \sqrt{\frac{1}{n_{sim} - 1} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \bar{\theta})^2} \tag{9.10}$$

The empirical standard errors were compared to average model standard errors, defined as $\sqrt{E\left[\widehat{Var}(\hat{\theta})\right]}$, and is estimated using Equation 9.11. The model SE should target the empirical SE so that $E(\text{Mod SE}^2) = \text{Emp SE}^2$ [348]. The corresponding relative percentage error in model SE was estimated using Equation 9.12.

$$\text{Average Mod SE} = \sqrt{\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \widehat{Var}(\hat{\theta}_i)} \tag{9.11}$$

$$\text{Relative \% error in Mod SE} = 100 \left(\frac{\widehat{ModSE}}{\widehat{EmpSE}} - 1 \right) \tag{9.12}$$

9.4 Results

9.4.1 Scenario 1

Table 9.1 details the results for models from Scenario 1. As a reminder, for these models the true linear association parameter was $\alpha_1 = 0.4$ and the quadratic association parameters were $\alpha_2 = 0.05, 0.1$, and 0.2 for models 1, 2 and 3 respectively. For each specification of α_2 in Scenario 1, a total of 990 out of 1000 models converged successfully.

Predicted survival probabilities for the misspecified model at baseline for each treatment group were unbiased across all choices of α_2. For all models in Scenario 1, misspecification of the quadratic association resulted in underestimation of the predicted survival probabilities at years 1 to 5. In general greater bias was

TABLE 9.1. Simulation results of bias, empirical standard error and relative % error in model SE for predicted survival probabilities from misspecified models in Scenario 1

Model	α_1	α_2	Treatment group	Time, years	True survival probability	Bias	% Bias	Monte Carlo SE	Empirical SE	Monte Carlo SE	Average Model SE	Monte Carlo SE	Relative % error*	Monte Carlo SE
1	0.4	0.05	0	0	1	-	-	-	-	-	-	-	-	-
				1	0.9253	-0.0020	-0.22	0.0003	0.0103	0.0002	0.0106	0.00003	2.77	2.33
				2	0.7842	-0.0067	-0.85	0.0006	0.0193	0.0004	0.0199	0.00003	3.03	2.32
				3	0.6164	-0.0124	-2.00	0.0008	0.0255	0.0006	0.0261	0.00002	2.54	2.31
				4	0.4489	-0.0166	-3.69	0.0009	0.0292	0.0007	0.0296	0.00003	1.42	2.28
				5	0.3014	-0.0176	-5.83	0.0010	0.0299	0.0007	0.0300	0.00004	0.35	2.26
			1	0	1	-	-	-	-	-	-	-	-	-
				1	0.9540	-0.0016	-0.17	0.0002	0.0071	0.0002	0.0073	0.00002	2.92	2.34
				2	0.8629	-0.0054	-0.63	0.0005	0.0147	0.0003	0.0152	0.00003	3.21	2.33
				3	0.7457	-0.0107	-1.44	0.0007	0.0217	0.0005	0.0223	0.00003	2.71	2.31
				4	0.6152	-0.0161	-2.62	0.0009	0.0280	0.0006	0.0284	0.00003	1.63	2.29
				5	0.4832	-0.0200	-4.14	0.0010	0.0329	0.0007	0.0330	0.00003	0.53	2.26
2		0.1	0	0	1	-	-	-	-	-	-	-	-	-
				1	0.9240	-0.0058	-0.63	0.0003	0.0109	0.0002	0.0110	0.00003	0.24	2.27
				2	0.7795	-0.0166	-2.13	0.0007	0.0208	0.0005	0.0207	0.00003	-0.70	2.24
				3	0.6065	-0.0277	-4.56	0.0009	0.0272	0.0006	0.0271	0.00003	-0.38	2.24
				4	0.4336	-0.0341	-7.86	0.0010	0.0300	0.0007	0.0303	0.00003	0.88	2.27
				5	0.2823	-0.0334	-11.84	0.0009	0.0289	0.0007	0.0296	0.00005	2.36	2.31
			1	0	1	-	-	-	-	-	-	-	-	-
				1	0.9532	-0.0040	-0.42	0.0002	0.0074	0.0002	0.0076	0.00003	2.63	2.33
				2	0.8598	-0.0122	-1.42	0.0005	0.0156	0.0004	0.0159	0.00003	2.24	2.31
				3	0.7384	-0.0224	-3.04	0.0007	0.0229	0.0005	0.0234	0.00003	2.05	2.30
				4	0.6024	-0.0317	-5.26	0.0009	0.0291	0.0007	0.0296	0.00003	1.90	2.29
				5	0.4643	-0.0372	-8.00	0.0011	0.0332	0.0007	0.0337	0.00003	1.75	2.29
3		0.2	0	0	1	-	-	-	-	-	-	-	-	-
				1	0.9213	-0.0156	-1.70	0.0004	0.0119	0.0003	0.0120	0.00003	0.85	2.28
				2	0.7696	-0.0378	-4.91	0.0007	0.0226	0.0005	0.0226	0.00003	-0.14	2.25
				3	0.5862	-0.0548	-9.35	0.0009	0.0292	0.0007	0.0292	0.00003	-0.05	2.25
				4	0.4025	-0.0591	-14.68	0.0010	0.0312	0.0007	0.0314	0.00004	0.42	2.26
				5	0.2444	-0.0495	-20.26	0.0009	0.0283	0.0006	0.0286	0.00006	0.82	2.28
			1	0	1	-	-	-	-	-	-	-	-	-
				1	0.9515	-0.0113	-1.19	0.0003	0.0082	0.0002	0.0085	0.00003	3.74	2.35
				2	0.8532	-0.0297	-3.48	0.0005	0.0172	0.0004	0.0178	0.00003	3.54	2.34
				3	0.7234	-0.0485	-6.71	0.0008	0.0251	0.0006	0.0258	0.00003	2.96	2.32
				4	0.5758	-0.0616	-10.70	0.0010	0.0312	0.0007	0.0318	0.00003	1.99	2.30
				5	0.4255	-0.0644	-15.13	0.0011	0.0344	0.0008	0.0347	0.00004	0.99	2.27

SE Standard Error
*in model SE

observed in the placebo group when compared to the treatment group. Levels of bias increased as survival probabilities were predicted for longer lengths of follow-up. Survival probability predictions for model 1, which represents the weakest quadratic association, were minimally biased at year 1 (-0.22% in placebo and -0.17% in treatment group). Lower levels of bias were observed in the treatment group at year 2 (-0.63%) than for the placebo group (-0.85%). Survival probabilities at year 5 were the most biased with a percentage bias of -5.83% and -4.14% in the placebo and treatment groups respectively. Subsequent predicted survival estimates from models with greater misspecification were incrementally more biased. Model 2 produced survival probabilities at years 1 through to 5 which were more biased than model 1. The largest level of bias for Model 2 was observed at year 5 (-11.84% and -8.00% for placebo and treatment groups respectively). Survival probabilities from Model 3 were the most biased. A bias of greater than 10% was observed at years 4 and 5. Year 4 survival probabilities were underestimated by -14.68% in the placebo group and -10.70% in the treatment group, whilst at year 5 the percentage bias observed was -20.26% and -15.13% in the placebo and treatment groups respectively.

Empirical standard error estimates show greater variation for predictions for the placebo group than treatment group. Variation increased for each year survival was predicted across all models. The model standard errors systematically missed targeting the empirical standard error. Generally model SEs were overestimated, slightly more so for the treatment group than the placebo. For model 1, the largest relative percentage error in model SE was seen at year 2 (3.03%, MC SE: 2.32 and 3.21% MC SE: 2.33 in the placebo and treatment group respectively). Whilst the smallest relative percentage error in model SE was observed at year 5 (0.35%, MC SE: 2.26 and 0.53% MC SE: 2.26 for placebo and treatment

group respectively). A greater quadratic association resulted in a smaller relative percentage error in model SE for the placebo group but larger for the treatment group. For Model 2 model SEs were over estimated by less than 1% for years 1 and 4 in the placebo group, but underestimated by less than 1% for years 2 and 3 in the placebo group. On the other hand the relative percentage error in model SE at year 5 in the placebo group was 2.36% (MC SE: 2.31). In the treatment group the relative percentage error in model SE ranged from 1.75% to 2.63% at years 1 to 5. The relative percentage error in model for survival predictions from model 3 were similar to those form model 2, with the exception of estimates at year 5. Model SEs were underestimated by less than 1% in the placebo group at years 2 and 3 and overestimated by less than 1% at years 1,4 and 5. In the treatment group the model SE missed targeting the empirical SE at time-points 1 to 5, resulting in a relative percentage error in model SE ranging from 0.99% at year 5 to 3.74% at year 1.

Figure 9.4 shows the distribution of the difference between predicted and true survival predictions for each model in Scenario 1. The spread in differences tends to increase as follow-up time increases under treatment, whilst at follow-up times 1 and 2 years there was less variation in the difference between predicted and true survival for the treatment group when compared with the placebo group.

9.4.2 Scenario 2

For each α_2 specified in Scenario 2, a total of 984 of 1000 models converged successfully. Table 9.2 presents results for Scenario 2, assuming a decrease in the biomarker is related to an increase in the risk of an event. For these models the true linear association parameter was $\alpha_1 = -0.6$ and the quadratic association parameters were $\alpha_2 = -0.05, -0.2$, and -0.3 for models 1, 2 and 3 respectively.

TABLE 9.2. Simulation results of bias, empirical standard error and relative % error in model SE for predicted survival probabilities from misspecified models in Scenario 2

Model	α_1	α_2	Treatment group	Time, years	True survival probability	Bias	% Bias	Monte Carlo SE	Empirical SE	Monte Carlo SE	Average model SE	Monte Carlo SE	Relative % error*	Monte Carlo SE
1	-0.6	-0.05	0	0	1	-	-	-	-	-	-	-	-	-
				1	0.9597	0.0016	0.17	0.0002	0.0075	0.0002	0.0077	0.00003	3.10	2.36
				2	0.8910	0.0045	0.51	0.0005	0.0142	0.0003	0.0145	0.00004	2.18	2.32
				3	0.8143	0.0070	0.86	0.0006	0.0199	0.0004	0.0202	0.00004	1.64	2.30
				4	0.7383	0.0084	1.13	0.0008	0.0256	0.0006	0.0259	0.00004	1.33	2.29
				5	0.6673	0.0086	1.29	0.0010	0.0314	0.0007	0.0318	0.00004	1.16	2.28
			1	0	1	-	-	-	-	-	-	-	-	-
				1	0.9753	0.0011	0.12	0.0002	0.0048	0.0001	0.0052	0.00002	7.19	2.47
				2	0.9324	0.0033	0.35	0.0003	0.0098	0.0002	0.0105	0.00003	6.67	2.43
				3	0.8829	0.0053	0.60	0.0005	0.0148	0.0003	0.0156	0.00004	5.57	2.39
				4	0.8319	0.0067	0.81	0.0007	0.0201	0.0005	0.0209	0.00004	4.24	2.36
				5	0.7824	0.0073	0.93	0.0008	0.0256	0.0006	0.0264	0.00005	3.04	2.33
2		-0.2	0	0	1	-	-	-	-	-	-	-	-	-
				1	0.9617	0.0048	0.50	0.0002	0.0071	0.0002	0.0071	0.00003	0.92	2.32
				2	0.8982	0.0131	1.46	0.0004	0.0133	0.0003	0.0133	0.00004	0.07	2.28
				3	0.8295	0.0199	2.40	0.0006	0.0189	0.0004	0.0188	0.00004	-0.34	2.26
				4	0.7636	0.0233	3.05	0.0008	0.0245	0.0006	0.0244	0.00004	-0.45	2.25
				5	0.7043	0.0230	3.26	0.0010	0.0303	0.0007	0.0302	0.00005	-0.37	2.25
			1	0	1	-	-	-	-	-	-	-	-	-
				1	0.9766	0.0028	0.29	0.0002	0.0048	0.0001	0.0048	0.00003	-0.83	2.30
				2	0.9370	0.0080	0.85	0.0003	0.0098	0.0002	0.0097	0.00004	-1.75	2.25
				3	0.8928	0.0124	1.39	0.0005	0.0146	0.0003	0.0144	0.00004	-1.52	2.24
				4	0.8491	0.0150	1.77	0.0006	0.0195	0.0004	0.0194	0.00005	-0.78	2.25
				5	0.8085	0.0151	1.87	0.0008	0.0245	0.0006	0.0245	0.00005	0.06	2.27
3		-0.3	0	0	1	-	-	-	-	-	-	-	-	-
				1	0.9630	0.0059	0.61	0.0002	0.0068	0.0002	0.0069	0.00003	0.51	2.32
				2	0.9027	0.0153	1.69	0.0004	0.0132	0.0003	0.0128	0.00004	-2.59	2.22
				3	0.8389	0.0220	2.62	0.0006	0.0191	0.0004	0.0182	0.00005	-5.14	2.15
				4	0.7791	0.0240	3.08	0.0008	0.0254	0.0006	0.0236	0.00005	-7.09	2.10
				5	0.7267	0.0210	2.89	0.0010	0.0320	0.0007	0.0294	0.00006	-8.20	2.08
			1	0	1	-	-	-	-	-	-	-	-	-
				1	0.9774	0.0037	0.37	0.0001	0.0046	0.0001	0.0046	0.00003	-0.34	2.32
				2	0.9398	0.0099	1.05	0.0003	0.0094	0.0002	0.0092	0.00004	-2.82	2.23
				3	0.8989	0.0147	1.63	0.0005	0.0144	0.0003	0.0137	0.00004	-4.57	2.17
				4	0.8595	0.0166	1.93	0.0006	0.0196	0.0004	0.0185	0.00005	-6.09	2.13
				5	0.8239	0.0152	1.84	0.0008	0.0253	0.0006	0.0235	0.00006	-7.43	2.10

SE Standard Error
*in model SE

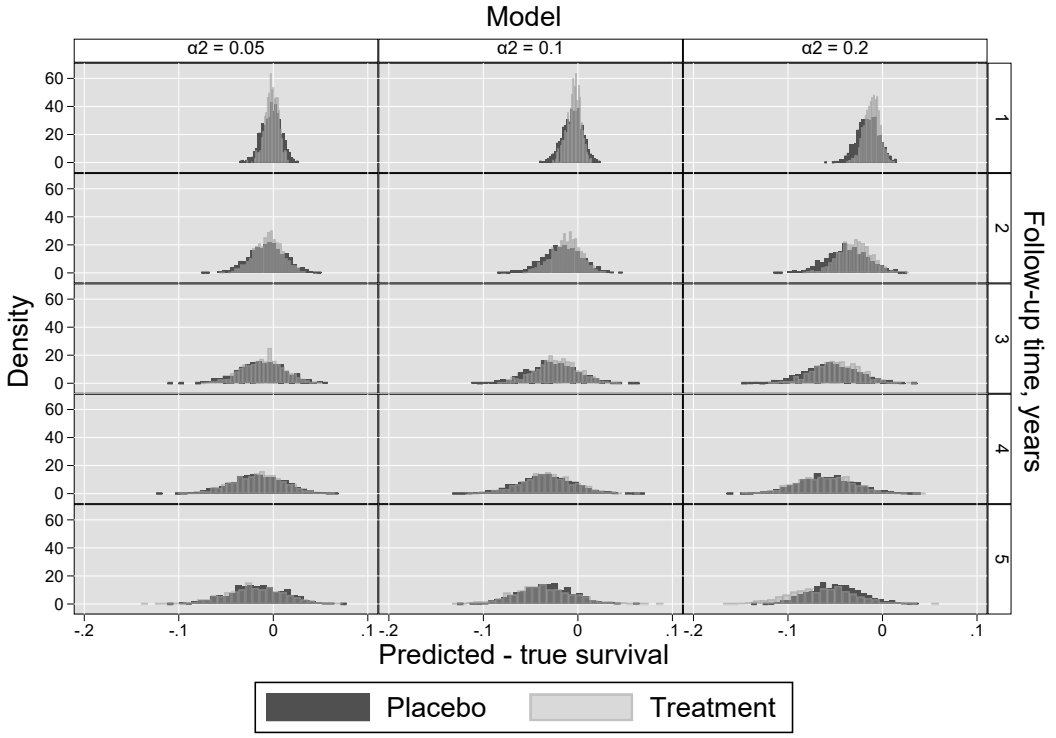


FIGURE 9.4. Difference between predicted and true survival probabilities for each model in Scenario 1, by treatment group (dark grey: placebo, light grey: treatment) and time-point)

Overall, misspecified models in Scenario 2 overestimated survival probabilities at year 1 through to 5. Survival predictions at baseline were unbiased across all models. For the model with the weakest quadratic association, $\alpha_2 = -0.05$, the predicted survival probabilities were the least biased. Bias estimates were less than 1% up to year three and year 4 in the placebo and treatment groups respectively. At year 5 percentage biases of 1.29% in the placebo group and 0.93% in the treatment group were observed. Predictions of survival for Model 2 were more biased for the placebo group than the treatment group. The largest level of bias in the placebo group was 3.26% at year 5 and correspondingly 1.87% also at year 5 in the treatment group. Bias estimates for Model 3 were similar to that

of Model 2. Percentage bias for each group increased from 0.61% (placebo) and 0.37% (treatment) at year 1 to 2.89% (placebo) and 1.84% (treatment) at year 5. Again survival predictions were more biased in the placebo group than the treatment group. The greatest % bias was observed at year 4, 3.08% and 1.93% for placebo and treatment groups respectively.

The empirical standard error estimates demonstrated slightly greater long-run variation for estimates for the placebo group than treatment group. Variation increased for each year survival was predicted across all models. Model SEs missed targeting the empirical SE for all models to varying degrees. For model 1, which included the smallest quadratic effect $\alpha_2 = -0.05$, the relative percentage error in model SE was larger in the treatment group than in the placebo. For both groups the model SEs were overestimated with larger error at year 1, and decreasing to the lowest at year 5. The relative percentage error in model SE was 3.10% (MC SE: 2.36) in the placebo group at year 1 decreasing to 1.16% (MC SE: 2.28) at year 5. In the treatment group the relative percentage error in model SE was 7.19% (MC SE: 2.47) at year 1 decreasing to 3.04% (MC SE: 2.33) at year 5. Model 2 underestimated model SEs at years 3, 4 and 5 in the placebo group and years 1, 2, 3 and 4 in the treatment group. The smallest relative percentage error in model SE was observed at year 2 (0.07%, MC SE: 2.28) in the placebo group and at year 5 (0.06% MC SE: 2.27) in the treatment group. Model SEs were underestimated by less than 1% in the placebo group and less than 2% in the treatment group across the yearly predictions. Finally the largest relative percentage error in model SE was observed for predictions of survival for model 3, for which $\alpha_2 = -0.3$. Generally model SEs were underestimated. The percentage error was fairly small at year 1, - 0.51% (MC SE: 2.32) in the placebo group and -0.34% (MC SE: 2.32) in the treatment group. The largest relative

percentage error in model SE was observed at year 5, -8.20% (MC SE: 2.08) and -7.43% (MC SE: 2.10) in the placebo and treatment groups respectively.

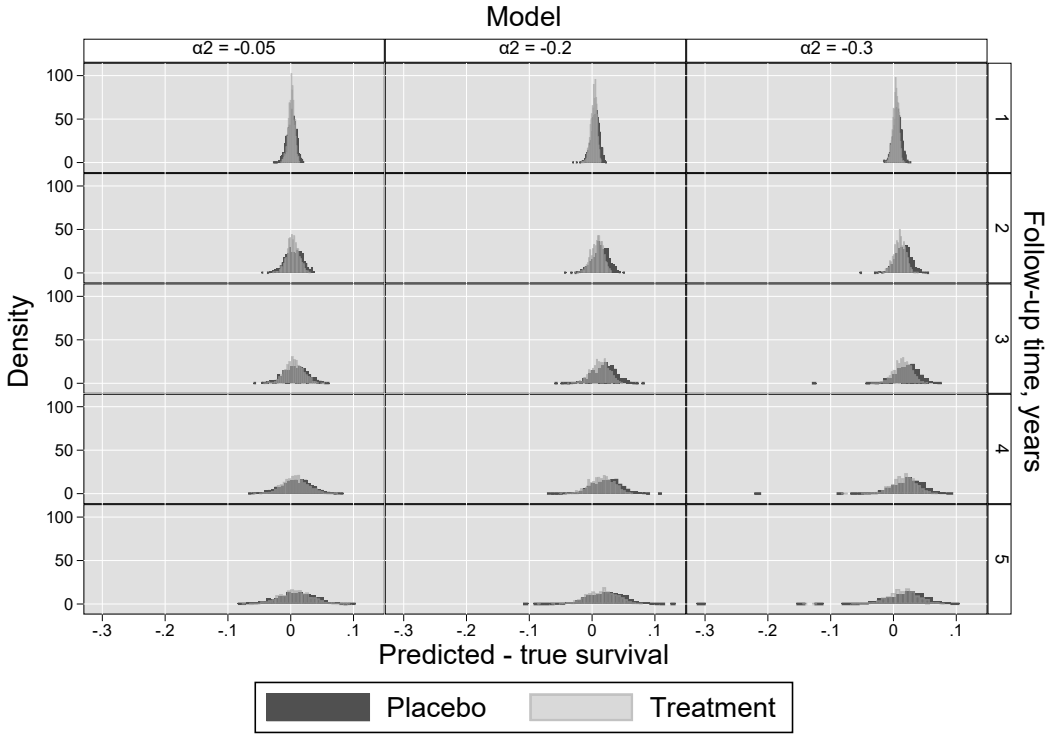


FIGURE 9.5. Difference between predicted and true survival probabilities for each model in Scenario 2, by treatment group (dark grey: placebo, light grey: treatment) and time-point)

Figure 9.5 shows the distribution of the difference between predicted and true survival predictions for each model in Scenario 2. As with Scenario 1, the variation in differences between predicted and true survival increase as follow-up time increases. The spread of differences was generally similar across treatment groups, although at year 1 and 2 there was less variation in differences for the treatment group.

9.5 Discussion

When implementing the joint longitudinal-survival model, often simplistic assumptions about the nature of the dependency between the survival and longitudinal outcomes are made. In practice, most likely due to convention, a linear association structure is assumed with little consideration for the possibility of a non-linear dependency between biomarker and time-to-event outcome. This is despite consideration of non-linear effects for baseline covariates being the norm. In this simulation study the effect of misspecifying a non-linear association structure was investigated. It has been shown that, for particular cases of quadratic hazard-longitudinal trajectory associations, bias is introduced when predicting survival from misspecified models which incorrectly assume a linear association. Two scenarios were evaluated for misspecification in this simulation study; a quadratic association between the hazard and a biomarker for which higher values are detrimental and conversely a biomarker for which lower values infer greater risk. For Scenario 1 larger levels of bias were observed at later follow-up times, when compared with the first two years of follow-up. Survival probabilities were consistently underestimated, with more biased survival predictions observed when a linear association structure was fitted to simulated data which assumed a greater

strength of quadratic association. This pattern was not as pronounced for Scenario 2, for which survival predictions from misspecified models were generally equally biased across models. Even so, increase in bias was observed between models 1 and 2. Survival probability predictions were generally less biased for the treatment group as opposed to the placebo group for both scenarios. Model SEs consistently missed targeting empirical SEs, suggesting bias was introduced in the estimation of model SEs. In Scenario 1, model SEs were generally overestimated, with the largest levels of error introduced for misspecified models fitted to simulated data assuming the smallest quadratic effect. Model SEs were overestimated to a greater extent in the treatment group. Larger levels of error in the estimation of model SE were observed for Scenario 2, particularly in the treatment groups for misspecified models 1 and 3.

9.5.1 Strengths and limitations

This simulation study is by no means exhaustive, however it does represent a novel investigation into non-linear association structures in the context of joint models. Non-linear effects are routinely considered when developing models for prognostic or prediction purposes. Yet in the case of a biomarker, modelled via a longitudinal trajectory, questions about non-linear associations are rarely asked. This study shows that ignoring non-linearity by incorrectly specifying the association structure in the survival submodel, introduces bias into the estimation of subsequent survival predictions, as well as their corresponding standard errors. Previous simulation studies have focussed on the consequences of misspecification in relation to model estimates. Here an alternative approach was taken to evaluate model misspecification by investigating the impact on model predictions. Joint longitudinal-survival models are very often implemented for their unique predictive capabilities, first because they allow predictions of survival which depend

on the (appropriately modelled) longitudinal biomarker of interest, and secondly allow predictions at the individual level when the BLUPs for the random effects are estimated. With such widescale use, it is important to consider how misspecification can affect resulting predictions, especially as a misspecified model does not necessarily lead to both biased model estimates and model predictions. This study evaluated how robust survival predictions, setting random effects to 0, are to misspecification.

There are several issues which arose in the course of this simulation study. For one the data generating mechanism was based on the PBC dataset of 312 individuals. This is smaller than the 500 individuals for which data was ultimately simulated. Longitudinal and survival models were simplified as much as possible, assuming a linear biomarker trajectory and a Weibull baseline hazard, to encourage model convergence both for linear and non-linear association structure models. It could be argued that a Weibull baseline hazard is unlikely to be flexible enough to capture the baseline hazard in real-life data. Certainly biomarkers rarely follow a linear trajectory, requiring non-linear terms such as polynomials or splines to be modelled appropriately (see section 7.5.3). These models are computationally difficult to estimate, and adding non-linear effects increases this burden. Small sample sizes and inadequate power can prohibit correct joint model model specification.

For each of Scenarios 1 and 2 a small percentage of models did not converge (1.0% and 1.6% respectively). In each case the estimates of performance were obtained for only those models which converged successfully, in what is termed a 'pure method evaluation' [348]. An alternative approach would be to continue to fine-tune the model estimates until they converged, e.g. for example adjusting quadrature points to achieve convergence [52]. This would be difficult to achieve

in a simulation study setting. For a complex hazard function as generated for this simulation it was difficult to identify just one set of parameter values which allowed models to converge the majority of the time. In fact, further data generating mechanisms for a cubic association structure were defined, however for this scenario the majority of models did not converge. Ultimately, it is likely that the sample size did not support complex non-linear associations which increase the complexity of the hazard function and so computational burden. A larger sample size is necessary then when the main interest is to investigate the impact of incorrect functional form in the joint model setting.

Original attempts at defining the data generating mechanism relied on the General Cycle Collection data described in Chapter 7. This dataset, though similar in size to the PBC dataset, is unlikely to resemble scenarios outside of a clinical trial or assisted pregnancy setting. As previously discussed in Chapter 7 the GCC dataset is singular in terms of the frequency and length for which hCG observations were collected. This is compounded by the addition of delayed entry, which is necessary to accommodate women falling pregnant and then becoming at risk of loss. A delayed entry model then requires an additional set of numerical integration on top of that which is required for the standard joint model [52]. The added features of this data and model made it difficult to fit non-linear associations, hence the use of PBC dataset as the basis for data simulation.

True model values were selected by fitting separate longitudinal and survival models to the PBC dataset. No further covariates were considered for either submodel, other than a treatment variable for the survival model. Depending on the medical application, there will be any number of variables associated with one or both outcomes. In the presence of censoring it has been shown that omission of important covariates can introduce bias into treatment effect estimates [349].

Model selection procedures for joint models, discussed briefly in section 5.10, are currently still in their infancy but should be considered when developing a model [225].

A simulation study can never cover all possible scenarios, and here the focus was on a quadratic non-linear association structure. Other alternative specifications of non-linear association structures are of course possible, however as they are by definition non-linear similar issues would be expected in those scenarios too. The quadratic associations evaluated in this simulation study followed a J shape, however U shaped associations with turning points are not uncommon. BMI and blood pressure risk profiles are characterised by turning points, with very high and low observations indicating an increase in the hazard rate.

9.5.2 Extensions

The focus here was on the misspecification of the non-linear association. A core component of defining the non-linear hazard function is the longitudinal submodel. As previously noted the joint model has been shown to be relatively robust to misspecification in the context of the current value association, however less so when the slope association is specified [52]. In light of this, misspecification of the longitudinal model is likely to have an impact on parameter estimates when a complex non-linear association structure is assumed. This would represent an interesting extension to the simulation study.

For this simulation survival probabilities were considered independently of the random effects. With conditional survival probabilities an important output of the joint model (see section 5.11.2) it would be of interest to extend the simulation to also evaluate the impact of misspecification on marginal survival predictions, which are estimated by integrating out the variances of the random-effects. This allows estimation of the BLUPs for the random effects, capturing variability at

the subject level, ultimately allowing for individualised predictions. If bias is introduced into estimation of the variance components, then resulting subject-specific probabilities would not be fit for use.

Further non-linear association structures should also be evaluated for misspecification, though the complexity of fitting such models makes this a challenging task. Certainly a larger dataset from which to base the simulation on, and ultimately a larger simulated dataset would improve model convergence.

9.5.3 Conclusions

This simulation study presents evidence that misspecification of a non-linear association structure can introduce bias into the estimation of predicted survival probabilities. However this is based on evaluating misspecification of only a quadratic association structure. The computational burden of fitting models with complex non-linear associations means it is unsurprising that they have not yet been addressed. Additional scenarios of non-linear associations would be useful to further generalise the results of this simulation study and provide further guidance. When fitting a joint longitudinal-survival model it is important to consider whether a non-linear association structure could explain the dependency between time-to-event and longitudinal outcomes. A simpler, more stable model is always encouraged, but sensitivity analyses are advised to avoid a misspecified model and subsequently biased predictions.

Chapter 10

Discussion

10.1 Chapter Overview

This chapter will summarise the work carried out in this thesis and how this fulfils the aims that were laid out in section 1.8. The overarching aim was to apply both established and cutting-edge joint modelling techniques to novel pregnancy datasets, in order to answer several hypotheses related to the early pregnancy setting. In doing so consequences of a key modelling assumption made when fitting the joint longitudinal-survival model was evaluated via a simulation study, namely that the association between the biomarker trajectory function and survival is linear, when this may not truly be the case. The strengths and limitations of the thesis will be discussed and areas of future work will be considered.

10.2 Summary of the thesis

The motivation for this thesis evolved from the General Cycle Collection study data introduced in section 7.3. The novel study, conducted by SPD Development Company Ltd, prospectively followed over 2000 women as they attempted to conceive. This presented a rare opportunity to access an intensive collection of daily urinary hormone observations on a scale which has yet to be seen in publicly funded clinical trials. More importantly the relationship between urinary hCG and early pregnancy loss, discussed in Chapter 2, is well-established; though

modelling the two outcomes simultaneously has received little attention up to now. This is largely because the joint longitudinal-survival model, introduced in Chapter 5, is a relatively recent development in statistics. The modelling technique combines conventional statistical models (see Chapters 4 and 3) to model the dependency between longitudinal and time-to-event data, where the repeatedly measured biomarker can feasibly be used as a surrogate for the survival outcome. The shared random effects parameter framework addresses issues of biomarker measurement error, intermittent observations and uncertainty. They also provide a foundation for individualised risk prediction, an important joint model extension which has yet to be addressed in the early pregnancy outcome literature.

The joint model was first applied to a smaller dataset, similar to the GCC data, which was previously analysed with the naive but perhaps more accessible two-stage model approach (see section 5.3.3) [44]. The analysis, presented in Chapter 6, was an opportunity to compare estimates for naive methods against the joint model (see 6.14). It also served as a learning experience, allowing exploration of the best way to model the longitudinal trajectory, the baseline hazard of the survival submodel and the impact this had on conditional survival predictions. When the two outcomes were simultaneously modelled via the joint model framework, a unit-increase in log hCG was associated with a 66% decrease in the rate of miscarriage, which was similar to the two-stage model estimate though with a larger, more appropriate standard error, but a larger effect than when modelled by the time-varying covariate approach. It was found during modelling, however, that the quadratic did not adequately model the longitudinal trajectory of log hCG, as the predicted decrease in log hCG came too early for healthy pregnancies. An additional quadratic random slope effect may have allowed for this

variation but could not be fitted at the joint modelling stage (see section 6.5.2). The survival submodel included a simpler Weibull baseline hazard, as opposed to a flexible baseline hazard modelled using restricted cubic splines, due to problems with model convergence. As a result the model was likely misspecified and resulting conditional survival probabilities were possibly biased and predicted with greater levels of uncertainty (see section 6.4.8). This analysis brought to light the complexity of the joint model and the need for a sufficient sample size for computation. All of this information informed the analysis of the GCC data.

The GCC data was analysed in Chapter 7. The approach to modelling aimed to address issues encountered in the previous analysis, in larger and more detailed data. Joint model calibration and discrimination measures were also considered. Clinical knowledge from those working in the pregnancy field was combined with standard model selection procedures to build trajectory and survival submodels which both provided good fit, but also included known confounding variables. Here improvements were made by modelling the longitudinal trajectory with restricted cubic splines, which better modelled the tails of the data. An interesting and possibly useful association between miscarriage and two other hormones, FSH3 and P3G, was identified during the modelling process, however this still corresponded to a 50% reduction in the rate of miscarriage for a unit increase in log hCG. Furthermore the slope of hCG was identified as an important factor in early pregnancy outcomes. Subject-specific survival predictions obtained for viable, early and late losses emphasized the changeable nature of hCG and the importance of continued biomarker observations to minimise uncertainty around predictions. Survival probabilities tended to predict the correct outcome even for a ten-day window. Establishing the optimal timing of observations was investigated, with a two to three-day window between updates to the longitudinal

trajectory giving enough time for some change to occur but not to the detriment of the pregnancy. This was highly dependent on the shape of the log hCG trajectory, however. Calibration and discrimination measures were introduced and implemented. Survival probability cut-offs for each five-day prediction window ranged from 0.65 to 0.98. In windows where there were fewer events the cut-off tended to be lower. Sensitivities and specificities were variable across the prediction windows, with sensitivities and specificities over 90% attained after 20 days of follow-up. The joint model was extended to a multiple marker context, modelling both log P3G and log hCG simultaneously with time-to-miscarriage. Ultimately, likely due to inconsistencies between the timing of P3G and hCG observations, increases in log P3G were found to increase the rate of miscarriage.

In Chapter 8 daily intercourse diary data, collected as part of the GCC study, was analysed relative to miscarriage. Acts in the luteal phase (post ovulation) and implantation window were associated with an increase in the rate of miscarriage, though these findings were not statistically significant. The timing of intercourse was further used as an (imperfect) proxy for sperm age (once in the female reproductive tract) and quality. If the last act recorded in the fertile window (assumed to have led to conception) was more in advance of the day of ovulation then the rate of miscarriage was decreased. Conversely more acts in the fertile window, assuming increased sperm quality, actually increased the rate of loss. Again neither of these associations were statistically significant. These results can be interpreted in two ways. Firstly, the study was not powered to find an association between timing of intercourse and miscarriage, with a number of assumptions made around recording of intercourse data, and in the definitions of the intercourse windows. Hence significant results should not be expected. Secondly, this adds to the current opinion that there is no evidence that intercourse

during pregnancy increases the likelihood of a loss [308]. No news therefore may be good news.

When including a continuous covariate in a survival model, the functional form is investigated to establish non-linear associations. For example, younger and older individuals may experience a higher event rate compared with individuals in between. This would require modelling age using polynomials or restricted cubic splines to capture turning points where the risk profile increases and then decreases. Joint model applications have tended not to deviate from the current value association structure, and the assumption that the relationship between a continuous biomarker and hazard is linear. The simulation study carried out in Chapter 9 investigated the effect of assuming linearity on predicted survival probabilities (setting random effects to zero) when the true association is quadratic. It was found that bias was introduced into the estimation of predicted survival probabilities for misspecified models, more so as the level of the quadratic association increased and at each subsequent follow-up time. Model standard errors were consistently over- or underestimated. The results of the simulation study demonstrated that survival predictions are not robust to misspecification of a non-linear association structure. Where possible, it is recommended that the possibility of a non-linear association structure is explored when fitting a joint model, just as non-linear covariates are. Sensitivity analyses should always be carried out to assess the validity of model assumptions.

10.3 Strengths and Limitations

The analyses presented in this thesis are the first examples of the application of the joint longitudinal-survival model to the early pregnancy outcome setting. The novel aspect comes from prospectively collected urinary hCG data from women

as they attempted to conceive, recorded daily for individuals, allowing for consistent capture of changes in relation to pregnancy progression. Evidence of an association between hCG and pregnancy loss is well-established in the literature, but has never been quantified utilising advanced joint modelling techniques [242; 350]. Where outcomes have been modelled in tandem, the hCG trajectory has been limited to a maximum of three observations or focussed on serum observations [45; 46]. In applying the joint model to this data, several modelling challenges were identified. This serves to emphasise the complexity of using joint models to analyse such data, and highlights the kind of challenges which analysts may encounter during modelling.

The analyses conducted in Chapters 6 and 7 have demonstrated the value of using urinary hCG to a greater extent in an early pregnancy setting. The improvements in accounting for error and uncertainty by implementing the shared parameter joint model framework over the two-stage model approach were well recognised [44; 241]. Software updates in recent years also made it possible to explore an alternative association structure and obtain subject-specific predictions to illustrate the dynamic monitoring capabilities of the joint longitudinal survival model [38; 200; 202]. Model calibration and discrimination measures are only now beginning to be implemented in software packages and have never been considered in a pregnancy outcome setting before now [43; 205; 290]. This thesis, though not all-encompassing, is an important step in prediction model building for early pregnancy outcomes.

There are several barriers to use of urinary hCG, however, in clinical practice. Urinary observations may be cheaper to observe, but at the scale observed in the GCC data would be impractical and costly, but for a subset of women who may be undergoing IVF and/or have experienced recurrent losses previously. In

natural pregnancy settings it is not usual to measure hCG regularly unless a loss is suspected, and certainly not from first detection of hCG (implantation). The day of conception would be unknown in a typical pregnancy and in reality clinicians work from the date of the last menstrual period to date a pregnancy rather than hCG. IVF pregnancies are tracked closely from implantation and so this type of hCG monitoring for viability is feasible and likely most beneficial in an assisted pregnancy setting. It is also possible that the General Cycle Collection study may have (inadvertently) recruited women who were struggling to conceive, perhaps due to the study provision of ovulation testing kits for targeting of intercourse. This represents sample selection bias and has implications for the generalisability of the results of analyses in this thesis.

The GCC study was a large scale longitudinal collection, with significant cost implications. Women were recruited via the Clearblue website and so were from the start more likely to adhere to an intensive sample collection regime. Volunteers who did not comply were subsequently withdrawn from the study by SPD Development Company Ltd. Furthermore, great effort was made to ensure full data collection through contacting volunteers for outstanding samples, confirming volunteer history and diary data. This dataset is therefore unlikely to be representative of the general population. In routine practice, a greater amount of missing data would be expected for the biomarker measurements and for covariate data.

NICE guidelines do not recommend hCG as a diagnostic tool for early pregnancy loss, but there could be scope for its use as a monitoring tool in conjunction with established methods such as ultrasound [66]. In addition, more consistent monitoring has the potential to establish a more well-rounded picture of pregnancy progression. During the course of this thesis the association between the

biomarker hCG and early pregnancy has been established and this comes under the scope of research priorities around miscarriage published by Prior *et al.* [105]. Once viability is determined and women move out of the first trimester, hCG may be less useful, however we have seen throughout this thesis the potential of hCG in early loss [11].

Pregnancy outcomes were generally predictable via subject-specific survival probabilities predicted for three women from the GCC data, and these certainly improved upon subject-specific predictions presented in Chapter 6. It was difficult to ascertain the ideal prediction window to truly illustrate the capabilities of the model. Though timing of risk prediction updates were crudely simulated, ultimately predicting for individuals whose data was used to build the model cannot evaluate model performance. Discrimination and calibration estimates were obtained, but these again were data dependent. To evaluate how well the model performs an external validation dataset would really be required to establish whether the risk of miscarriage can be predicted via hCG. Identifying a similar dataset would be challenging, given the singular nature of the GCC data. In the absence of an external dataset, however, internal validation could have been carried out via bootstrapping.

Several improvements to modelling could have been made during the course of analysis. All models were fitted on a complete case basis. Where models were fitted including P3G and FSH3 on the day of implantation, this resulted in a significant reduction in sample size from 367 to 312 individuals. This is because all samples were not tested for all hormones, and likely contributed to issues with model convergence. An alternative approach could be to impute missing P3G and/or FSH3 values. Secondly, the last day hCG was observed was assumed to be the day the miscarriage occurred. However, it may have been more

appropriate to assume interval censoring, assuming that the loss occurred between perhaps the date of the onset of a symptom, such as a bleed, and the last day hCG was observed. Finally, the hCG hormone can be elevated in circumstances unrelated to pregnancy, such as in menopause or in particular forms of cancer [351; 352; 353]. This could be modelled using a competing risks joint model, where in older women the two events are pregnancy or menopause, or in a more diverse age group pregnancy or cancer.

The issues which have been encountered across the analyses in this thesis can largely be attributed to the data. This is real-life data with its own quirks that ultimately do not fit into the mould of illustrative data presented in joint modelling tutorials. So, whilst joint models are attractive prospects for combining two types of outcome data with a built in framework for dynamic prediction, in practice they require a sample size which perhaps is prohibitive. Development of a joint model specific sample size equation by Chen *et al.* [354] has shown that power to detect an effect is impacted by several aspects of data collection; including the number of events observed, the timing of biomarker observation and the frequency of observation. In addition greater numbers of polynomial terms used to model the longitudinal trajectory in turn require a greater sample size for equivalent power [354].

Chen *et al.* have suggested that the 'ideal data collection strategy' for longitudinal biomarker data would require balancing the timing and frequency of measurements for a fixed follow-up period [354]. Though a longer follow-up time would lead to an increase in event numbers and in turn power, it is usual for follow-up to be pre-determined due to cost implications. Increasing the frequency of observations is a means of reducing the impact of within-subject variability or measurement error [354]. If measurement error is high, each individual would need

to contribute at least two measurements for a linear trajectory and more than four for a quadratic trajectory to maintain power and avoid biased estimates of the longitudinal effect [354]. An additional consideration should be the variance-covariance matrix, and whether this is known or unknown. If there is high heterogeneity with larger variances, sample size will likely be low, less heterogeneity will result in larger sample size requirements. Assuming the variance-covariance matrix is unknown results in reduced power against a known variance-covariance structure, even when the number of events increases [354].

The length of follow-up affects the number of events observed, but for the early pregnancy setting this is limited to the 13 weeks for which hCG is considered a relevant biomarker to predict early loss. Furthermore longitudinal observations were observed maximally, obtained from daily early morning urine where concentration of hCG would be the greatest. Urine could potentially have been collected at multiple times of day, but these would then need to be corrected for creatinine concentration, for comparability across urine samples of varying dilution. A ramped-up collection schedule would also increase the burden on the volunteer, most likely resulting in greater amounts of missing data. Furthermore, modelling the longitudinal trajectory using more complex restricted cubic splines would again reduce power to observe effects. Data collection can only be optimised then by collecting for a larger sample of volunteers. Though a substantial association between log hCG and miscarriage was observed in the GCC data, the number of confounding variables included in the survival submodel may be an example of overfitting. Certainly it is unlikely that the number of events support the inclusion of so many variables.

Joint longitudinal-survival models are complex in nature, requiring several levels of numerical integration (adaptive Gaussian quadrature), depending on the

model specification, to evaluate the likelihood [38]. This increases computational burden and for a small sample size, using too few quadrature nodes can prevent models from converging [214]. The analysis in Chapter 7 ultimately moved away from the frequentist joint model approach, in favour of Bayesian joint models, removing the need for numerical quadrature techniques entirely. Default vague priors in the JMBayes R package were utilised, however these could have been updated with informative prior distributions using the findings from the analysis conducted in Chapter 6 [221]. A Bayesian approach to model estimation also allows for a natural progression to the estimation of subject-specific conditional survival predictions using the approach presented by Rizopoulos [43].

The computational complexity of fitting a joint model to the pregnancy outcome data led to several compromises during the course of modelling. A less flexible but more stable baseline hazard was utilised in the survival submodels described in Chapters 6 and 7. Restricted cubic splines, however, have been advocated over distributional assumptions for capturing the baseline hazard [179; 182]. An appropriately modelled baseline hazard is important for accurate survival predictions, a key aspect of this thesis [178]. Misspecification of components of the joint model have been covered in Chapter 9, but to reiterate - a misspecified baseline hazard can lead to biased parameter estimates in the joint model [53]. Though improvements to the longitudinal model were made over the course of the two joint modelling analyses, an additional slope random effect could not be fitted. This meant only subject-specific deviations in linear slope were allowed for, possibly resulting in a misspecified longitudinal submodel. Bias may have been introduced into subsequent association estimates, particularly for the first derivative model fitted in Chapter 7 [52].

Model selection followed a combination of expert opinion and established model selection techniques, such as the AIC and BIC, for separate submodels. However, the influence of the individual components of the longitudinal submodel have on the survival submodel are not considered using this approach. An approach for simultaneous model selection proposes a partitioning of the joint model AIC (or BIC) into that for the longitudinal model and conditional (on the longitudinal submodel) survival submodel [223]. This is, however, currently only implemented in SAS. Given the number of modelling decisions made during joint model building, for both longitudinal and survival submodel and random effect selection, it is important sensitivity analyses are carried out to minimise model misspecification.

10.4 Future Work

10.4.1 Multiple markers

Several extensions could be made to the work presented in this thesis. For one this thesis has in the main considered only one pregnancy-specific biomarker in relation to early pregnancy loss. Yet, a number of biomarkers are key to the menstrual cycle and some key to maintaining a pregnancy. Progesterone was modelled along with hCG in Chapter 7 within a multivariate joint model framework. This met with limited success, due to the inconsistency in the timing of observations between the biomarkers. An extension to the Bayesian multivariate joint model which allowed for biomarkers measured at different and for varying numbers of time-points may result in altered findings. Progesterone encourages implantation and maintenance of the fertilised ovum, whilst falling levels can cause the pregnancy to be rejected [355]. High levels of progesterone then are indicative of a positive outcome, and several trials have investigated its benefits as a treatment

for women who experience recurrent loss [21; 80]. Progesterone in conjunction with hCG have been shown to assist in diagnosis of ectopic pregnancies [356], and there is evidence of each marker being useful to differentiate between viable and failing pregnancies in isolation [357]. However, there is less certainty about whether the markers together can be used for predicting pregnancy loss, perhaps as current studies have used significantly fewer observations than were available in the GCC study data [48; 280]. The Bayesian joint model formulation allows in theory informative prior distributions to be specified for the correlation between each of these longitudinal biomarkers.

10.4.2 Timing of observations

Subject-specific survival predictions from a joint longitudinal-survival model, have been discussed in the context of establishing the timing of observations. Extensive collection, such as that undertaken in the GCC study, is unlikely to be implementable outside of a trial setting. Therefore, work needs to be carried out to define the optimal collection schedule to capture significant changes in biomarker trajectory, whilst offsetting the cost and burden of collection. This was briefly investigated in Chapter 7, however a robust comparison of profiles with differing observation schedules is required to understand how often hCG should be measured for a given individual. Use of Bayesian optimal designs to select personalised screening intervals has been discussed in the literature [358]. This is based on maximising a utility function which estimates the difference between the predictive conditional distribution of the joint model with and without an additional biomarker measurement being observed. A larger deviation between the two distributions indicates an additional measurement should be observed imminently. A alternative approach proposes the use of decision modelling based on estimating optimal expected life-years [230].

10.4.3 Interval censoring

Throughout this thesis the assumption of right censoring has been made, however we could also think of losses as either occurring before the last observation or between two observation times. This would indicate either left-censoring or interval-censoring. The timing of loss was recorded as the date of the last hCG observation. By this point the miscarriage may have resolved, and so the loss most likely occurred prior to the censoring time. On the other hand, a biochemical pregnancy, may only be recognised by a declining hCG trajectory. In this case the loss may have occurred between the time at which hCG began to decrease and the last observed hCG measurement, resulting in interval censoring. Approaches to modelling interval censored time-to-event outcomes in a joint model setting have utilised both parametric and semi-parametric model approaches, with the former allowing for more robust and informative estimates in the presence of heavy interval-censoring [359; 360].

10.4.4 Joint longitudinal-survival hurdle models

The time since conception timeline has been a point of contention throughout this thesis. As a common anchoring point across women it does its job, but it also relies on the assumption that conception occurs the day after ovulation. This is likely reasonable, but even so is unlikely to hold for all women. The bigger issue is that though we assume conception on this day, pregnancy is not confirmed until hCG is observed above a certain detection limit, which leads to delayed entry. The joint longitudinal-survival hurdle model proposes a two-part hurdle model for the longitudinal biomarker, which first estimates the probability of exceeding the detection limit [361]. This could be a hCG level of 2mIU/ml for a lab test or even 25mIU/ml for a home pregnancy test. The second part estimates

the mean biomarker response conditional on having exceeded the detection limit. Threshold values are then excluded from the estimation of the mean response in the second part of the longitudinal modelling process. The resulting association structure with the survival submodel is based on both parts of the longitudinal hurdle model. This is an interesting extension to incorporate a common issue when measuring a biomarker.

10.4.5 Simulation study extensions

The simulation study presented in Chapter 9 could be extended in several ways. Firstly the effect of misspecification for further non-linear associations should be considered, although currently polynomials are the simplest to implement. This extension also relies on having a large enough dataset for which to simulate from to address issues around model convergence. The effect on marginal survival predictions could also be evaluated, as this would investigate the impact of misspecification on the variance components which estimate variability at the subject level. This feature is a main attraction of the joint model and bias introduced at this juncture would mean misspecified models are not fit for prognostic purposes. It may also be of interest to evaluate the impact of misspecifying the longitudinal submodel for a non-linear association structure, particularly as it has been shown that longitudinal model misspecification has an impact on estimates from a first derivative association structure [52].

10.5 Conclusion

In this thesis cutting-edge methods were utilised to quantify the association between longitudinal hCG and time-to-miscarriage. The joint longitudinal-survival model was uniquely applied to a rich dataset, with a view to allow prediction of subject-specific risk estimates. Each of the analyses carried out in Chapters 6 and

7 demonstrated the value of more consistent observation of urinary hCG measurements in an early pregnancy setting. The findings confirmed evidence of an association between longitudinal log hCG and time-to-miscarriage, and showed the potential of its use as a monitoring tool. In certain subsets of women where the collection burden is offset by need, for example for those undergoing IVF treatment or who have experienced recurrent loss, hCG tracking may be feasible to establish current risk of loss. This, however, would need to be implemented in conjunction with established diagnostic pathways such as ultrasound. There is clearly potential for urinary hCG to be utilised to a greater extent in practice than it currently is.

Joint models are complex in nature and require sufficient data to truly exploit the advantages of simultaneous modelling within a shared parameter framework. Where data allows, fitting a joint model accounts for uncertainty better than the two-stage model approach [28]. If prediction is an aim, then developments in software make individualised prediction and monitoring an accessible tool. However, the computational complexity of this software is also a barrier to implementation. In these instances, the two-stage model makes for a reasonable alternative, particularly as biological applications evolve in difficulty.

Appendix A

Nature Scientific Reports research paper

This Appendix contains a copy of the research paper titled "Jointly modelling longitudinally measured urinary human chorionic gonadotrophin and early pregnancy outcomes," which was published in Nature Scientific Reports under the DOI https://doi.org/10.1038/s41598-020-61461-w.

SCIENTIFIC REPORTS
nature research

OPEN

Jointly modelling longitudinally measured urinary human chorionic gonadotrophin and early pregnancy outcomes

N. B. Ashra[1*], L. Marriott[2], S. Johnson[2], K. R. Abrams[1] & M. J. Crowther[1]

Human chorionic gonadotrophin (hCG) is largely used to confirm pregnancy. Yet evidence shows that longitudinal hCG profiles are distinguishable between healthy and failing pregnancies. We retrospectively fitted a joint longitudinal-survival model to data from 127 (85 healthy and 42 failing pregnancies) US women, aged 18–45, who were attempting to conceive, to quantify the association between longitudinally measured urinary hCG and early miscarriage. Using subject-specific predictions, obtained uniquely from the joint model, we investigated the plausibility of adaptively monitoring early pregnancy outcomes based on updating hCG measurements. Volunteers collected daily early morning urine samples for their menstrual cycle and up to 28 days post day of missed period. The longitudinal submodel for log hCG included a random intercept and slope and fixed linear and quadratic time terms. The survival submodel included maternal age and cycle length covariates. Unit increases in log hCG corresponded to a 63.9% (HR 0.36, 95% CI 0.16, 0.47) decrease in the risk of miscarriage, confirming a strong association between hCG and miscarriage. Outputted conditional survival probabilities gave individualised risk estimates for the early pregnancy outcomes in the short term. However, longer term monitoring would require a larger sample size and prospectively followed up data, focusing on emerging extensions to the joint model, which allow assessment of the specificity and sensitivity.

Early miscarriage, defined in the UK as loss before week 13, is a frequent complication of pregnancy[1]. It affects 12% to 24% of clinically confirmed pregnancies, not counting those losses which occur prior to the date of the missed period - so-called biochemical pregnancies[2]. Women who suffer from a miscarriage are more likely to report symptoms associated with depression, with affected women ranging from 20% to a high of 55%[3]. Though the majority of losses are self-resolving, those that are not may require diagnostic tests, hospital treatment, surgical intervention and follow-up care[2]. This provides an incentive to identify potential early losses as early as possible by exploring more patient-centred monitoring strategies.

The recently published priorities for research within miscarriage ranked highest the identification of effective interventions to prevent miscarriage[4]. This encompasses the plausibility of using biomarkers to track pregnancy progression through viability or miscarriage. Several potential biomarkers have been identified to predict miscarriage, with human chorionic gonadotrophin (hCG) a strong contender[5]. The hormone tends to rise rapidly and reliably in early pregnancy, doubling every 1.5 days in the first 5 weeks post conception and then every 3.5 days from week 7, before plateauing around week 10[5,6]. Its use is more prevalent in tracking early pregnancy progress in an *in vitro* fertilisation (IVF) population and for identifying ectopic pregnancies[7]. However, evidence suggests that longitudinal profiles of hCG can be utilised to distinguish between viable and failing pregnancies, with similar patterns of hCG noted across maternal serum and urine[8].

The repeated collection of a continuous biomarker, such as hCG, over time gives rise to intermittently observed longitudinal data which are subject to measurement error[9,10]. Conventionally, this data is analysed using linear mixed effects models, with time-to-event outcomes analysed using survival models[11,12]. However, when interest lies in quantifying the association between the repeatedly measured biomarker and time-to-event outcome, separate analyses ignore the dependency between the longitudinal and time-to-event processes[13].

[1]Department of Health Sciences, University of Leicester, Leicester, UK. [2]Clinical Research Department, SPD Development Company Ltd., Bedford, UK. *email: nbba1@le.ac.uk

Variables	Healthy (n = 85)	Miscarried (n = 44)	Overall (n = 129)
Age, years	29.95 (4.15)	32.34 (4.60)	30.77 (4.44)
Ethnicity, n (%)			
White	75 (88.24)	34 (77.27)	109 (84.50)
Black	3 (3.53)	7 (15.91)	10 (7.75)
Asian	4 (4.71)	2 (4.55)	6 (4.65)
Mixed	3 (3.53)	1 (2.27)	4 (3.10)
Education, n (%)			
High School	4 (4.71)	2 (4.55)	6 (4.65)
Graduate	69 (81.18)	28 (63.64)	97 (75.19)
Postgraduate	12 (14.12)	14 (31.82)	26 (20.16)
Occupation, n (%)			
Homemaker	12 (14.12)	3 (6.82)	15 (11.63)
Student	1 (1.18)	1 (2.27)	2 (1.55)
Skilled labourer	2 (2.35)	2 (4.55)	4 (3.10)
Office admin	8 (9.41)	5 (11.36)	13 (10.08)
Professional	60 (70.59)	31 (70.45)	91 (70.54)
Other	2 (2.35)	2 (4.55)	4 (3.10)
Cycle length	29.94 (2.95)	28.66 (3.21)	29.50 (3.09)
Previous pregnancies	1.00 (1.05)	1.11 (1.15)	1.04 (1.08)
Previous live births	0.62 (0.76)	0.70 (0.88)	0.65 (0.80)
Time to conceive, months	4.36 (5.83)	4.55 (5.98)	4.43 (5.86)
Previous miscarriage, n (%)	11 (12.94)	4 (9.76)	15 (11.90)

Table 1. Baseline demographics by pregnancy viability group. All values are mean(SD) unless otherwise stated.

Acknowledging an association between a longitudinal biomarker and survival outcome implies that very high or low values of the biomarker are indicative of adverse outcomes[14]. Fitting a simple survival model to the event, including all of the longitudinal biomarker information, tells us how a change in biomarker value affects survival over follow-up time. However, the variation in biomarker observations between individuals is not incorporated into the model, so inferences for individuals cannot be drawn. Secondly, the implicit changes in biomarker values between each physically observed measurement are ignored, resulting in a failure to build a complete biomarker profile. The linear mixed effects model can build this biomarker trajectory and inbuilt random effects allow estimation of personalised risk as an output of the model. Recognizing the advantages of both types of model, combining both the linear mixed effects and survival models through a shared dependence structure via the joint longitudinal-survival model is essential. This allows the association to be appropriately modelled, whilst taking into account the intermittent nature of observations and measurement error. The model, through estimation of individual trajectories, can aid monitoring and potentially prediction of outcomes.

The aim of this paper is to retrospectively apply the classical joint model framework to data of pregnant women, who were followed up from before conception, to quantify the association between longitudinal urinary hCG observations and early miscarriage. The paper will also consider whether estimation of conditional survival probabilities from the joint model could provide the basis for dynamic monitoring of patients in the very early stages of pregnancy prior to other symptoms manifesting.

Results

A total of 44 (17.6%) women suffered miscarriages. The dataset used for analysis consists of 85 randomly selected viable pregnancies and 44 miscarried pregnancies. A summary of demographic variables is given in Table 1. Overall, the two groups were comparable. Women who experienced healthy pregnancies were slightly younger (mean $\pm$ SD: 29.95 ± 4.15) than those who miscarried (mean $\pm$ SD: 32.34 ± 4.60). The majority of women in either group were from a White European background (88.24% and 77.27% respectively). A slightly higher proportion of women who had viable pregnancies had previously experienced a miscarriage, compared to women who miscarried (12.94%, and 9.76%). Of the women who miscarried, 18 (14.2%) experienced biochemical pregnancies and 24 (57.1%) women suffered early miscarriages. Two women who miscarried did not contribute hCG measurements and were not included in the joint modelling analysis.

The remaining 127 women all contributed repeated hCG measurements. For women who miscarried the average number of hCG observations was 17.5 (SD: 8.6) and for women who experienced viable pregnancies the average number of measurements was higher at 23.6 (SD 3.9).

Profiles of log hCG measurements for viable and failing pregnancies are presented in Fig. 1. The general trajectory shows an initial rise after conception, which continues through the first three weeks of the healthy pregnancies before slowing in rise. There was greater variation in profiles for women who miscarried, who also

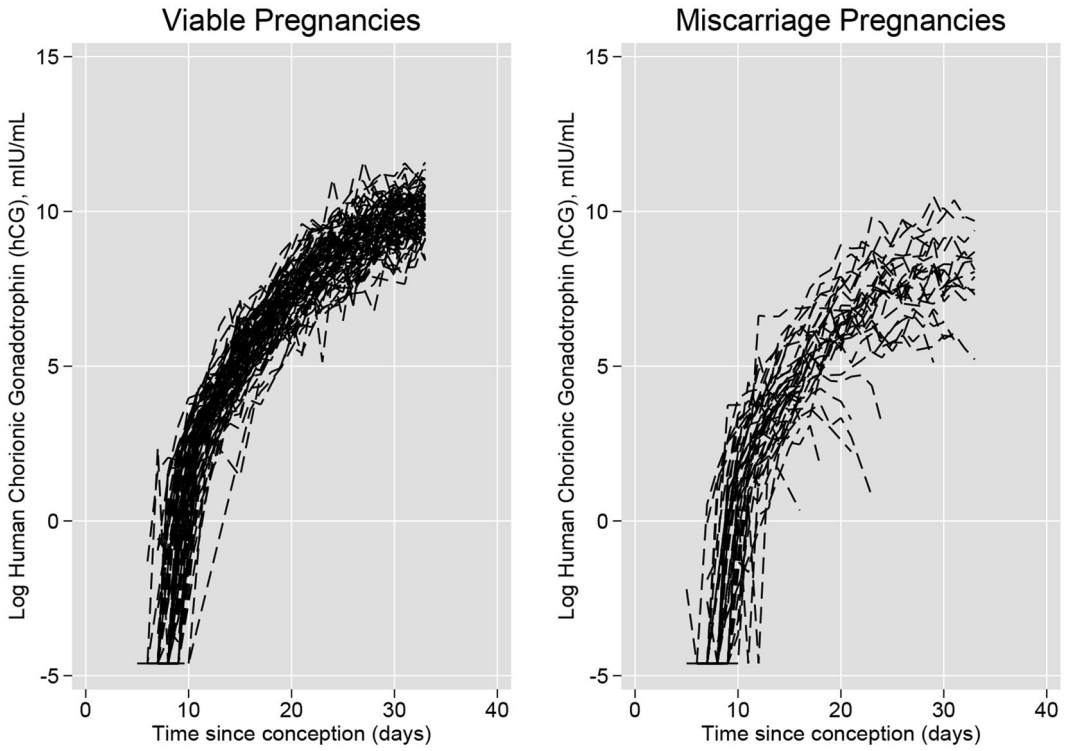


Figure 1. Log human chorionic gonadotrophin trajectories for viable pregnancies and miscarriage pregnancies.

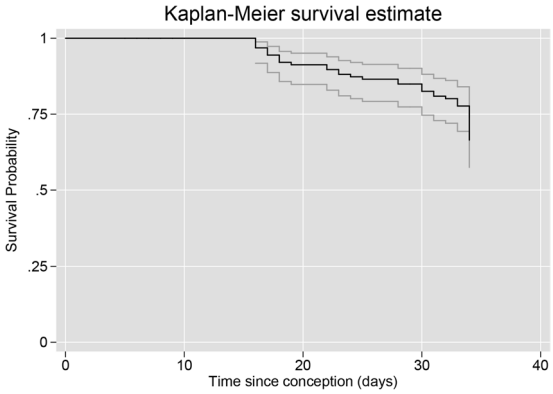


Figure 2. Kaplan-Meier survival probabilities for time-to-miscarriage.

presented with an initial rise after conception. However, some women experienced a sharp drop in hCG, whilst others experienced a more gradual rise in hCG in comparison with women who had healthy pregnancies.

Overall Kaplan-Meier survival estimates for time to miscarriage are shown in Fig. 2. An approximate 15 day lag is evident before an event is seen, due to the use of time since conception as a timeline.

Modelling longitudinal profile. Inclusion of a quadratic time variable was necessary to appropriately capture the shape of the log hCG profile. Results from an initial fitted linear mixed effects model, including a grouping variable for pregnancy outcome, confirmed that mean log hCG was −1.66 mIU/mL (95% CI −2.14, −1.18) lower in the biochemical pregnancy group and −1.13 mIU/mL (95% CI −1.48, −0.78) lower in the early miscarriage group, when compared with the healthy pregnancies. Results are presented in Table 2.

Joint longitudinal-survival model. A joint longitudinal-survival model was fitted to the data. Estimates for the model with current value association structure are given in Table 3. A unit increase in absolute value of log hCG corresponded to a 66.1% (HR 0.339, 95% CI 0.257, 0.447) decrease in the risk of miscarriage at time t. A one-year increase in maternal age at conception resulted in a 7.6% (HR 1.076, 95% CI 0.998, 1.159) increase in the risk of miscarriage. A one-day increase in cycle length was associated with a 15.6% (HR 0.844, 95% CI 0.739, 0.965) decrease in the risk of miscarriage.

Longitudinal model	Mean change in log hCG MIu/ml	95% Confidence Interval
Time since conception, days	1.431	1.396, 1.466
Quadratic time since conception, days	−0.025	−0.026, −0.024
Group		
Healthy	—	—
Biochemical loss	−1.656	−2.135, −1.176
Early loss	−1.132	−1.484, −0.781

Table 2. Model estimates from a linear mixed effects model.

Survival submodel	Hazard Ratio	95% Confidence Interval
Age, years	1.076	0.998, 1.159
Usual cycle length, days	0.844	0.739, 0.965
Expected current value of log hCG	0.339	0.257, 0.447
Longitudinal submodel	**Mean**	**95% Confidence Interval**
Time since conception, days	1.431	1.396, 1.466
Quadratic time since conception, days	−0.025	−0.026, −0.024

Table 3. Model estimates from a joint longitudinal-survival model with current value association structure.

Model	Standard error	Hazard Ratio for log hCG	95% Confidence Interval
Time-varying covariate	0.036	0.439	0.373, 0.516
Two-stage model	0.040	0.440	0.368, 0.527
Joint model	0.142	0.339	0.257, 0.447

Table 4. Survival estimates from a standard survival model with time-varying covariate, two-stage model and joint model.

A comparison of log hCG association parameters for various models are presented in Table 4. The association between log hCG and time to miscarriage was attenuated when fitting a survival model with time-varying covariate (HR: 0.439, 95% CI: 0.373, 0.516) and the two-stage model (HR: 0.440 95% CI: 0.368, 0.527). Furthermore, standard errors for both the standard survival model and two-stage model were 0.036 and 0.040 respectively compared to a larger 0.142 for the joint model.

Conditional survival predictions. Conditional survival probabilities were obtained from the joint model, which included the current value association structure (see Table 2). Probabilities estimated for the ten-day window after the last observed hCG measurement are shown in Fig. 3 and for a two-day window in Fig. 4. Participants A and B experienced biochemical and early losses respectively, whilst participant C experienced a healthy pregnancy. For both participants A and B a similar number of measurements were observed over comparable time periods, with similar average cycle lengths (28 and 30 days respectively) and ages (42 and 38 years respectively). Based on observed hCG measurements as well as age and cycle information, both were predicted to experience miscarriages.

For participant C estimates confirmed an 80% survival probability for the pregnancy two days post last observed hCG measurement. Depending on the cut-off used for low risk this may not be considered a high enough survival probability for a healthy pregnancy. As follow-up did not continue it was not possible to update probabilities to look at longer-term outcomes.

Discussion

Principal findings. This analysis builds on the two-stage model approach implemented by Marriott *et al.*[15]. By utilising the more advanced joint longitudinal-survival framework, the association between longitudinally measured urinary hCG and time to miscarriage is modelled, accounting for both measurement error and the intermittent nature of observations. This improves upon the two-stage model, which assumed that measurements remained constant between observation times.

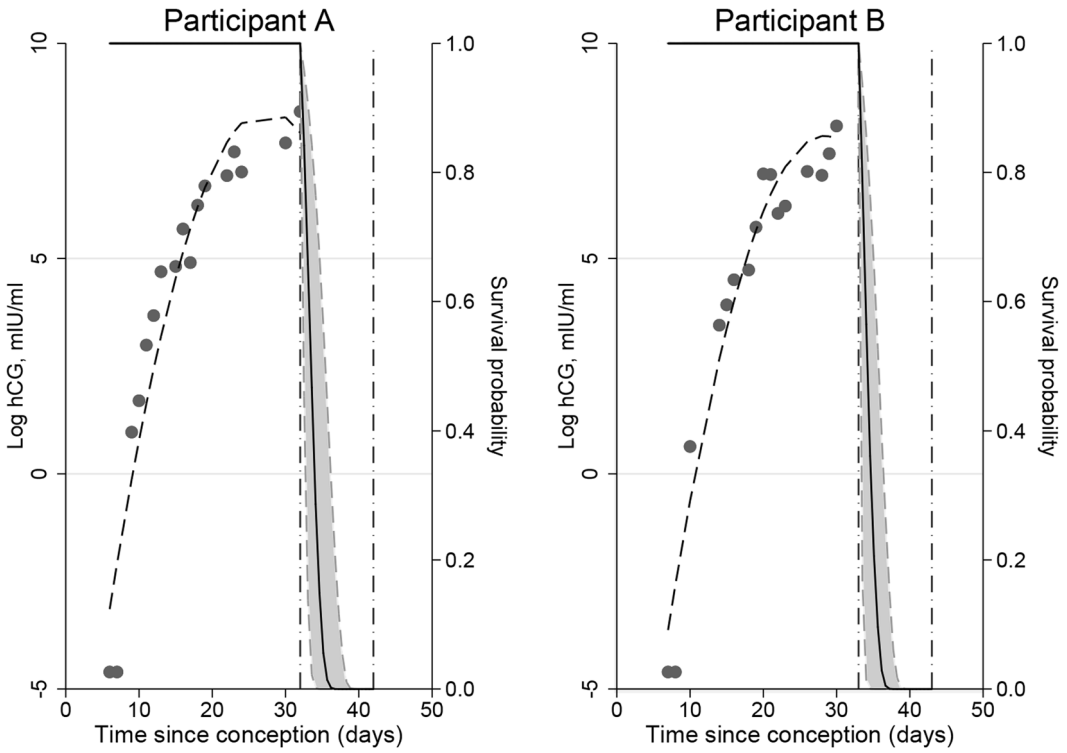


Figure 3. Conditional survival probability curves for participants A and B who experienced biochemical and early losses, respectively.

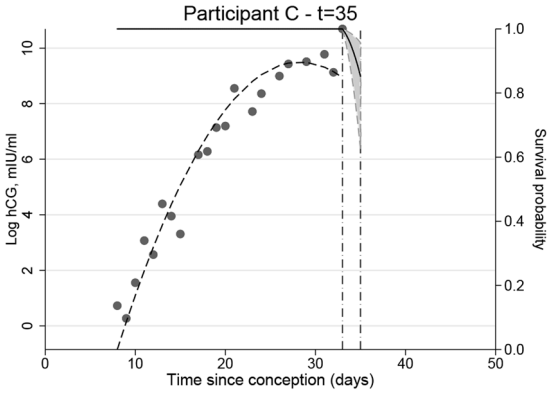


Figure 4. Conditional survival probability curve for participant C who experienced a healthy pregnancy.

With the emphasis now on personalised care, it is becoming standard practice to use the joint model in favour of singular or two-stage analyses to model the association between longitudinal and failure processes, to both maximise efficiency and minimise the potential for bias[16]. The mainstream use of joint models coincides with improvements in software making these complicated models increasingly easier to fit, with packages available in both R (JM, JoineRML) and Stata (`stjm, merlin`)[13,17–19]. This makes the estimation of conditional survival probabilities from such models more accessible.

This paper investigates whether urinary hCG could be used to monitor pregnancy viability prospectively in early pregnancy from first detection of hCG. Tracking at this early stage presents an adjunct to diagnosis by ultrasound later on in the pregnancy. This analysis echoes research suggesting declines in hCG can be noted even prior to other symptoms presenting[20]. There is also potential for this monitoring to occur prior to conception, with a recent study finding that a lag between the luteal phase and hCG production can be indicative of a biochemical pregnancy, possibly due to early or delayed implantation[21].

Tracking of hCG by pregnant women is practicable, as demonstrated by Foo *et al*. who employed a fertility monitor that also provide semi-quantitative analysis of hCG levels on pregnancy tests that were used daily in women who conceived[21]. Retrospective analysis of the semi-quantitative data indicated that non-viable pregnancies had different hCG profiles to viable pregnancies. Serial tracking could have the potential to cause stress,

although women using tests to track ovulation for fertility purposes do not appear to have higher stress levels than those not employing tests[22,23]. Nevertheless, it is likely that tracking would initially be of benefit in high risk pregnancies, where anxiety levels are already high and there would be a willingness and reason to track. Further research would be required to understand the psychological impact of tracking.

Monitoring from first detection has the potential to be useful in cases of recurrent miscarriage, particularly as research into treatment gains traction. A recently published feasibility study assessing the effectiveness of the diabetes drug sitagliptin as a treatment for recurrent miscarriage, presented promising findings[24]. This trial builds on previous research, which found that in some cases of recurrent miscarriage it is the deterioration of stem-like cells in the uterus which contribute to pregnancy loss. When adjusted for age and baseline colony forming unit (CFU) counts, the CFU count was higher (RR 1.52, 95% CI 1.32, 1.75) in the sitagliptin group compared to placebo, pointing to successful regeneration of cells. These findings could revolutionise treatment for unexplained recurrent miscarriage, particularly as the more established progesterone therapy has not been shown to significantly impact the rate of live births (PROMISE and PRISM trials)[25,26].

Not all miscarriage is likely to be predictable due to the diverse aetiology of the condition. Some causes can be directly related to reduced hCG levels, e.g. conditions that affect rate of embryonic development such as chromosomal abnormalities, or inadequate placentation. Other causes, for example, where infectious agents or trauma are involved, may have no forewarning.

The demographic factors that add to the model have plausibility. The association between chronological age and miscarriage is well documented, and short follicular phase has also been associated with miscarriage by other authors[27,28]. Short cycle length may represent a surrogate marker for advanced reproductive age as the initial transition to peri-menopause can be characterised by a shortening of cycle length[29].

Strengths and limitations. The two-stage model did not allow the investigation of the nature of the association between miscarriage and hCG, something which is possible with the joint model. Although attempted it was not feasible to sensibly fit a joint model with a first derivative association structure, possibly due to the small sample size. This is something that requires further investigation in a larger dataset, particularly as there is evidence in the literature, which suggests that the overall profile of hCG is important as opposed to changes in absolute values of hCG. Certainly, both recent papers utilising Bayesian non-parametric models, and mixed effects penalized splines model approaches, focused on classification of each type of pregnancy based on complete longitudinal profiles[30,31].

A Weibull model was utilised to model the baseline hazard, however ideally more flexibility would be desirable. This could be achieved by using restricted cubic splines to model the baseline hazard. Model selection was carried out using forwards stepwise selection, which is known to introduce bias[32]. Alternative selection methods should be considered in future, specific to the joint modelling context. Selection based on the log likelihood contribution for the longitudinal part and conditional survival model have been proposed, but are currently only implemented in the SAS statistical software[33]. The example dataset was relatively small, and so fitting a model as complex as the joint model was challenging. Results must therefore be interpreted with caution. As this was a retrospective analysis of data with limited follow-up measurements, it was not possible to update predictions as measurements were observed. Predictions were therefore inaccurate for wider time periods. With the small sample size, it was also not viable to split the dataset for development and validation of the model. Attempting to utilise such data for diagnostic or monitoring purposes also requires careful consideration of the potential for false positives. This study did not take into account the sensitivity and specificity of the fitted model, however this is an important component for planned future analyses in line with developments in joint model methodology[34,35].

When utilising the joint model framework, it is essential to think about adjustments that need to be made to the model to truly reflect the biological reality of the biomarker and disease processes. In this analysis considerations were made for the timeline on which miscarriage was modelled and how this affected the inclusion of fixed and random effects. Due to limitations of the software, the models included fixed and random intercepts, though no hCG would have been detectable at time zero. Date of conception would also be unknown in a natural pregnancy setting, making this analysis more suited to an IVF setting. This, however, could be adjusted for by using the last menstrual period (LMP) as a timeline in a natural pregnancy setting.

Employing two separate modelling techniques for longitudinal and survival data requires larger sample size requirements in a clinical trial setting. The increased efficiency of simultaneously modelling the two outcomes has the advantage of maintaining desired power at a lower sample size[36]. This makes designing clinical trials around a joint model framework an attractive prospect.

Conclusions

The novel extension to this analysis concerns the subject-specific predictions. This study is an initial investigation into whether women at high risk of miscarriage could be adaptively monitored via their urinary hCG concentration. Though the effectiveness of possible treatments, particularly for recurrent miscarriage, remain uncertain; the joint model is well placed for dynamic monitoring. However long term follow-up observations are required, along with access to a larger dataset for a model to be developed and subsequently validated. Future analyses should also consider the sensitivity and specificity of the fitted predictive model, to minimise the likelihood of false diagnoses of miscarriage.

Methods

Description of dataset. Women attempting to conceive were asked to collect daily early morning urine samples for their entire menstrual cycle and up to 28 days after the day of their missed period if they became pregnant. Women recruited were aged between 18–45 years and were not excluded on the basis of existing fertility issues. Intra-individual variation in the concentration of first morning urine is much lower, than when

considering all urine voids. In addition, the exponential rise of hCG in early pregnancy from <1 mIU/ml to >150 000 mIU/ml renders fluctuations in urine concentration as having minimal effect on the trajectory of rise. Therefore, correction for urine concentration differences, e.g. using creatinine, was not required.

Urinary concentration of hCG was quantified using a validated quantitative immunoassay system (AutoDELFIA; PerkinElmer, Waltham, USA). The concentration of luteinising hormone (LH) in the urine were analysed by a panel of experts from a range of disciplines, including statisticians, endocrinologists and clinical scientists, to determine the day of LH surge, which occurs approximately 24 hours prior to ovulation. It was assumed conception occurred the day after the LH surge[37]. Details of sample collection and storage have been described previously[15]. Additional maternal demographics, menstrual and pregnancy history data were recorded. The study was carried out in accordance with the ethical principles of the Declaration of Helsinki. Written, informed consent was obtained from all individual participants involved in the study.

Statistical analyses. The data utilised in this analysis have been analysed previously using a two-stage model approach[15]. The two-stage model utilises existing modelling techniques by first fitting a linear mixed effects model to the longitudinal data. Subject-specific predictions are then obtained from the mixed model and included as a time-varying covariate in a survival model. This method incorrectly assumes that the biomarker remains stagnant between measurements and gives too precise estimates with unrealistically small standard errors[16]. This analysis will now be extended using a joint model framework, which offers a number of advantages over the two-stage model approach.

Longitudinal modelling. The joint model is made up of two component models – the longitudinal linear mixed effects model and proportional hazards survival model[36]. The longitudinal model for urinary log hCG forms a trajectory function, which estimates the unobserved values of log hCG for the i^{th} patient at time t to form complete profiles. The formulation of the fitted model is as follows,

$$log\ hCG_i(t) = m_i(t) + e_{ij}$$

$$m_i(t) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\ time + \beta_2\ time^2$$

$$b_i \sim MVN(0, \Omega_u) \qquad e_{ij} \sim N(0, \sigma_e^2)$$

The model is made of fixed effects parameters including a fixed intercept (β_0) and linear and quadratic time since conception terms, with parameter estimates, β_1 and β_2 respectively. The random effects parameters allow each individual i to vary at baseline via a random intercept (b_{oi}) and over time through a random linear time since conception term, with parameter estimate b_{1i}. The possibility of measurement error, as with any continuous biomarker, is accounted for via the residual error term, e_{ij}, which is normally distributed. The random effects b_i are multivariate normally distributed. An unstructured correlation matrix was assumed.

Survival modelling. A proportional hazards survival submodel was assumed, conditional on $M_i(t) = \{m_i(s),\ 0 \le s \le t\}$, which denotes the history of the true unobserved longitudinal measurements up to time t and additional covariates v_i[14]. The specific fitted model is given by,

$$h(t|M_i(t), v_i) = h_0(t)\exp[\gamma_1 age + \gamma_2\ usual\ cycle\ length + \alpha\ m_i(t)]$$

The baseline hazard, $h_0(t)$, was assumed to follow a Weibull distribution. Maternal age and usual cycle length were included as covariates in the survival model, after a forwards model selection procedure was carried out at the 5% significance level. The inclusion of $m_i(t)$ in the survival submodel estimates the change in absolute log hCG values and is termed the current value parameterisation, with association parameter α. By including the longitudinal model within the survival submodel, we effectively link the expected value of log hCG to the miscarriage or censoring time, where typically an hCG response would not have been observed. Various association structures were explored, including the first derivative association structure, which models the rate of change of log hCG.

To allow for comparison a standard survival model with log hCG included as a time-varying covariate was fitted, as well as a two-stage model using subject specific predictions from the longitudinal model, as defined for the joint model, in a survival model.

Subject-specific survival probabilities dependent on maternal age and longitudinal log hCG measurements were obtained from the sample on which the joint model was fitted, using the Stata package `stjm`[13]. Conditional survival predictions can potentially be updated as measurements are observed, giving a real-time risk of miscarriage, or dynamic predictions. All models were fitted in Stata IC version 15.1.

Funding and ethical approval. This was a diagnostic accuracy study on a sample bank collected from a multicentre, prospective study, conducted by Radiant Research (USA) on behalf of the sponsor SPD Development Company Ltd. (UK). The study was approved by Quorum Review Committee on 30th November 2009; clinical trial number NCT01077583. This analysis was conducted by N.B.A as part of a doctoral training programme jointly funded by MRC IMPACT and SPD Development Company Ltd.

Data availability

The datasets analysed during the current study are not publicly available due to confidentiality restrictions in place between SPD Development Company Ltd. and the University of Leicester. Stata code written/used to perform the analysis are available from the corresponding author on reasonable request.

Received: 18 September 2019; Accepted: 26 February 2020;
Published online: 12 March 2020

References

1. National Institute for Health and Care Excellence. Scenario: Managing suspected miscarriage. https://cks.nice.org.uk/miscarriage#!scenario (2018).
2. Jurkovic, D., Overton, C., & Bender-Atik, R. Diagnosis and management of first trimester miscarriage. *BMJ*. **346** (2013).
3. Robinson, G. E. Pregnancy loss. *Best. Pract. Res. Clin. Obstet. Gynaecology.* **28**, 169–178 (2014).
4. Prior, M. *et al*. Priorities for research in miscarriage: a priority setting partnership between people affected by miscarriage and professionals following the James Lind Alliance methodology. *BMJ Open.* **7**, e016571, https://doi.org/10.1136/bmjopen-2017-016571 (2017).
5. Barnhart, K. T. *et al*. Symptomatic Patients With an Early Viable Intrauterine Pregnancy: hCG Curves Redefined. *Obstet. Gynecology.* **104**, 50–55 (2004).
6. Fritz, M. A. & Guo, S. Doubling time of human chorionic gonadotropin (hCG) in early normal pregnancy: relationship to hCG concentration and gestational age. *Fertil. Steril.* **47**, 584–589 (1987).
7. Senapati, S. & Barnhart, K. T. Biomarkers for ectopic pregnancy and pregnancy of unknown location. *Fertil. Sterility.* **99**, 1107–1116 (2013).
8. Norman, R. J., Menabawey, M., Lowings, C., Buck, R. H. & Chard, T. Relationship between blood and urine concentrations of intact human chorionic gonadotropin and its free subunits in early pregnancy. *Obstet. Gynecology.* **69**, 590–593 (1987).
9. Asar, Ö., Ritchie, J., Kalra, P. A. & Diggle, P. J. Joint modelling of repeated measurement and time-to-event data: an introductory tutorial. *Int. J. Epidemiology.* **44**, 334–344 (2015).
10. Wulfsohn, M. S. & Tsiatis, A. A. A joint model for survival and longitudinal data measured with error. *Biometrics.* **53**, 330–339 (1997).
11. Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. Applied longitudinal analysis 2nd edition. (Wiley, 2004).
12. Elandt-Johnson, R. C. & Johnson, N. L. Survival models and data analysis. (John Wiley & Sons 1999).
13. Crowther, M. J., Abrams, K. R. & Lambert, P. C. Joint modelling of longitudinal and survival data. *Stata Journal.* **13**, 165–184 (2013).
14. Rizopoulos, D. Joint models for longitudinal and time-to-event data: with applications in R. 50–51 (Chapman and Hall/CRC, 2012).
15. Marriott, L., Zinaman, M., Abrams, K. R., Crowther, M. J. & Johnson, S. Analysis of urinary human chorionic gonadotrophin concentrations in normal and failing pregnancies using longitudinal, Cox proportional hazards and two-stage modelling. *Ann. Clin. Biochemistry.* **54**, 548–557 (2017).
16. Sweeting, M. J. & Thompson, S. G. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal.* **53**, 750–763 (2011).
17. Rizopoulos, D. JM: An R package for the joint modelling of longitudinal and time-to-event data. *J. Stat. Software.* **35**, 1–33, https://doi.org/10.18637/jss.v035.i09 (2010).
18. Hickey, G. L., Philipson, P., Jorgensen, A. & Kolamunnage-Dona, R. joineRML: a joint model and software package for time-to-event and multivariate longitudinal outcomes. *BMC Med. Res. Methodol.* **18**, 50 (2018).
19. Crowther, M. J. merlin-a unified modelling framework for data analysis and methods development in Stata. Preprint at https://arxiv.org/abs/1806.01615 [Submitted] (2018).
20. Johnson, S. & Marriott, L. hCG levels can decline pre- or post- onset of bleeding in early loss. PO99 @ Fertility 2019 Conference (2019).
21. Foo, L. *et al*. Peri-implantation urinary hormone monitoring distinguishes between types of first-trimester spontaneous pregnancy loss. *Paediatric and Perinatal Epidemiology*. **00** (Special Issue: Leveraging Technology), 1–8 (2019).
22. Tiplady, S., Jones, G., Campbell, M., Johnson, S. & Ledger, W. Home ovulation tests and stress in women trying to conceive: a randomized controlled trial. *Hum. Reproduction.* **28**, 138–151 (2013).
23. Weddell, S. *et al*. Home ovulation test use and stress during subfertility evaluation: subarm of a randomized controlled trial. *Women's Health.* **15**, 1745506519838363 (2019).
24. Tewary, S. *et al*. Impact of sitagliptin on endometrial mesenchymal stem-like progenitor cells: a randomised, double-blind placebo-controlled feasibility trial. *E. Bio. Medicine.* **51**, 102597 (2020).
25. Coomarasamy, A. *et al*. A Randomized Trial of Progesterone in Women with Recurrent Miscarriages. *N. Engl. J. Medicine.* **373**, 2141–2148 (2015).
26. Coomarasamy, A. *et al*. A Randomized Trial of Progesterone in Women with Bleeding in Early Pregnancy. *N. Engl. J. Medicine.* **380**, 1815–1824 (2019).
27. Jukic, A. M. Z., Weinberg, C. R., Baird, D. D. & Wilcox, A. J. Lifestyle and reproductive factors associated with follicular phase length. *J. Women's Health.* **16**, 1340–1347 (2007).
28. Small, C. M. *et al*. Menstrual cycle characteristics: associations with fertility and spontaneous abortion. *Epidemiology.* **17**, 52–60 (2006).
29. Miro, F. *et al*. Sequential classification of endocrine stages during reproductive aging in women: the FREEDOM study. *Menopause.* **12**, 281–290 (2005).
30. Gaskins, J. T., Fuentes, C. & De la Cruz, R. A Bayesian Nonparametric Model for Predicting Pregnancy Outcomes Using Longitudinal Profiles. Preprint https://arxiv.org/abs/1711.01512 (2017).
31. De la Cruz, R., Fuentes, C., Meza, C., Lee, D. & Arribas-Gil, A. Predicting pregnancy outcomes using longitudinal information: a penalized splines mixed-effects model approach. *Stat. Medicine.* **36**, 2120–2134 (2017).
32. Lukacs, P. M., Burnham, K. P. & Anderson, D. R. Model selection bias and Freedman's paradox. *Ann. Inst. Stat. Mathematics.* **62**, 117–125 (2010).
33. Zhang, D., Chen, M., Ibrahim, J. G., Boye, M. E. & Shen, W. JMFit: a SAS macro for joint models of longitudinal and survival data. *Journal of Statistical Software*. **71** (2016).
34. Rizopoulos, D. Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data. *Biometrics* **67**, 819–829 (2011).
35. Andrinopoulou, E., Eilers, P. H. C., Takkenberg, J. J. M. & Rizopoulos, D. Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using P-splines. *Biometrics.* **74**, 685–693 (2018).
36. Ibrahim, J. G., Chu, H. & Chen, L. M. Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data. *J. Clin. Oncology.* **28**, 2796–2801 (2010).
37. Johnson, S. *et al*. Development of the first urinary reproductive hormone ranges referenced to independently determined ovulation day. *Clin. Chem. Laboratory Medicine.* **53**, 1099–1108 (2015).

Author contributions

N.B.A. conducted the statistical analysis, drafted and reviewed the manuscript. L.M. processed the data and reviewed the data and manuscript. S.J. developed the sample collection protocol, managed the clinical study and reviewed the data and manuscript. M.J.C. and K.R.A. provided support with statistical analysis and reviewed the manuscript.

Competing interests

M.J.C. and K.R.A. were awarded funding by SPD Development Company Ltd (Bedford, UK) as part of an MRC IMPACT PhD studentship. N.B.A. is the recipient of the MRC IMPACT PhD studentship funded in collaboration with SPD Development Company Ltd. (Bedford, UK). L.M. and S.J. are employees of SPD Development Company Ltd. (Bedford, UK).

Additional information

Correspondence and requests for materials should be addressed to N.B.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix B

Modelling checks for intercourse analyses

Appendix B.1 contains Schoenfeld residual, Martingale residual, deviance residual and delta-beta plots for the multivariable model for acts during implantation.

Appendix B.2 contains Schoenfeld residual, Martingale residual, deviance residual and delta-beta plots for the multivariable model for acts during the three-day implantation window.

Appendix B.3 contains Schoenfeld residual, Martingale residual, deviance residual and delta-beta plots for the multivariable model for the last act in the fertile window.

Appendix B.4 contains Schoenfeld residual, Martingale residual, deviance residual and delta-beta plots for the multivariable model for the number of acts in the fertile window.

B.1 Model checks for acts during the implantation window

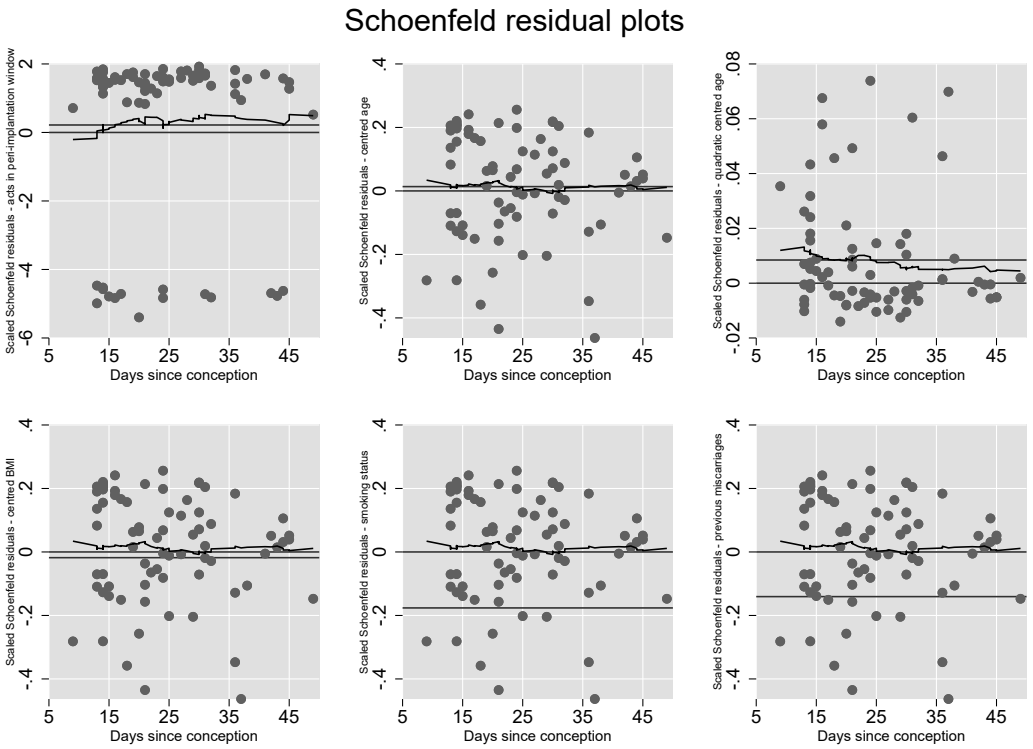


FIGURE B.11. Schoenfeld residual plots to assess the proportional hazards assumption for the multivariable model for acts in the peri-implantation window

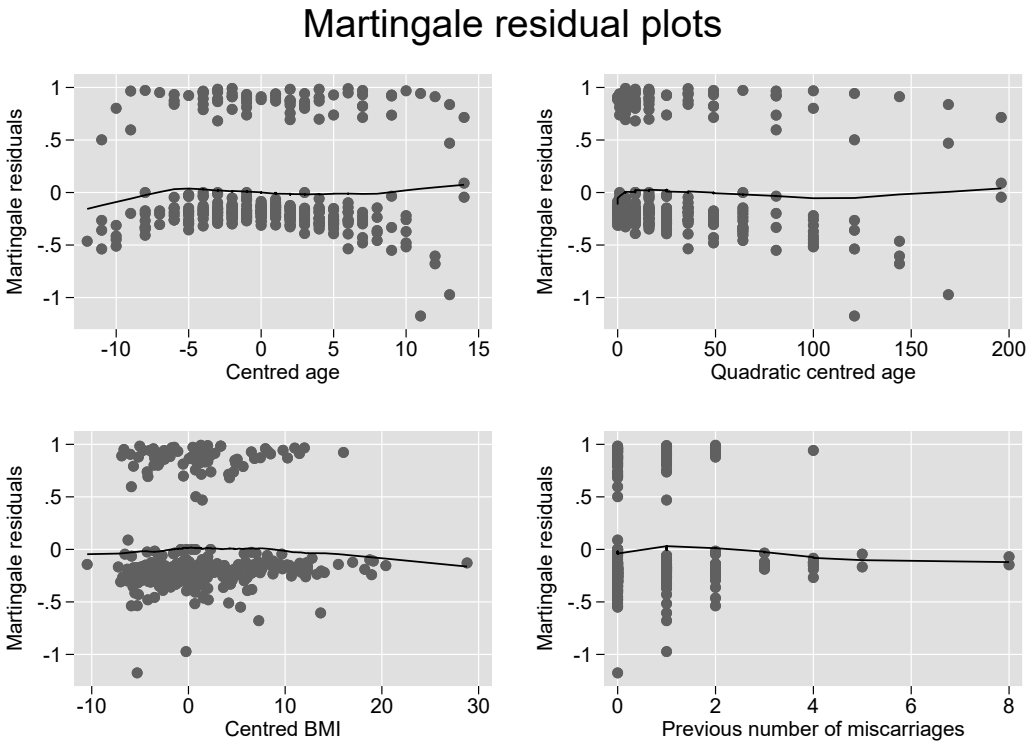


FIGURE B.12. Martingale residual plots to assess the functional forms of variables included in the multivariable model for acts in the peri-implantation window

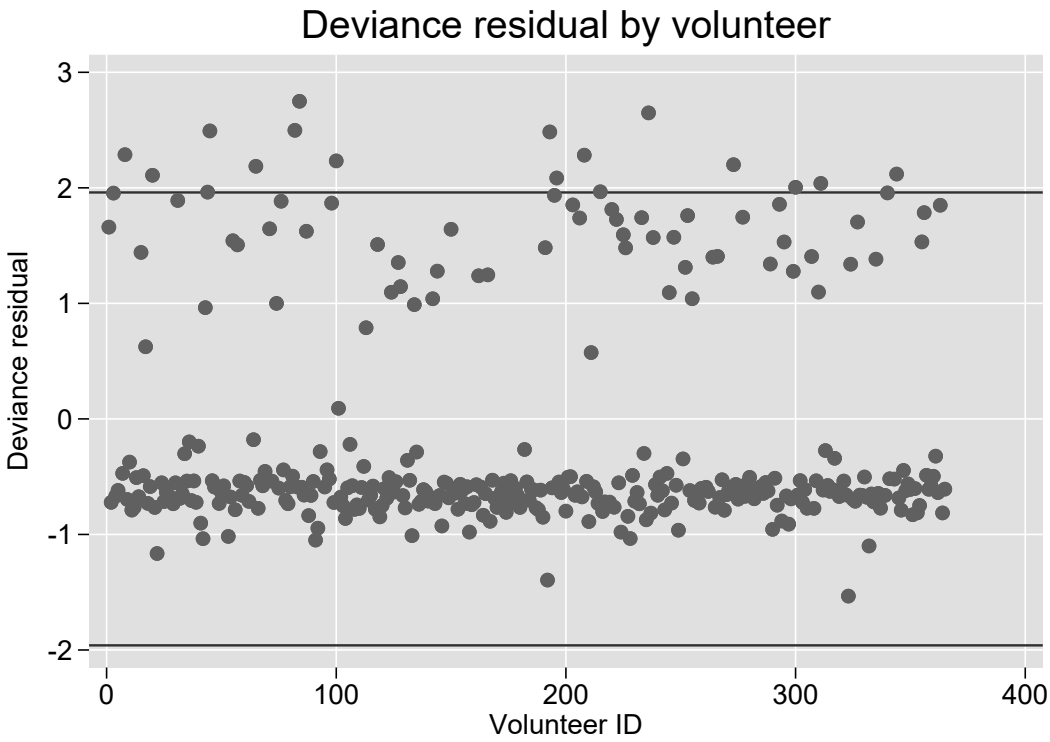


FIGURE B.13. Deviance residual plot for multivariable model for acts in the peri-implantation window

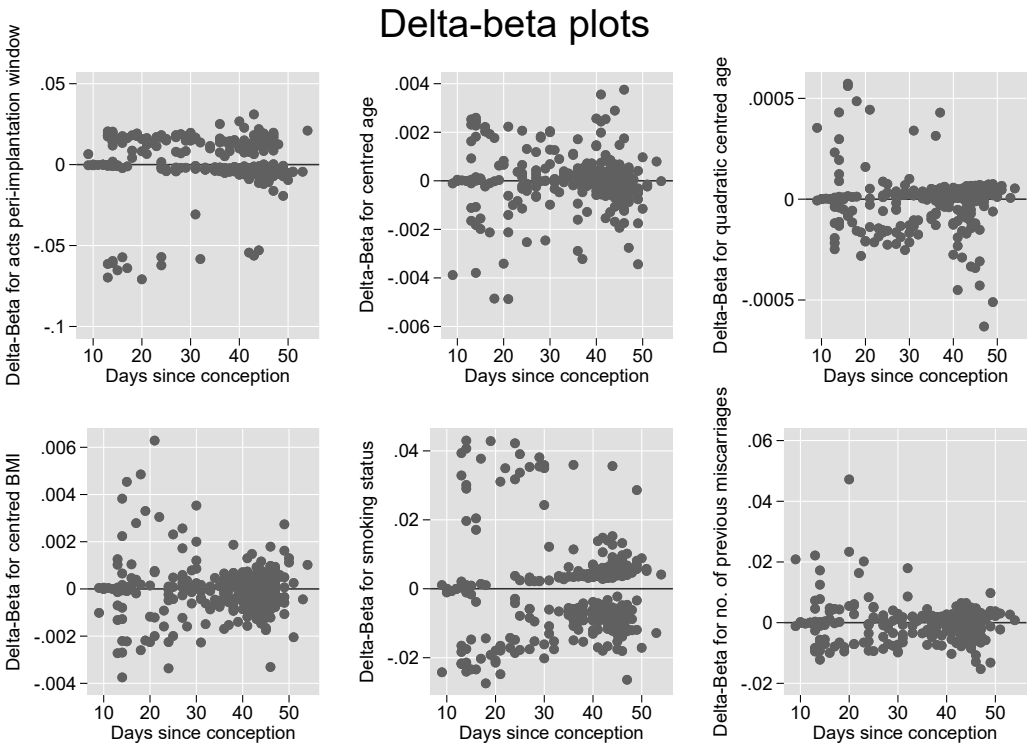


FIGURE B.14. Delta-beta plot for multivariable model for acts in the per-implantation window

B.2 Model checks for acts during the three-day implantation window

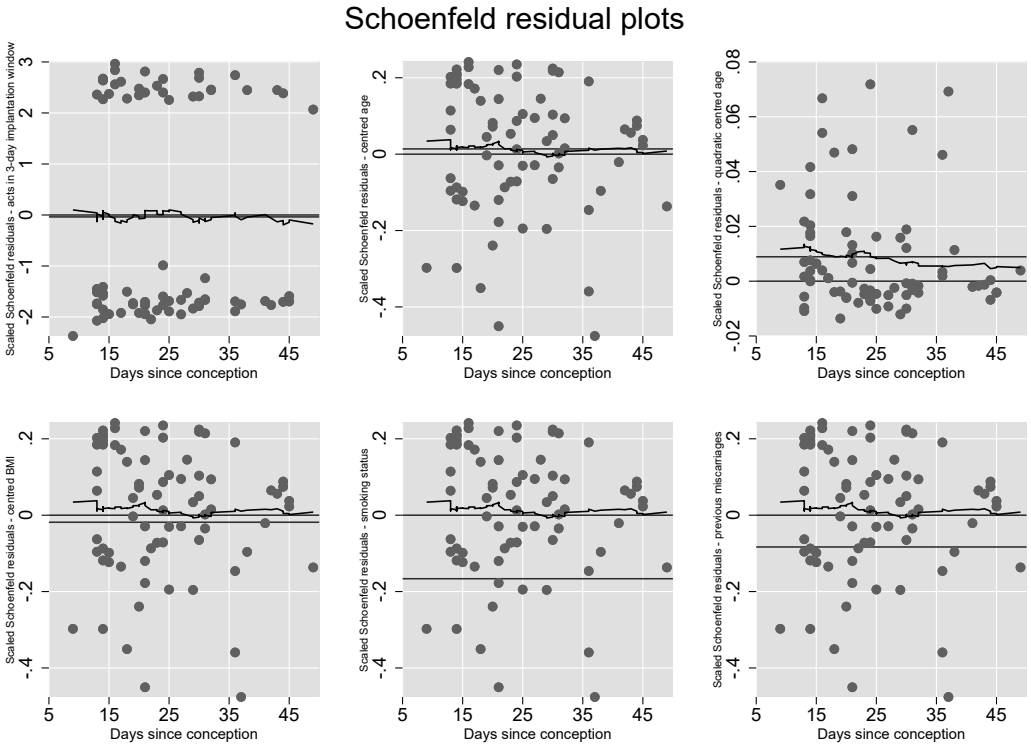


FIGURE B.25. Schoenfeld residual plots to assess the proportional hazards assumption for the multivariable model for acts in the three-day implantation window

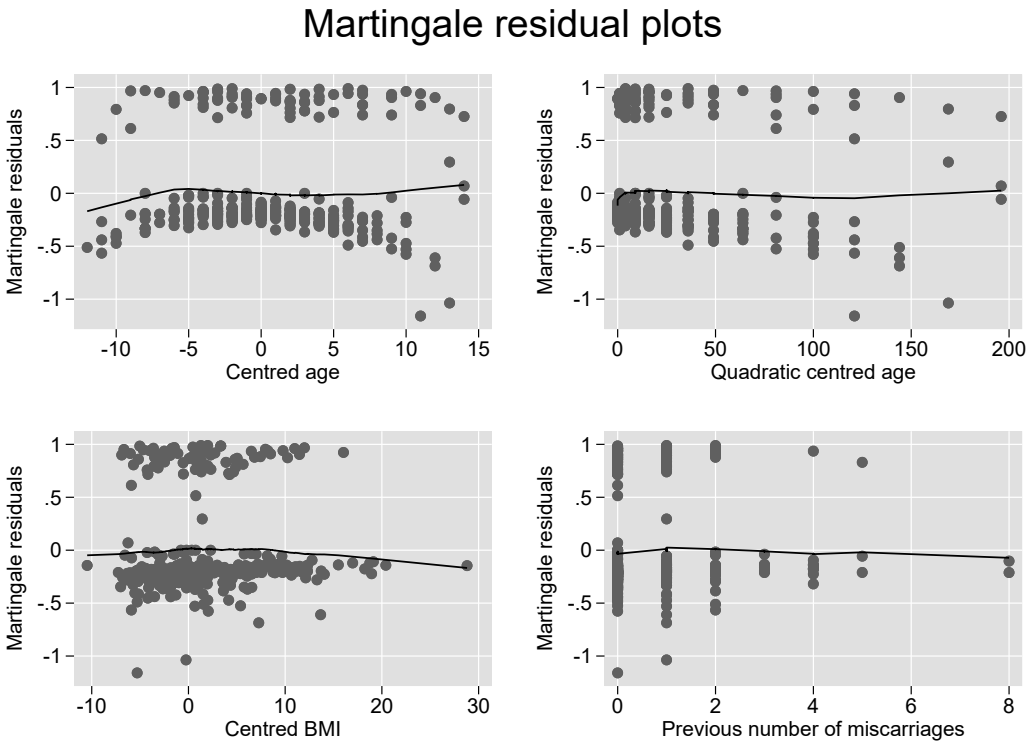


FIGURE B.26. Martingale residual plots to assess the functional forms of variables included in the multivariable model for acts in the three-day implantation window

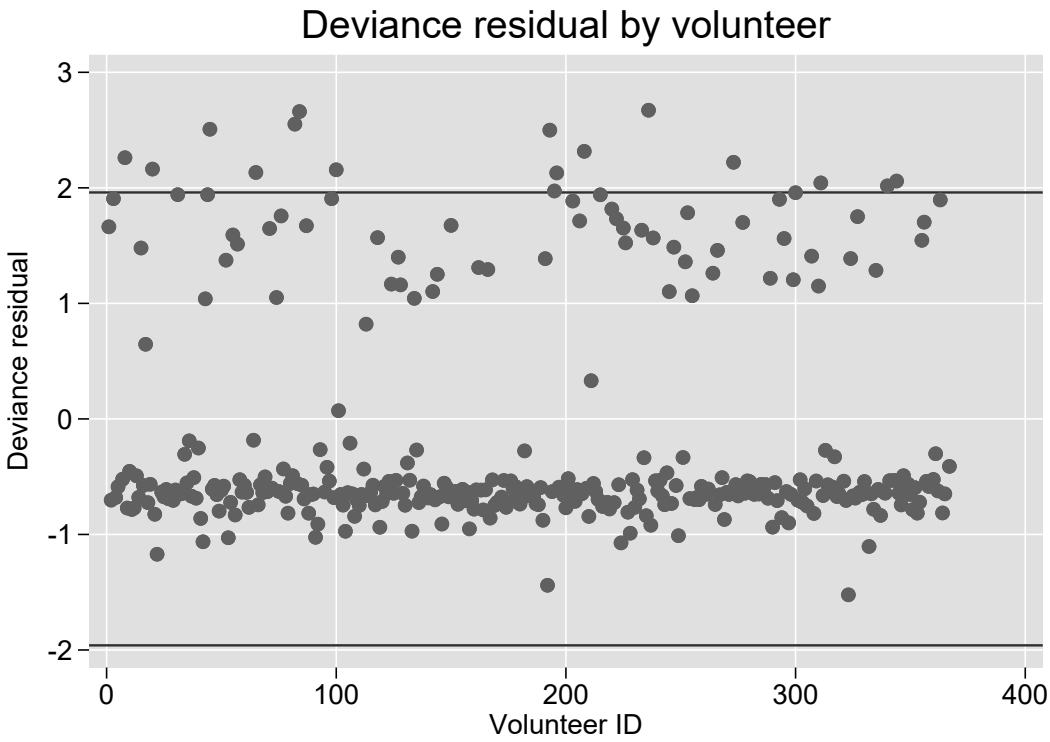


FIGURE B.27. Deviance residual plot for multivariable model for acts in the three-day implantation window

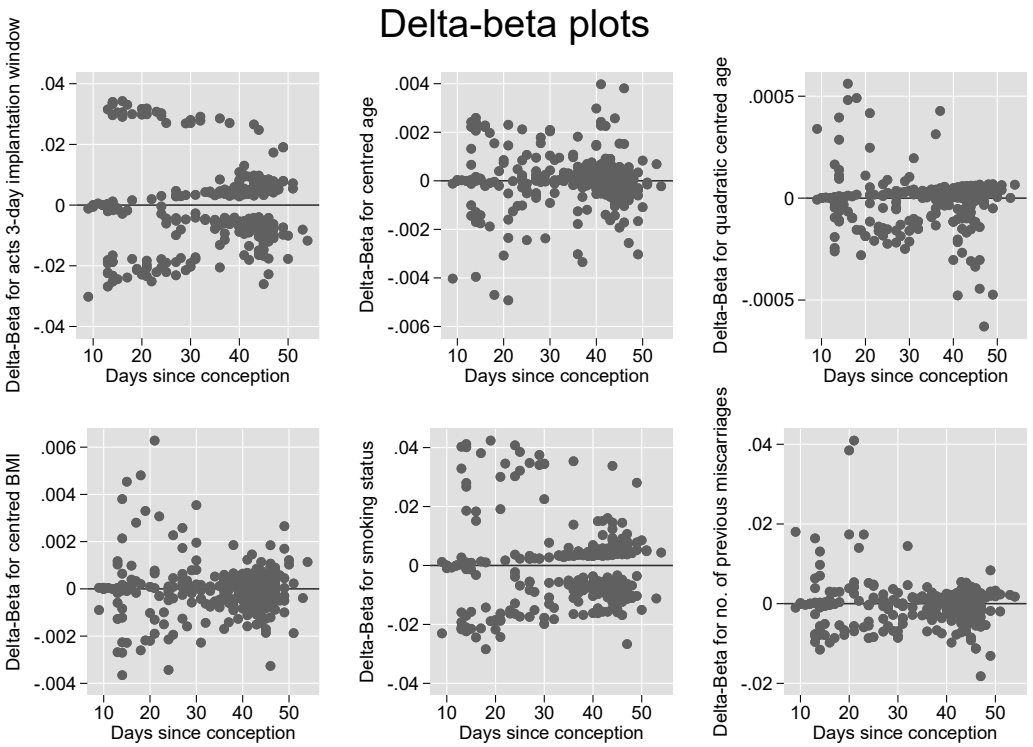


FIGURE B.28. Delta-beta plot for multivariable model for acts in the three-day implantation window

B.3 Model checks for last act in fertile window

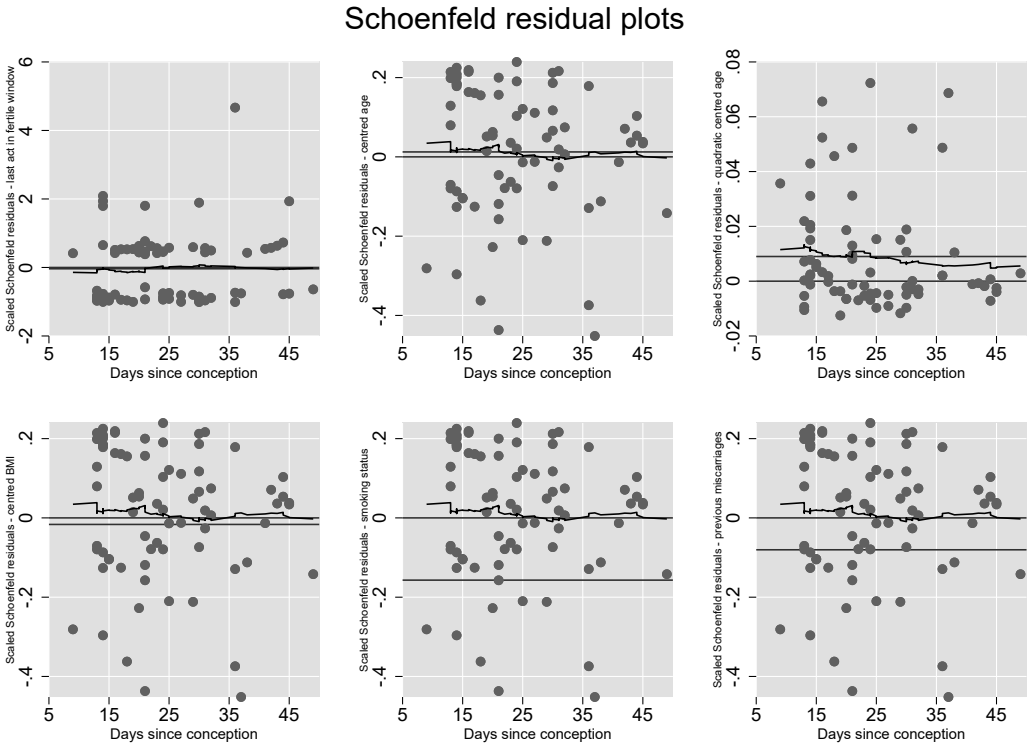


FIGURE B.39. Schoenfeld residual plots to assess the proportional hazards assumption for the multivariable model for last act in fertile window

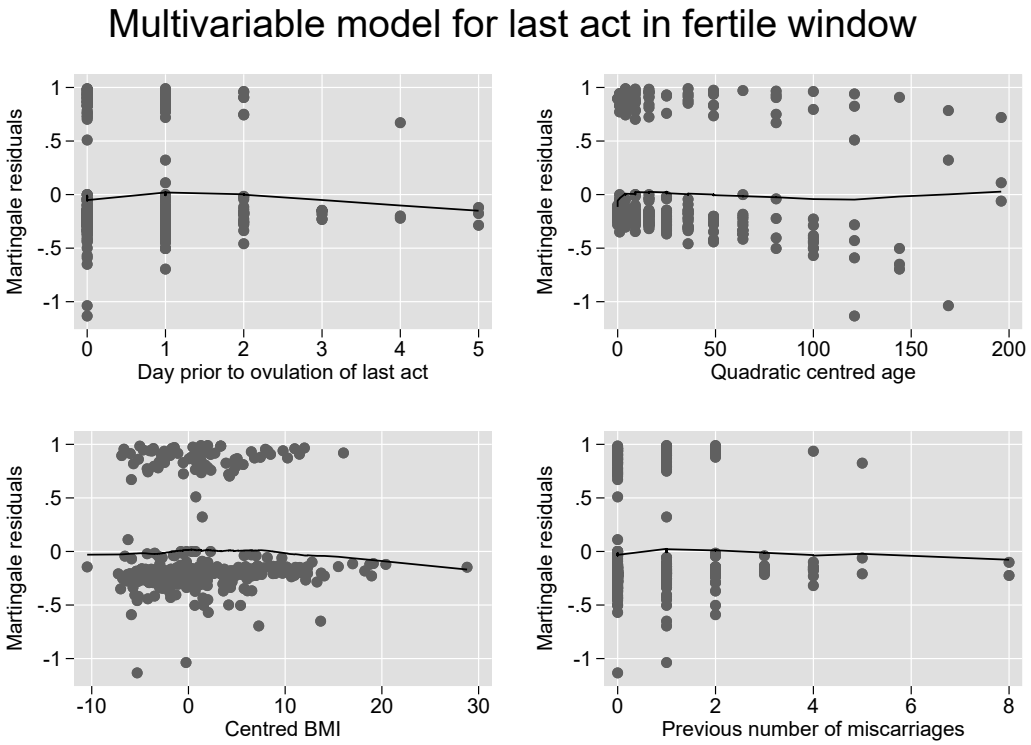


FIGURE B.310. Martingale residual plots to assess the functional forms of variables included in the multivariable model for last act in the fertile window

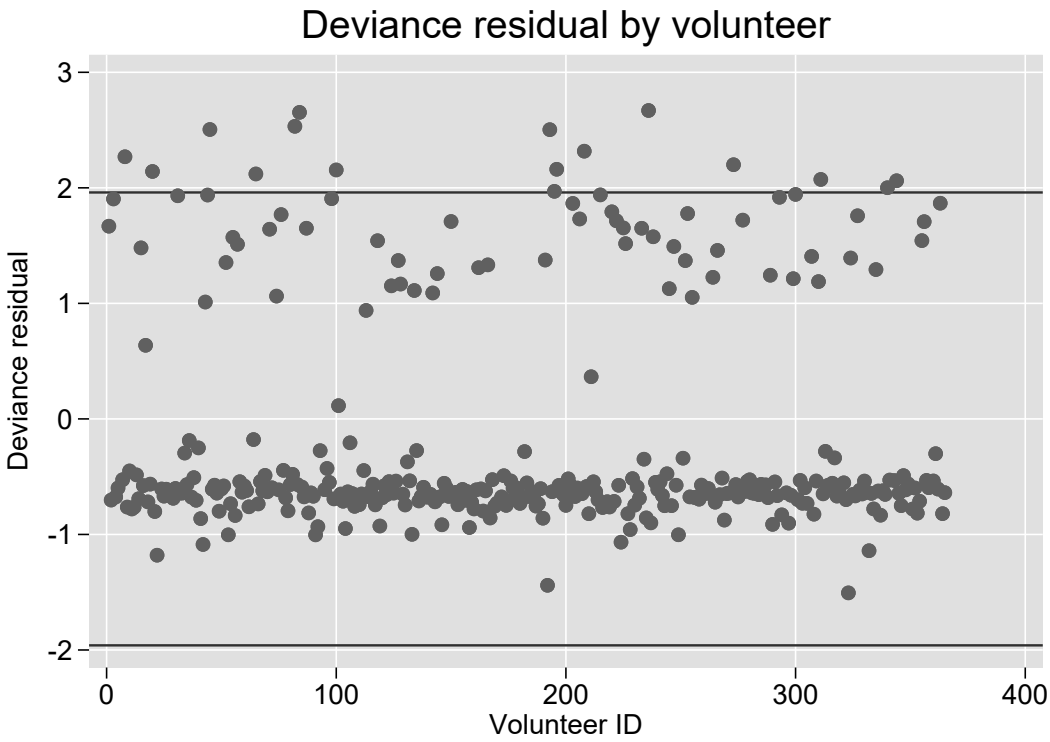


FIGURE B.311. Deviance residual plot for multivariable model for last act in the fertile window

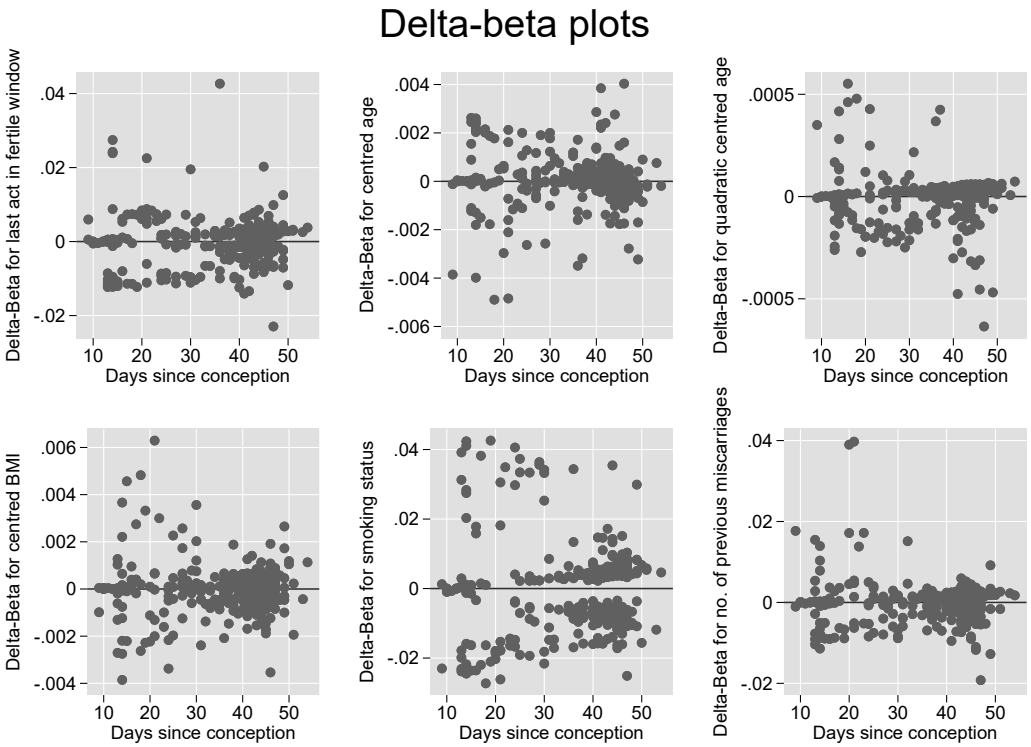


FIGURE B.312. Delta-beta plot for multivariable model for last act in the fertile window

B.4 Model checks for number of acts in the fertile window

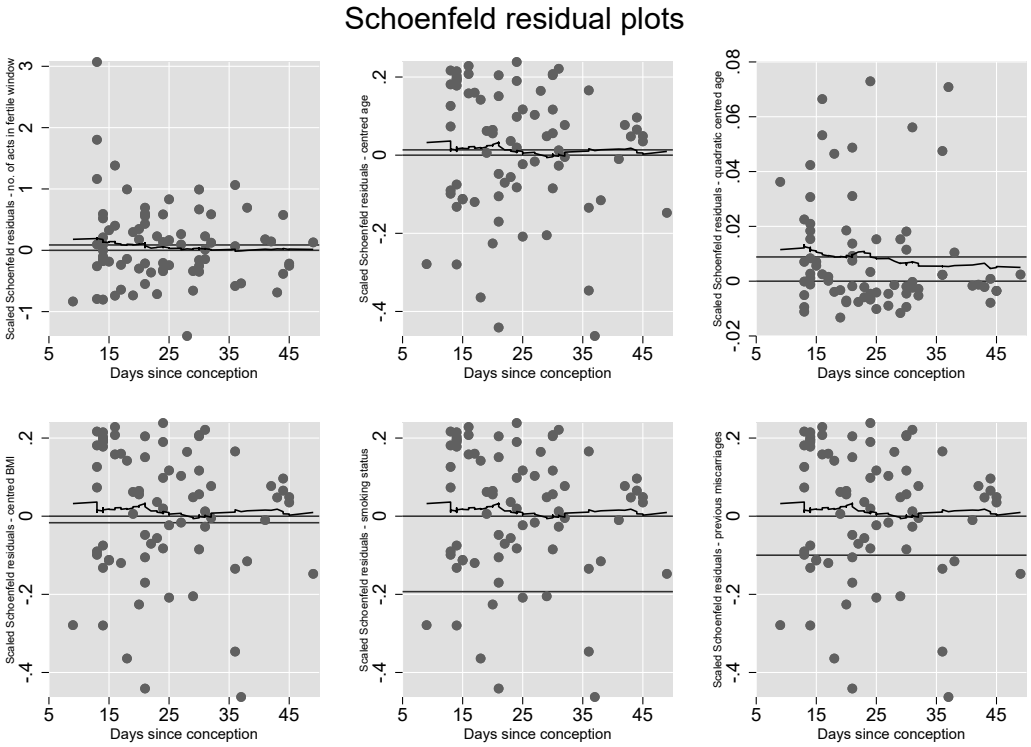


FIGURE B.413. Schoenfeld residual plots to assess the proportional hazards assumption for the multivariable model for number of acts in the fertile window

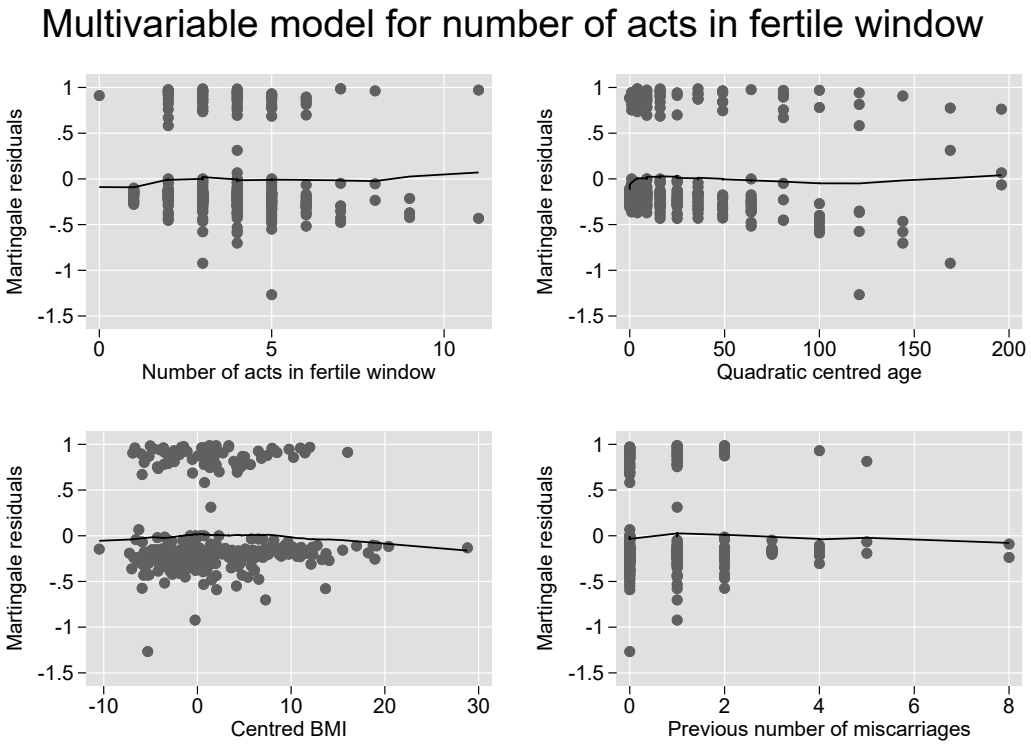


FIGURE B.414. Martingale residual plots to assess the functional forms of variables included in the multivariable model for number of acts in the fertile window

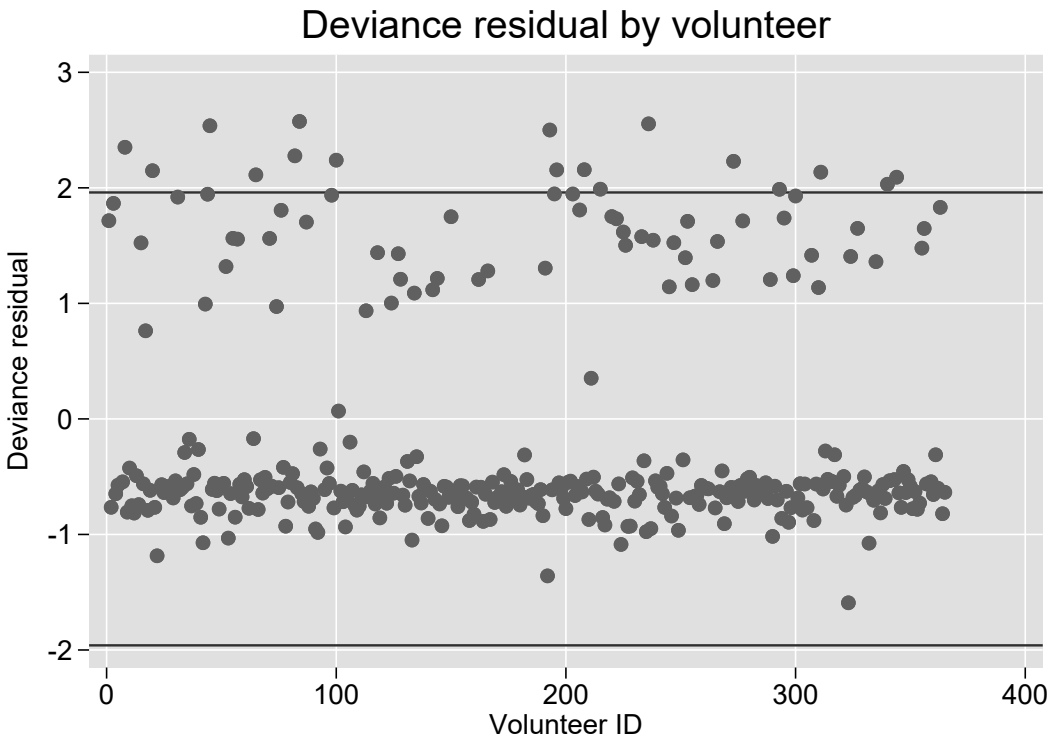


FIGURE B.415. Deviance residual plot for multivariable model for number of acts in the fertile window

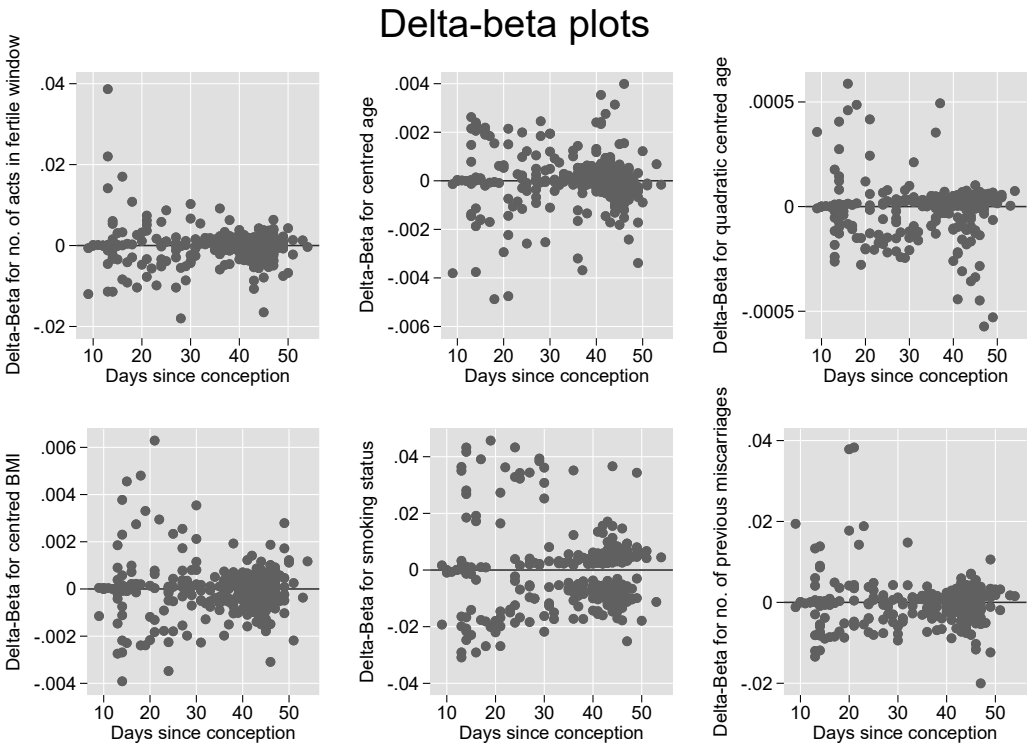


FIGURE B.416. Delta-beta plot for multivariable model for number of acts in the fertile window

Bibliography

[1] L. Ray, "The menstrual cycle: more than just your period," Online, Clue, Dec 2019. [Online]. Available: https://helloclue.com/articles/cycle-a-z/the-menstrual-cycle-more-than-just-the-period

[2] J. Merlo, B. Chaix, M. Yang, J. Lynch, and L. Råstam, "A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon," *Journal of Epidemiology & Community Health*, vol. 59, no. 6, pp. 443–449, 2005.

[3] M. J. Crowther, "Development and application of methodology for the parametric analysis of complex survival and joint longitudinal-survival data in biomedical research," Ph.D. dissertation, University of Leicester, Leicester, 2014.

[4] C. D. Lynch, L. W. Jackson, and G. M. Buck Louis, "Estimation of the day-specific probabilities of conception: current state of the knowledge and the relevance for epidemiological research," *Paediatric and Perinatal Epidemiology*, vol. 20, no. s1, pp. 3–12, 2006.

[5] D. Jurkovic, C. Overton, and R. Bender-Atik, "Diagnosis and management of first trimester miscarriage," *BMJ*, vol. 346, 2013.

[6] American College of Obstetricians and Gynecologists, "Acog practice bulletin no. 200: early pregnancy loss," *Obstetrics and Gynecology*, vol. 132, no. 5, pp. e197–e207, 2018.

[7] A. B. Hooker, H. Aydin, H. A. Brölmann, and J. A. Huirne, "Long-term complications and reproductive outcome after the management of retained products of conception: a systematic review," *Fertility and Sterility*, vol. 105, no. 1, pp. 156–164, 2016.

[8] O. B. Van den Akker, "The psychological and social consequences of miscarriage," *Expert Review of Obstetrics & Gynecology*, vol. 6, no. 3, pp. 295–304, 2011.

[9] T. Li, M. Makris, M. Tomsu, E. Tuckerman, and S. Laird, "Recurrent miscarriage: aetiology, management and prognosis," *Human Reproduction Update*, vol. 8, no. 5, pp. 463–481, 2002.

[10] S. R. Johnson, F. Miro, S. Barrett, and J. E. Ellis, "Levels of urinary human chorionic gonadotrophin (hCG) following conception and variability of menstrual cycle length in a cohort of women attempting to conceive," *Current Medical Research and Opinion*, vol. 25, no. 3, pp. 741–748, 2009.

[11] R. N. Pillai, J. C. Konje, D. G. Tincello, and N. Potdar, "Role of serum biomarkers in the prediction of outcome in women with threatened miscarriage: a systematic review and diagnostic accuracy meta-analysis," *Human Reproduction Update*, vol. 22, no. 2, pp. 228–239, 2016.

[12] M. Memtsa, D. Jurkovic, and E. Jauniaux, "Diagnostic biomarkers for predicting adverse early pregnancy outcomes," *BJOG*, vol. 126, no. 3, pp. E107–E113, 2019.

[13] D. Jurkovic, L. Valentin, and S. Vyas, Eds., *Gynaecological ultrasound in clinical practice: ultrasound imaging in the management of gynaecological conditions.* RCOG, 2009.

[14] K. T. Barnhart, M. D. Sammel, P. F. Rinaudo, L. Zhou, A. C. Hummel, and W. Guo, "Symptomatic patients with an early viable intrauterine pregnancy: hCG curves redefined," *Obstetrics & Gynecology*, vol. 104, no. 1, pp. 50–55, 2004.

[15] L. Foo, S. Johnson, L. Marriott, T. Bourne, P. Bennett, and C. Lees, "Peri-implantation urinary hormone monitoring distinguishes between types of

first-trimester spontaneous pregnancy loss," *Paediatric and Perinatal Epidemiology*, vol. 34, no. 5, pp. 495–503, 2020.

[16] D. Hay, M. Gronow, A. Lopata, and J. Brown, "Monitoring early production of chorionic gonadotrophin (hCG) following in vitro fertilization and embryo transfer," *Australian and New Zealand Journal of Obstetrics and Gynaecology*, vol. 24, no. 3, pp. 206–209, 1984.

[17] National Insititute for Clinical Excellence, "NICE guideline [NG126] Ectopic pregnancy and miscarriage: diagnosis and initial management," Apr 2019. [Online]. Available: https://www.nice.org.uk/guidance/ng126

[18] K. T. Barnhart, "Early pregnancy failure: beware of the pitfalls of modern management," *Fertility and Sterility*, vol. 98, no. 5, pp. 1061–1065, 2012.

[19] A. Weeks and K. G. Danielsson, "Spontaneous miscarriage in the first trimester," pp. 1223–1224, 2006.

[20] S. L. Bailey, J. Boivin, Y. C. Cheong, E. Kitson-Reynolds, C. Bailey, and N. Macklon, "Hope for the best... but expect the worst: a qualitative study to explore how women with recurrent miscarriage experience the early waiting period of a new pregnancy," *BMJ Open*, vol. 9, no. 5, p. e029354, 2019.

[21] A. Coomarasamy, A. J. Devall, V. Cheed, H. Harb, L. J. Middleton, I. D. Gallos, H. Williams, A. K. Eapen, T. Roberts, C. C. Ogwulu *et al.*, "A randomized trial of progesterone in women with bleeding in early pregnancy," *New England Journal of Medicine*, vol. 380, no. 19, pp. 1815–1824, 2019.

[22] A. Coomarasamy, A. J. Devall, J. J. Brosens, S. Quenby, M. D. Stephenson, S. Sierra, O. B. Christiansen, R. Small, J. Brewin, T. E. Roberts *et al.*, "Micronized vaginal progesterone to prevent miscarriage: a critical evaluation of randomized evidence," *American Journal of Obstetrics and Gynecology*,

vol. 223, no. 2, pp. 167–176, 2020.

[23] S. Tewary, E. S. Lucas, R. F., P. K. Kimani, A. Polanco, P. J. Brighton, J. Muter, K. J. Fishwick, M. J. Da Costa, L. J. Ewington, L. Lacey, S. Takeda, J. J. Brosens, and S. Quenby, "Impact of sitagliptin on endometrial mesenchymal stem-like progenitor cells: a randomised, double-blind placebo-controlled feasibility trial," *EBioMedicine*, vol. 51, p. 102597, 2020.

[24] A. Taylor, "ABC of subfertility: extent of the problem," *BMJ*, vol. 327, no. 7412, pp. 434–436, 2003.

[25] R. Norman, M. Menabawey, C. Lowings, R. Buck, and T. Chard, "Relationship between blood and urine concentrations of intact human chorionic gonadotropin and its free subunits in early pregnancy." *Obstetrics and Gynecology*, vol. 69, no. 4, pp. 590–593, 1987.

[26] G. D. Braunstein, "The long gestation of the modern home pregnancy test," *Clinical Chemistry*, vol. 60, no. 1, pp. 18–21, 2014.

[27] Clearblue, "Pregnancy test ultra early," Swiss Precision Diagnostics, 2020. [Online]. Available: https://uk.clearblue.com/pregnancy-tests/early-detection

[28] M. J. Sweeting and S. G. Thompson, "Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture," *Biometrical Journal*, vol. 53, no. 5, pp. 750–763, 2011.

[29] J. G. Ibrahim, H. Chu, and L. M. Chen, "Basic concepts and methods for joint models of longitudinal and survival data," *Journal of Clinical Oncology*, vol. 28, no. 16, pp. 2796–2801, 2010.

[30] D. Rizopoulos, *Joint models for longitudinal and time-to-event data: with applications in R.* CRC press, 2012.

[31] A. A. Tsiatis, V. Degruttola, and M. S. Wulfsohn, "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*, vol. 90, no. 429, pp. 27–37, 1995.

[32] R. Henderson, P. Diggle, and A. Dobson, "Joint modelling of longitudinal measurements and event time data," *Biostatistics*, vol. 1, no. 4, pp. 465–480, 2000.

[33] M. J. Sweeting, J. K. Barrett, S. G. Thompson, and A. M. Wood, "The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the aric study," *Statistics in Medicine*, vol. 36, no. 28, pp. 4514–4528, 2017.

[34] Y.-Y. Chi and J. G. Ibrahim, "Joint models for multivariate longitudinal and multivariate survival data," *Biometrics*, vol. 62, no. 2, pp. 432–445, 2006.

[35] X. Huang, G. Li, R. M. Elashoff, and J. Pan, "A general joint model for longitudinal measurements and competing risks survival data with heterogeneous random effects," *Lifetime Data Analysis*, vol. 17, no. 1, pp. 80–100, 2011.

[36] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data.* John Wiley & Sons, 2002, ch. 6, pp. 193–217.

[37] D. Rizopoulos, *Joint models for longitudinal and time-to-event data: with applications in R.* CRC press, 2012, ch. 3, pp. 35–50.

[38] M. J. Crowther, K. R. Abrams, and P. C. Lambert, "Joint modeling of longitudinal and survival data," *The Stata Journal*, vol. 13, no. 1, pp. 165–184, 2013.

[39] F. E. Harrell, "Introduction to survival analysis," in *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*, 2nd ed. Springer, 2015, ch. 16, pp. 399–422.

[40] ——, "Parametric survival models," in *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*, 2nd ed. Springer, 2015, ch. 17, pp. 413–442.

[41] P. Royston and P. C. Lambert, *Flexible parametric survival analysis using Stata: beyond the Cox model*. Stata Press College Station, TX, 2011, vol. 347.

[42] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[43] D. Rizopoulos, "Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data," *Biometrics*, vol. 67, no. 3, pp. 819–829, 2011.

[44] L. Marriott, M. Zinaman, K. R. Abrams, M. J. Crowther, and S. Johnson, "Analysis of urinary human chorionic gonadotrophin concentrations in normal and failing pregnancies using longitudinal, cox proportional hazards and two-stage modelling," *Annals of Clinical Biochemistry*, vol. 54, no. 5, pp. 548–557, 2017.

[45] J. T. Gaskins, C. Fuentes, and R. De la Cruz, "A bayesian nonparametric model for classification of longitudinal profiles," *Biostatistics*, pp. 1–17, 2021.

[46] R. De la Cruz, C. Fuentes, C. Meza, D. Lee, and A. Arribas-Gil, "Predicting pregnancy outcomes using longitudinal information: a penalized splines mixed-effects model approach," *Statistics in Medicine*, vol. 36, no. 13, pp. 2120–2134, 2017.

[47] L. Duan, D. Yan, W. Zeng, X. Yang, and Q. Wei, "Predictive power progesterone combined with beta human chorionic gonadotropin measurements in the outcome of threatened miscarriage," *Archives of Gynecology and Obstetrics*, vol. 283, no. 3, pp. 431–435, 2011.

[48] C. Puget, Y. Joueidi, E. Bauville, B. Laviolle, C. Bendavid, V. Lavoué, and M. Le Lous, "Serial hCG and progesterone levels to predict early pregnancy outcomes in pregnancies of uncertain viability: a prospective study," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 220, pp. 100–105, 2018.

[49] J. D. Long and J. A. Mills, "Joint modeling of multivariate longitudinal data and survival data in several observational studies of Huntington's disease," *BMC Medical Research Methodology*, vol. 18, no. 1, pp. 1–15, 2018.

[50] P. R. Williamson, R. Kolamunnage-Dona, P. Philipson, and A. G. Marson, "Joint modelling of longitudinal and competing risks data," *Statistics in Medicine*, vol. 27, no. 30, pp. 6426–6438, 2008.

[51] Y.-T. Hwang, C.-H. Huang, C.-C. Wang, T.-Y. Lin, and Y.-K. Tseng, "Joint modelling of longitudinal binary data and survival data," *Journal of Applied Statistics*, vol. 46, no. 13, pp. 2357–2371, 2019.

[52] M. J. Crowther, T. M.-L. Andersson, P. C. Lambert, K. R. Abrams, and K. Humphreys, "Joint modelling of longitudinal and survival data: incorporating delayed entry and an assessment of model misspecification," *Statistics in Medicine*, vol. 35, no. 7, pp. 1193–1209, 2016.

[53] M. W. Arisido, L. Antolini, D. P. Bernasconi, M. G. Valsecchi, and P. Rebora, "Joint model robustness compared with the time-varying covariate cox model to evaluate the association between a longitudinal marker and a time-to-event endpoint," *BMC Medical Research Methodology*, vol. 19, no. 1, pp.

1–13, 2019.

[54] F. Hsieh, Y.-K. Tseng, and J.-L. Wang, "Joint modeling of survival and longitudinal data: likelihood approach revisited," *Biometrics*, vol. 62, no. 4, pp. 1037–1043, 2006.

[55] A. Z. Steiner, D. A. Pritchard, S. L. Young, and A. H. Herring, "Peri-implantation intercourse lowers fecundability," *Fertility and Sterility*, vol. 102, no. 1, pp. 178–182, 2014.

[56] M. D. Creinin, S. Keverline, and L. A. Meyn, "How regular is regular? an analysis of menstrual cycle regularity," *Contraception*, vol. 70, no. 4, pp. 289–292, 2004.

[57] L. Chiazze, F. T. Brayer, J. J. Macisco, M. P. Parker, and B. J. Duffy, "The length and variability of the human menstrual cycle," *JAMA*, vol. 203, no. 6, pp. 377–380, 1968.

[58] National Health Service, "Periods and fertility in the menstrual cycle," Online, Aug 2019. [Online]. Available: https://www.nhs.uk/conditions/periods/fertility-in-the-menstrual-cycle/

[59] K. Münster, L. Schmidt, and P. Helm, "Length and variation in the menstrual cycle—a cross-sectional study from a danish county," *BJOG*, vol. 99, no. 5, pp. 422–429, 1992.

[60] J. E. Holesh, A. N. Bass, and M. Lord. (2020) Physiology, ovulation. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK441996/

[61] R. L. Wiele, J. Bogumil, I. Dyrenfurth, M. Ferin, R. Jewelewicz, M. Warren, T. Rizkallah, and G. Mikhail, "Mechanisms regulating the menstrual cycle in women," in *Proceedings of the 1969 Laurentian Hormone Conference*. Elsevier, 1970, pp. 63–103.

[62] S. Hillier, "Current concepts of the roles of follicle stimulating hormone and luteinizing hormone in folliculogenesis," *Human Reproduction*, vol. 9, no. 2, pp. 188–191, 1994.

[63] A. J. Wilcox, D. Dunson, and D. D. Baird, "The timing of the "fertile window" in the menstrual cycle: day specific estimates from a prospective study," *BMJ*, vol. 321, no. 7271, pp. 1259–1262, 2000.

[64] S. Taraborrelli, "Physiology, production and action of progesterone," *Acta Obstetricia et Gynecologica Scandinavica*, vol. 94, pp. 8–16, 2015.

[65] D. Betz and K. Fane. (2020) Human chorionic gonadotropin. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK532950/

[66] National Institute for Health and Care Excellence, "Scenario: Managing suspected first trimester miscarriage," Nov 2020. [Online]. Available: https://cks.nice.org.uk/topics/miscarriage/management/managing-suspected-miscarriage/

[67] S. Bhattacharya and S. Bhattacharya, "Effect of miscarriage on future pregnancies," *Women's Health*, vol. 5, no. 1, pp. 5–8, 2009.

[68] M. C. Magnus, A. J. Wilcox, N.-H. Morken, C. R. Weinberg, and S. E. Håberg, "Role of maternal age and pregnancy history in risk of miscarriage: prospective register based study," *BMJ*, vol. 364, 2019.

[69] Practice Committee of the American Society for Reproductive Medicine, "Evaluation and treatment of recurrent pregnancy loss: a committee opinion," *Fertility and Sterility*, vol. 98, no. 5, pp. 1103–1111, 2012.

[70] N. Maconochie, P. Doyle, S. Prior, and R. Simmons, "Risk factors for first trimester miscarriage — results from a uk-population-based case-control study," *BJOG*, vol. 114, no. 2, pp. 170–186, 2007.

[71] H. Lashen, K. Fear, and D. Sturdee, "Obesity is associated with increased risk of first trimester and recurrent miscarriage: matched case–control study," *Human Reproduction*, vol. 19, no. 7, pp. 1644–1646, 2004.

[72] B. L. Pineles, E. Park, and J. M. Samet, "Systematic review and meta-analysis of miscarriage and maternal exposure to tobacco smoke during pregnancy," *American Journal of Epidemiology*, vol. 179, no. 7, pp. 807–823, 2014.

[73] A. J. Wilcox, C. R. Weinberg, J. F. O'Connor, D. D. Baird, J. P. Schlatterer, R. E. Canfield, E. G. Armstrong, and B. C. Nisula, "Incidence of early loss of pregnancy," *New England Journal of Medicine*, vol. 319, no. 4, pp. 189–194, 1988.

[74] M. J. Zinaman, E. D. Clegg, C. C. Brown, J. O'Connor, and S. G. Selevan, "Estimates of human fertility and pregnancy loss," *Fertility and Sterility*, vol. 65, no. 3, pp. 503–509, 1996.

[75] N. S. Macklon, J. P. Geraedts, and B. C. Fauser, "Conception to ongoing pregnancy: the 'black box' of early pregnancy loss," *Human Reproduction Update*, vol. 8, no. 4, pp. 333–343, 2002.

[76] A. B. Little, "There's many a slip 'twixt implantation and the crib," pp. 241–242, 1988.

[77] A. Garcıa-Enguıdanos, M. Calle, J. Valero, S. Luna, and V. Domınguez-Rojas, "Risk factors in miscarriage: a review," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 102, no. 2, pp. 111–119, 2002.

[78] The Miscarriage Association, "Recurrent miscarriage," 2011. [Online]. Available: https://www.miscarriageassociation.org.uk/wp-content/uploads/2016/10/Recurrent-Miscarriage.pdf

[79] L. Shahine and R. Lathi, "Recurrent pregnancy loss: evaluation and treatment," *Obstetrics and Gynecology Clinics*, vol. 42, no. 1, pp. 117–134, 2015.

[80] A. Coomarasamy, H. Williams, E. Truchanowicz, P. T. Seed, R. Small, S. Quenby *et al.*, "A randomized trial of progesterone in women with recurrent miscarriages," *New England Journal of Medicine*, vol. 373, no. 22, pp. 2141–2148, 2015.

[81] E. S. Lucas, N. P. Dyer, K. Murakami, Y. H. Lee, Y.-W. Chan, G. Grimaldi *et al.*, "Loss of endometrial plasticity in recurrent pregnancy loss," *Stem Cells*, vol. 34, no. 2, pp. 346–356, 2016.

[82] C. E. Gargett, "Uterine stem cells: what is the evidence?" *Human Reproduction Update*, vol. 13, no. 1, pp. 87–101, 2007.

[83] S. Brahem, M. Mehdi, H. Landolsi, S. Mougou, H. Elghezal, and A. Saad, "Semen parameters and sperm DNA fragmentation as causes of recurrent pregnancy loss," *Urology*, vol. 78, no. 4, pp. 792–796, 2011.

[84] M. Leach, R. J. Aitken, and G. Sacks, "Sperm DNA fragmentation abnormalities in men from couples with a history of recurrent miscarriage," *Australian and New Zealand Journal of Obstetrics and Gynaecology*, vol. 55, no. 4, pp. 379–383, 2015.

[85] E. Borges Jr, B. F. Zanetti, A. S. Setti, D. Braga, R. R. Provenza, and A. Iaconelli Jr, "Sperm DNA fragmentation is correlated with poor embryo development, lower implantation rate, and higher miscarriage rate in reproductive cycles of non–male factor infertility," *Fertility and Sterility*, vol. 112, no. 3, pp. 483–490, 2019.

[86] A. B. Hooker, M. Lemmers, A. L. Thurkow, M. W. Heymans, B. C. Opmeer, H. A. Brölmann, B. W. Mol, and J. A. Huirne, "Systematic review and meta-analysis of intrauterine adhesions after miscarriage: prevalence, risk

factors and long-term reproductive outcome," *Human Reproduction Update*, vol. 20, no. 2, pp. 262–278, 2014.

[87] G. E. Robinson, "Pregnancy loss," *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 28, no. 1, pp. 169 – 178, 2014.

[88] K. M. Shreffler, A. L. Greil, and J. McQuillan, "Pregnancy Loss and Distress Among U.S. Women," *Family Relations*, vol. 60, no. 3, pp. 342–355, 2011.

[89] F. Serrano and M. L. Lima, "Recurrent miscarriage: psychological and relational consequences for couples," *Psychology and Psychotherapy: Theory, Research and Practice*, vol. 79, no. 4, pp. 585–594, 2006.

[90] D. Greenfeld and V. Walther, "Psychological aspects of recurrent pregnancy loss," *Infertility and Reproductive Clinic of North America*, vol. 2, pp. 235–247, 1991.

[91] H. J. Janssen, M. C. Cuisinier, K. P. de Graauw, and K. A. Hoogduin, "A prospective study of risk factors predicting grief intensity following pregnancy loss," *Archives of General Psychiatry*, vol. 54, no. 1, pp. 56–61, 1997.

[92] M. Garel, B. Blondel, N. Lelong, S. Bonenfant, and M. Kaminski, "Long-term consequences of miscarriage: the depressive disorders and the following pregnancy," *Journal of Reproductive and Infant Psychology*, vol. 12, no. 4, pp. 233–240, 1994.

[93] E. R. Blackmore, D. Côté-Arsenault, W. Tang, V. Glover, J. Evans, J. Golding, and T. G. O'Connor, "Previous prenatal loss as a predictor of perinatal depression and anxiety," *The British Journal of Psychiatry*, vol. 198, no. 5, pp. 373–378, 2011.

[94] R. Hasan, D. D. Baird, A. H. Herring, A. F. Olshan, M. L. J. Funk, and K. E. Hartmann, "Association between first-trimester vaginal bleeding and miscarriage," *Obstetrics and Gynecology*, vol. 114, no. 4, pp. 860–867, 2009.

[95] Tommy's, "Miscarriage symptoms," Online, April 2020. [Online]. Available: https://www.tommys.org/baby-loss-support/miscarriage-information-and-support/miscarriage-symptoms

[96] B. S. Shapiro, M. Escobar, R. Makuch, G. Lavy, and A. H. DeCherney, "A model-based prediction for transvaginal ultrasonographic identification of early intrauterine pregnancy," *American Journal of Obstetrics and Gynecology*, vol. 166, no. 5, pp. 1495–1500, 1992.

[97] L. F. Smith, J. Frost, R. Levitas, H. Bradley, and J. Garcia, "Women's experiences of three early miscarriage management options a qualitative study," *British Journal of General Practice*, vol. 56, no. 524, pp. 198–205, 2006.

[98] J. Preisler, J. Kopeika, L. Ismail, V. Vathanan, J. Farren, Y. Abdallah, P. Battacharjee, C. Van Holsbeke, C. Bottomley, D. Gould *et al.*, "Defining safe criteria to diagnose miscarriage: prospective observational multicentre study," *BMJ*, vol. 351, 2015.

[99] E. McCarthy and S. Tong, "Diagnosing a miscarriage: When is it safe to make the call?" *BMJ*, vol. 351, 2015.

[100] D. D. Baird, C. R. Weinberg, D. R. McConnaughey, and A. J. Wilcox, "Rescue of the corpus luteum in human pregnancy," *Biology of Reproduction*, vol. 68, no. 2, pp. 448–456, 2003.

[101] J.-P. Grün, S. Meuris, P. De Nayer, and D. Glinoer, "The thyrotrophic role of human chorionic gonadotrophin (hCG) in the early stages of twin (versus single) pregnancies," *Clinical Endocrinology*, vol. 46, no. 6, pp. 719–725, 1997.

[102] Clearblue, "All you need to know about hcg levels in early pregnancy," Online, Swiss Precision Diagnostics, 2020. [Online]. Available: https:

//uk.clearblue.com/pregnancy-tests/hcg

[103] NHS, "Diagnosis: Miscarriage," Online, National Health Service, jun 2018. [Online]. Available: https://www.nhs.uk/conditions/miscarriage/diagnosis/

[104] S.-I. Cho, M. Goldman, L. Ryan, C. Chen, A. Damokosh, D. Christiani, B. Lasley, J. O'Connor, A. Wilcox, and X. Xu, "Reliability of serial urine hcg as a biomarker to detect early pregnancy loss," *Human Reproduction*, vol. 17, no. 4, pp. 1060–1066, 2002.

[105] M. Prior, C. Bagness, J. Brewin, A. Coomarasamy, L. Easthope, B. Hepworth-Jones *et al.*, "Priorities for research in miscarriage: a priority setting partnership between people affected by miscarriage and professionals following the james lind alliance methodology," *BMJ Open*, vol. 7, no. 8, 2017.

[106] F. Zegers-Hochschild, E. Altieri, C. Fabres, E. Fernandez, A. Mackenna, and P. Orihuela, "Predictive value of human chorionic gonadotrophin in the outcome of early pregnancy after in-vitro fertilization and spontaneous conception," *Human Reproduction*, vol. 9, no. 8, pp. 1550–1555, 1994.

[107] G. Marshall, R. De la Cruz-Mesía, F. A. Quintana, and A. E. Barón, "Discriminant analysis for longitudinal data with multiple continuous responses and possibly missing data," *Biometrics*, vol. 65, no. 1, pp. 69–80, 2009.

[108] P. J. Diggle, P. J. Heagerty, K.-Y. Liang, and S. L. Zeger, *Analysis of longitudinal data*. Oxford University Press, 2002.

[109] E. Paige, J. Barrett, D. Stevens, R. H. Keogh, M. J. Sweeting, I. Nazareth, I. Petersen, and A. M. Wood, "Landmark models for optimizing the use of repeated measurements of risk factors in electronic health records to predict future disease risk," *American Journal of Epidemiology*, vol. 187, no. 7, pp.

1530–1538, 2018.

[110] A. J. Karter, M. M. Parker, H. H. Moffet, M. M. Spence, J. Chan, S. L. Ettner, and J. V. Selby, "Longitudinal study of new and prevalent use of self-monitoring of blood glucose," *Diabetes Care*, vol. 29, no. 8, pp. 1757–1763, 2006.

[111] J. Golding, "The Avon Longitudinal Study of Parents and Children (ALSPAC) - study design and collaborative opportunities," *European Journal of Endocrinology*, vol. 151, no. s3, pp. U119–U123, 2004.

[112] J. K. Lindsey, *Models for repeated measurements*, 2nd ed. Oxford University Press, 1999.

[113] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*, 3rd ed. John Wiley & Sons, 2019.

[114] L. Billingham and K. Abrams, "Simultaneous analysis of quality of life and survival data," *Statistical Methods in Medical Research*, vol. 11, no. 1, pp. 25–48, 2002.

[115] J. N. Matthews, "Summary measures analysis of longitudinal data," *Encyclopedia of Biostatistics*, vol. 8, 2005.

[116] P. Schober and T. R. Vetter, "Repeated measures designs and analysis of longitudinal data: if at first you do not succeed—try, try again," *Anesthesia and Analgesia*, vol. 127, no. 2, p. 569, 2018.

[117] J. Matthews, D. G. Altman, M. Campbell, and P. Royston, "Analysis of serial measurements in medical research." *British Medical Journal*, vol. 300, no. 6719, pp. 230–235, 1990.

[118] D. B. Allison, F. Paultre, C. Maggio, N. Mezzitis, and F. X. Pi-Sunyer, "The use of areas under curves in diabetes research," *Diabetes Care*, vol. 18, no. 2, pp. 245–250, 1995.

[119] S. Senn, "Change from baseline and analysis of covariance revisited," *Statistics in medicine*, vol. 25, no. 24, pp. 4334–4344, 2006.

[120] E. J. Caruana, M. Roman, J. Hernández-Sánchez, and P. Solli, "Longitudinal studies," *Journal of Thoracic Disease*, vol. 7, no. 11, p. E537, 2015.

[121] S. R. Searle, G. Casella, and C. E. McCulloch, *Variance components*. John Wiley & Sons, 2009, ch. 4, pp. 112–167.

[122] G. Keppel, *Design and analysis: A researcher's handbook*. Prentice-Hall, Inc, 1991.

[123] J. De Leeuw, E. Meijer, and H. Goldstein, *Handbook of multilevel analysis*. Springer, 2008.

[124] S. W. Raudenbush and A. S. Bryk, *Hierarchical linear models: applications and data analysis methods*. Sage, 2002.

[125] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982.

[126] H. Goldstein, *Multilevel statistical models*. John Wiley & Sons, 2011.

[127] A. G. Barnett, N. Koper, A. J. Dobson, F. Schmiegelow, and M. Manseau, "Using information criteria to select the correct variance–covariance structure for longitudinal data in ecology," *Methods in Ecology and Evolution*, vol. 1, no. 1, pp. 15–24, 2010.

[128] S. Rabe-Hesketh and A. Skrondal, *Multilevel and longitudinal modeling using Stata*. Stata Press College Station, TX, 2008.

[129] T. A. Snijders and R. J. Bosker, *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Sage, 2011.

[130] J. H. Ryoo, J. D. Long, G. W. Welch, A. Reynolds, and S. M. Swearer, "Fitting the fractional polynomial model to non-gaussian longitudinal data," *Frontiers in Psychology*, vol. 8, p. 1431, 2017.

[131] J. J. Hox, M. Moerbeek, and R. Van de Schoot, *Multilevel analysis: techniques and applications*. Routledge, 2017.

[132] H. Jacqmin-Gadda, S. Sibillot, C. Proust, J.-M. Molina, and R. Thiébaut, "Robustness of the linear mixed model to misspecified error distribution," *Computational Statistics & Data Analysis*, vol. 51, no. 10, pp. 5142–5154, 2007.

[133] C. E. McCulloch and J. M. Neuhaus, "Misspecifying the shape of a random effects distribution: why getting it wrong may not matter," *Statistical Science*, vol. 26, no. 3, pp. 388–402, 2011.

[134] C. J. Maas and J. J. Hox, "Robustness issues in multilevel regression analysis," *Statistica Neerlandica*, vol. 58, no. 2, pp. 127–137, 2004.

[135] A. Bell, M. Fairbrother, and K. Jones, "Fixed and random effects models: making an informed choice," *Quality & Quantity*, vol. 53, no. 2, pp. 1051–1074, 2019.

[136] H. Schielzeth, N. J. Dingemanse, S. Nakagawa, D. F. Westneat, H. Allegue, C. Teplitsky *et al.*, "Robustness of linear mixed-effects models to violations of distributional assumptions," *Methods in Ecology and Evolution*, vol. 11, no. 9, pp. 1141–1152, 2020.

[137] F. Gumedze and T. Dunne, "Parameter estimation and inference in the linear mixed model," *Linear Algebra and its Applications*, vol. 435, no. 8, pp. 1920–1944, 2011.

[138] H. Goldstein, "Likelihood estimation in multilevel models," in *The SAGE handbook of multilevel modeling*, M. A. Scott, J. S. Simonoff, and B. D. Marx, Eds. Sage, 2013, ch. 3, pp. 39–52.

[139] M. J. Lindstrom and D. M. Bates, "Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data," *Journal of the*

American Statistical Association, vol. 83, no. 404, pp. 1014–1022, 1988.

[140] W. Gould, J. Pitblado, and W. Sribney, *Maximum likelihood estimation with Stata*. Stata Press College Station, TX, 2006.

[141] S. R. Searle, G. Casella, and C. E. McCulloch, *Variance components*. John Wiley & Sons, 2009, ch. 6, pp. 232–257.

[142] W. A. Thompson, "The problem of negative estimates of variance components," *Annals of Mathematical Statistics*, vol. 33, no. 1, pp. 273–289, 1962.

[143] R. Steele, "Model selection in multilevel models," in *The SAGE handbook of multilevel modeling*, M. A. Scott, J. S. Simonoff, and B. D. Marx, Eds. Sage, 2013, ch. 7, pp. 109–126.

[144] "The matrix handling of BLUE and BLUP in the mixed linear model, author=Searle, Shayle R," *Linear Algebra and its Applications*, vol. 264, pp. 291–311, 1997.

[145] A. Skrondal and S. Rabe-Hesketh, "Prediction in multilevel generalized linear models," *Journal of the Royal Statistical Society: Series A*, vol. 172, no. 3, pp. 659–687, 2009.

[146] D. Steffey and R. E. Kass, "That blup is a good thing: the estimation of random effects: Comment," *Statistical Science*, vol. 6, no. 1, pp. 45–47, 1991.

[147] S. Rabe-Hesketh, A. Skrondal, and A. Pickles, "Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects," *Journal of Econometrics*, vol. 128, no. 2, pp. 301–323, 2005.

[148] S. L. Zeger, K.-Y. Liang, and P. S. Albert, "Models for longitudinal data: a generalized estimating equation approach," *Biometrics*, vol. 44, no. 4, pp.

1049–1060, 1988.

[149] A. E. Hubbard, J. Ahern, N. L. Fleischer, M. Van der Laan, S. A. Satariano, N. Jewell, T. Bruckner, and W. A. Satariano, "To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health," *Epidemiology*, vol. 21, no. 4, pp. 467–474, 2010.

[150] G. A. Ballinger, "Using generalized estimating equations for longitudinal data analysis," *Organizational Research Methods*, vol. 7, no. 2, pp. 127–150, 2004.

[151] D. Collett, *Modelling survival data in medical research*. CRC press, 2015.

[152] M. Arnold, M. J. Rutherford, A. Bardot, J. Ferlay, T. M. Andersson, T. Å. Myklebust *et al.*, "Progress in cancer survival, mortality, and incidence in seven high-income countries 1995–2014 (icbp survmark-2): a population-based study," *The Lancet Oncology*, vol. 20, no. 11, pp. 1493–1505, 2019.

[153] L. Hogen, D. Vicus, S. E. Ferguson, L. T. Gien, S. Nofech-Mozes, G. K. Lennox, and M. Q. Bernardini, "Patterns of recurrence and impact on survival in patients with clear cell ovarian carcinoma," *International Journal of Gynecologic Cancer*, vol. 29, no. 7, 2019.

[154] J. Emmerson and J. Brown, "Understanding survival analysis in clinical trials," *Clinical Oncology*, vol. 33, no. 1, pp. 12–14, 2021.

[155] P. Schober and T. R. Vetter, "Survival analysis and interpretation of time-to-event data: the tortoise and the hare," *Anesthesia and Analgesia*, vol. 127, no. 3, p. 792, 2018.

[156] D. G. Kleinbaum and M. Klein, *Survival analysis: a self-learning text*, 2nd ed. Springer, 2005.

[157] F. Campigotto and E. Weller, "Impact of informative censoring on the Kaplan-Meier estimate of progression-free survival in phase II clinical trials," *Journal of Clinical Oncology*, vol. 32, no. 27, p. 3068, 2014.

[158] A. J. Templeton, E. Amir, and I. F. Tannock, "Informative censoring - a neglected cause of bias in oncology trials," *Nature Reviews Clinical Oncology*, vol. 17, no. 6, pp. 327–328, 2020.

[159] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, "Survival analysis part i: basic concepts and first analyses," *British Journal of Cancer*, vol. 89, no. 2, pp. 232–238, 2003.

[160] A. C. Thiébaut and J. Bénichou, "Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study," *Statistics in Medicine*, vol. 23, no. 24, pp. 3803–3820, 2004.

[161] R. Lamarca, J. Alonso, G. Gomez, and Á. Muñoz, "Left-truncated data with age as time scale: an alternative for survival analysis in the elderly population," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 53, no. 5, pp. M337–M343, 1998.

[162] M. Cleves, W. Gould, W. W. Gould, R. Gutierrez, and Y. Marchenko, *An introduction to survival analysis using Stata*. Stata Press College Station, TX, 2008.

[163] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[164] R. C. Elandt-Johnson and N. L. Johnson, *Survival models and data analysis*, 1st ed. John Wiley & Sons, 1980.

[165] E. T. Lee and J. Wang, *Statistical methods for survival data analysis*. John Wiley & Sons, 2003, vol. 476.

[166] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data*. John Wiley & Sons, 2002, ch. 1, pp. 1–30.

[167] ——, *The statistical analysis of failure time data*. John Wiley & Sons, 2002.

[168] T. M. Therneau, *Modeling survival data : extending the Cox model*. Springer, 2000.

[169] D. R. Cox, "Partial likelihood," *Biometrika*, vol. 62, no. 2, pp. 269–276, 1975.

[170] N. Breslow, "Covariance analysis of censored survival data," *Biometrics*, vol. 30, no. 1, pp. 89–99, 1974.

[171] D. Lin, "On the breslow estimator," *Lifetime Data Analysis*, vol. 13, no. 4, pp. 471–480, 2007.

[172] F. E. Harrell, *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*, 2nd ed. Springer, 2015.

[173] M. J. Crowther and P. C. Lambert, "A general framework for parametric survival analysis," *Statistics in Medicine*, vol. 33, no. 30, pp. 5280–5297, 2014.

[174] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[175] M. Stone, "Comments on model selection criteria of Akaike and Schwarz," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 41, no. 2, pp. 276–278, 1979.

[176] E. A. Mohammed, C. Naugler, and B. H. Far, "Emerging business intelligence framework for a clinical laboratory through big data analytics,"

Emerging trends in computational biology, bioinformatics, and systems biology: algorithms and software tools, pp. 577–602, 2015.

[177] C. A. Bellera, G. MacGrogan, M. Debled, C. T. de Lara, V. Brouste, and S. Mathoulin-Pélissier, "Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer," *BMC Medical Research Methodology*, vol. 10, no. 1, pp. 1–12, 2010.

[178] M. J. Rutherford, M. J. Crowther, and P. C. Lambert, "The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study," *Journal of Statistical Computation and Simulation*, vol. 85, no. 4, pp. 777–793, 2015.

[179] P. Royston and M. K. Parmar, "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects," *Statistics in Medicine*, vol. 21, no. 15, pp. 2175–2197, 2002.

[180] S. R. Hinchliffe and P. C. Lambert, "Extending the flexible parametric survival model for competing risks," *The Stata Journal*, vol. 13, no. 2, pp. 344–355, 2013.

[181] S. Durrleman and R. Simon, "Flexible regression models with cubic splines," *Statistics in Medicine*, vol. 8, no. 5, pp. 551–561, 1989.

[182] P. C. Lambert and P. Royston, "Further development of flexible parametric models for survival analysis," *The Stata Journal*, vol. 9, no. 2, pp. 265–290, 2009.

[183] Clinical Practice Research Datalink, *CPRD Aurum January 2021 (Version 2021.01.001) [Data set]*. Clinical Practice Research Datalink, 2021. [Online]. Available: https://doi.org/10.11581/D86M-GD03

[184] W. K. Redekop and D. Mladsi, "The faces of personalized medicine: a framework for understanding its meaning and scope," *Value in Health*, vol. 16, no. 6, pp. S4–S9, 2013.

[185] NHS England, "Improving outcomes through personalised medicine," Sep 2016. [Online]. Available: https://www.england.nhs.uk/wp-content/uploads/2016/09/improving-outcomes-personalised-medicine.pdf

[186] J. K. Aronson, "Biomarkers and surrogate endpoints," p. 491, 2005.

[187] J. M. Taylor and Y. Wang, "Surrogate markers and joint models for longitudinal and survival data," *Controlled Clinical Trials*, vol. 23, no. 6, pp. 626–634, 2002.

[188] C. Proust-Lima and J. M. Taylor, "Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach," *Biostatistics*, vol. 10, no. 3, pp. 535–549, 2009.

[189] L. Ferrer, H. Putter, and C. Proust-Lima, "Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment," *Statistical methods in medical research*, vol. 28, no. 12, pp. 3649–3666, 2019.

[190] D. Renard, H. Geys, G. Molenberghs, T. Burzykowski, M. Buyse, T. Vangeneugden, and L. Bijnens, "Validation of a longitudinally measured surrogate marker for a time-to-event endpoint," *Journal of Applied Statistics*, vol. 30, no. 2, pp. 235–247, 2003.

[191] R. B. Geskus, "Which individuals make dropout informative?" *Statistical Methods in Medical Research*, vol. 23, no. 1, pp. 91–106, 2014.

[192] M. S. Wulfsohn and A. A. Tsiatis, "A joint model for survival and longitudinal data measured with error," *Biometrics*, vol. 53, no. 1, pp. 330–339,

1997.

[193] Ö. Asar, J. Ritchie, P. A. Kalra, and P. J. Diggle, "Joint modelling of repeated measurement and time-to-event data: an introductory tutorial," *International Journal of Epidemiology*, vol. 44, no. 1, pp. 334–344, 2015.

[194] R. L. Prentice, "Covariate measurement errors and parameter estimation in a failure time regression model," *Biometrika*, vol. 69, no. 2, pp. 331–342, 1982.

[195] S. Self and Y. Pawitan, "Modeling a marker of disease progression and onset of disease," in *AIDS epidemiology*. Springer, 1992, pp. 231–255.

[196] A. A. Tsiatis and M. Davidian, "Joint modeling of longitudinal and time-to-event data: an overview," *Statistica Sinica*, vol. 14, pp. 809–834, 2004.

[197] C. L. Faucett and D. C. Thomas, "Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach," *Statistics in medicine*, vol. 15, no. 15, pp. 1663–1685, 1996.

[198] E. Brown and J. Ibrahim, "A bayesian semiparametric joint hierarchical model for longitudinal and survival data," *Biometrics*, vol. 59, no. 2, pp. 221–228, 2003.

[199] Y.-K. Tseng, F. Hsieh, and J.-L. Wang, "Joint modelling of accelerated failure time and longitudinal data," *Biometrika*, vol. 92, no. 3, pp. 587–603, 2005.

[200] M. J. Crowther, "merlin—a unified modeling framework for data analysis and methods development in stata," *The Stata Journal*, vol. 20, no. 4, pp. 763–784, 2020.

[201] D. Rizopoulos, "JM: An R package for the joint modelling of longitudinal and time-to-event data," *Journal of Statistical Software (Online)*, vol. 35, no. 9, pp. 1–33, 2010.

[202] ——, "The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC," *Journal of Statistical Software, Articles*, vol. 72, no. 7, pp. 1–46, 2016. [Online]. Available: https://www.jstatsoft.org/v072/i07

[203] P. J. Diggle, P. J. Heagerty, K.-Y. Liang, and S. L. Zeger, *Analysis of longitudinal data*. Oxford University Press, 2002.

[204] G. L. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona, "joinerml: a joint model and software package for time-to-event and multivariate longitudinal outcomes," *BMC Medical Research Methodology*, vol. 18, no. 1, p. 50, 2018.

[205] E.-R. Andrinopoulou, P. H. Eilers, J. J. Takkenberg, and D. Rizopoulos, "Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using p-splines," *Biometrics*, vol. 74, no. 2, pp. 685–693, 2018.

[206] E. P. Pulkstenis, T. R. Ten Have, and J. R. Landis, "Model for the analysis of binary longitudinal pain data subject to informative dropout through remedication," *Journal of the American Statistical Association*, vol. 93, no. 442, pp. 438–450, 1998.

[207] D. Rizopoulos, *Joint models for longitudinal and time-to-event data: with applications in R*. CRC press, 2012, ch. 5, pp. 97–115.

[208] ——, *Joint models for longitudinal and time-to-event data: with applications in R*. CRC press, 2012, ch. 4, pp. 61–77.

[209] K. Lange, "The EM algorithm," in *Optimization*. Springer, 2013, pp. 221–244.

[210] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. Springer, 2009.

[211] S. Rabe-Hesketh and A. Skrondal, "Generalized linear mixed models," in *International Encyclopedia of Education*, 3rd ed., P. Peterson, E. Baker, and B. McGaw, Eds., 2010, pp. 171–177.

[212] J. C. Naylor and A. F. Smith, "Applications of a method for the efficient computation of posterior distributions," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 31, no. 3, pp. 214–225, 1982.

[213] F. Tuerlinckx, F. Rijmen, G. Verbeke, and P. De Boeck, "Statistical inference in generalized linear mixed models: a review," *British Journal of Mathematical and Statistical Psychology*, vol. 59, no. 2, pp. 225–255, 2006.

[214] S. Rabe-Hesketh, A. Skrondal, and A. Pickles, "Reliable estimation of generalized linear mixed models using adaptive quadrature," *The Stata Journal*, vol. 2, no. 1, pp. 1–21, 2002.

[215] J. C. Pinheiro and D. M. Bates, "Approximations to the log-likelihood function in the nonlinear mixed-effects model," *Journal of computational and Graphical Statistics*, vol. 4, no. 1, pp. 12–35, 1995.

[216] D. Rizopoulos, "Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule," *Computational Statistics & Data Analysis*, vol. 56, no. 3, pp. 491–501, 2012.

[217] A. Perperoglou, W. Sauerbrei, M. Abrahamowicz, and M. Schmid, "A review of spline function procedures in R," *BMC Medical Research Methodology*, vol. 19, no. 1, pp. 1–16, 2019.

[218] D. Laurie, "Calculation of Gauss-Kronrod quadrature rules," *Mathematics of Computation*, vol. 66, no. 219, pp. 1133–1145, 1997.

[219] W. Gautschi and T. J. Rivlin, "A family of gauss-kronrod quadrature formulae," *Mathematics of computation*, vol. 51, no. 184, pp. 749–754, 1988.

[220] P. Philipson, I. Sousa, P. Diggle, P. Williamson, R. Kolamunnage-Dona, R. Henderson, and G. Hickey, "Package "joineR" v 1.2.6," *R Foundation for Statistical Computing*, 2021.

[221] D. Rizopoulos, "Package "JMbayes2" v 0.1.3," *R Foundation for Statistical Computing*, 2021.

[222] A. Garcia-Hernandez and D. Rizopoulos, "JM: A SAS Macro to Fit Jointly Generalized Mixed Models for Longitudinal Data and Time-to-Event Responses," *Journal of Statistical Software*, vol. 84, no. 12, pp. 1–29, 2018.

[223] D. Zhang, M.-H. Chen, J. G. Ibrahim, M. E. Boye, and W. Shen, "JMFit: a SAS macro for joint models of longitudinal and survival data," *Journal of Statistical Software*, vol. 71, no. 3, 2016.

[224] M. Sudell, R. Kolamunnage-Dona, and C. Tudur-Smith, "Joint models for longitudinal and time-to-event data: a review of reporting quality with a view to meta-analysis," *BMC Medical Research Methodology*, vol. 16, no. 1, pp. 1–11, 2016.

[225] D. Zhang, M.-H. Chen, J. G. Ibrahim, M. E. Boye, P. Wang, and W. Shen, "Assessing model fit in joint models of longitudinal and survival data with applications to cancer clinical trials," *Statistics in Medicine*, vol. 33, no. 27, pp. 4715–4733, 2014.

[226] Z. He, W. Tu, S. Wang, H. Fu, and Z. Yu, "Simultaneous variable selection for joint models of longitudinal and survival outcomes," *Biometrics*, vol. 71, no. 1, pp. 178–187, 2015.

[227] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[228] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[229] L. Ferrer, H. Putter, and C. Proust-Lima, "Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment," *Statistical Methods in Medical Research*, vol. 28, no. 12, pp. 3649–3666, 2019.

[230] M. J. Sweeting, "Using predictions from a joint model for longitudinal and survival data to inform the optimal time of intervention in an abdominal aortic aneurysm screening programme," *Biometrical Journal*, vol. 59, no. 6, pp. 1247–1260, 2017.

[231] M. J. Pencina, R. B. D'Agostino Sr, R. B. D'Agostino Jr, and R. S. Vasan, "Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond," *Statistics in Medicine*, vol. 27, no. 2, pp. 157–172, 2008.

[232] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher, "Assessment and comparison of prognostic classification schemes for survival data," *Statistics in Medicine*, vol. 18, no. 17-18, pp. 2529–2545, 1999.

[233] R. Henderson, P. Diggle, and A. Dobson, "Identification and efficacy of longitudinal markers for survival," *Biostatistics*, vol. 3, no. 1, pp. 33–50, 2002.

[234] C. Proust-Lima, M. Séne, J. M. Taylor, and H. Jacqmin-Gadda, "Joint latent class models for longitudinal and time-to-event data: A review," *Statistical Methods in Medical Research*, vol. 23, no. 1, pp. 74–90, 2014.

[235] H. Lin, B. W. Turnbull, C. E. McCulloch, and E. H. Slate, "Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate

cancer," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 53–65, 2002.

[236] C. Proust-Lima, P. Joly, J.-F. Dartigues, and H. Jacqmin-Gadda, "Joint modelling of multivariate longitudinal outcomes and a time-to-event: a nonlinear latent class approach," *Computational Statistics & Data Analysis*, vol. 53, no. 4, pp. 1142–1154, 2009.

[237] H. Putter and H. C. van Houwelingen, "Landmarking 2.0: bridging the gap between joint models and landmarking," *Statistics in Medicine*, pp. 1–17, 2022.

[238] H. C. Van Houwelingen, "Dynamic prediction by landmarking in event history analysis," *Scandinavian Journal of Statistics*, vol. 34, no. 1, pp. 70–85, 2007.

[239] H. van Houwelingen and H. Putter, *Dynamic prediction in clinical survival analysis*. CRC Press, 2011, ch. 8, pp. 121–133.

[240] K. Suresh, J. M. Taylor, D. E. Spratt, S. Daignault, and A. Tsodikov, "Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model," *Biometrical Journal*, vol. 59, no. 6, pp. 1277–1300, 2017.

[241] N. Ashra, L. Marriott, S. Johnson, K. Abrams, and M. Crowther, "Jointly modelling longitudinally measured urinary human chorionic gonadotrophin and early pregnancy outcomes," *Scientific Reports*, vol. 10, no. 1, pp. 1–9, 2020.

[242] K. Barnhart, M. D. Sammel, K. Chung, L. Zhou, A. C. Hummel, and W. Guo, "Decline of serum human chorionic gonadotropin and spontaneous complete abortion: defining the normal curve," *Obstetrics & Gynecology*, vol. 104, no. 5, pp. 975–81, 2004.

[243] J. Kerin, "Ovulation detection in the human." *Clinical reproduction and fertility*, vol. 1, no. 1, pp. 27–54, 1982.

[244] P. B. Miller and M. R. Soules, "The usefulness of a urinary lh kit for ovulation prediction during menstrual cycles of normal women," *Obstetrics & Gynecology*, vol. 87, no. 1, pp. 13–17, 1996.

[245] H. M. Behre, J. Kuhlage, C. Gaβner, B. Sonntag, C. Schem, H. P. Schneider, and E. Nieschlag, "Prediction of ovulation by urinary hormone measurements with the home use clearplan® fertility monitor: comparison with transvaginal ultrasound scans and serum hormone measurements," *Human Reproduction*, vol. 15, no. 12, pp. 2478–2482, 2000.

[246] A. Nardi and M. Schemper, "Comparing Cox and parametric models in clinical studies," *Statistics in Medicine*, vol. 22, no. 23, pp. 3597–3610, 2003.

[247] A.-M. N. Andersen, J. Wohlfahrt, P. Christens, J. Olsen, and M. Melbye, "Maternal age and fetal loss: population based register linkage study," *BMJ*, vol. 320, no. 7251, pp. 1708–1712, 2000.

[248] A. M. Z. Jukic, C. R. Weinberg, D. D. Baird, and A. J. Wilcox, "Lifestyle and reproductive factors associated with follicular phase length," *Journal of Women's Health*, vol. 16, no. 9, pp. 1340–1347, 2007.

[249] C. M. Small, A. K. Manatunga, M. Klein, H. S. Feigelson, C. E. Dominguez, R. McChesney, and M. Marcus, "Menstrual cycle characteristics: associations with fertility and spontaneous abortion," *Epidemiology*, no. 1, pp. 52–60, 2006.

[250] F. Miro, S. W. Parker, L. J. Aspinall, J. Coley, P. W. Perry, and J. E. Ellis, "Sequential classification of endocrine stages during reproductive aging in women: the FREEDOM study," *Menopause*, vol. 12, no. 3, pp. 281–290, 2005.

[251] S. Feodor Nilsson, P. Andersen, K. Strandberg-Larsen, and A.-M. Nybo Andersen, "Risk factors for miscarriage from a prevention perspective: a nationwide follow-up study," *BJOG*, vol. 121, no. 11, pp. 1375–1385, 2014.

[252] S. Senapati and K. T. Barnhart, "Biomarkers for ectopic pregnancy and pregnancy of unknown location," *Fertility and Sterility*, vol. 99, no. 4, pp. 1107–1116, 2013.

[253] S. Tong, E. M. Wallace, and L. Rombauts, "Association between low day 16 hCG and miscarriage after proven cardiac activity," *Obstetrics & Gynecology*, vol. 107, no. 2, pp. 300–304, 2006.

[254] J. Gardosi, "Dating of pregnancy: time to forget the last menstrual period," *Ultrasound in Obstetrics & Gynecology*, vol. 9, no. 6, pp. 367–368, 1997.

[255] S. Johnson and L. Marriott, "hCG levels can decline pre- or post- onset of bleeding in early loss," in *Fertility 2019*, 2019, p. PO 99.

[256] S. Tiplady, G. Jones, M. Campbell, S. Johnson, and W. Ledger, "Home ovulation tests and stress in women trying to conceive: a randomized controlled trial," *Human Reproduction*, vol. 28, no. 1, pp. 138–151, 2013.

[257] S. Weddell, G. L. Jones, S. Duffy, C. Hogg, S. Johnson, and W. Ledger, "Home ovulation test use and stress during subfertility evaluation: subarm of a randomized controlled trial," *Women's Health*, vol. 15, 2019.

[258] P. M. Lukacs, K. P. Burnham, and D. R. Anderson, "Model selection bias and Freedman's paradox," *Annals of the Institute of Statistical Mathematics*, vol. 62, no. 1, p. 117, 2010.

[259] R. H. Gray and L. Y. Wu, "Subfertility and risk of spontaneous abortion." *American journal of public health*, vol. 90, no. 9, p. 1452, 2000.

[260] S. Kamalanathan, J. P. Sahoo, and T. Sathyapalan, "Pregnancy in polycystic ovary syndrome," *Indian Journal of Endocrinology and Metabolism*,

vol. 17, no. 1, p. 37, 2013.

[261] E.-R. Andrinopoulou, D. Rizopoulos, J. J. Takkenberg, and E. Lesaffre, "Combined dynamic predictions using joint models of two longitudinal outcomes and competing risk data," *Statistical Methods in Medical Research*, vol. 26, no. 4, pp. 1787–1801, 2017.

[262] C. Gnoth, D. Godehardt, E. Godehardt, P. Frank-Herrmann, and G. Freundl, "Time to pregnancy: results of the german prospective study and impact on the management of infertility," *Human Reproduction*, vol. 18, no. 9, pp. 1959–1966, 2003.

[263] B. Lunenfeld, W. Bilger, S. Longobardi, V. Alam, T. D'Hooghe, and S. K. Sunkara, "The development of gonadotropins for clinical use in the treatment of infertility," *Frontiers in Endocrinology*, vol. 10, p. 429, 2019.

[264] S. Johnson, L. Marriott, and M. Zinaman, "Can apps and calendar methods predict ovulation with accuracy?" *Current Medical Research and Opinion*, vol. 34, no. 9, pp. 1587–1594, 2018.

[265] S. Johnson, S. Weddell, S. Godbert, G. Freundl, J. Roos, and C. Gnoth, "Development of the first urinary reproductive hormone ranges referenced to independently determined ovulation day," *Clinical Chemistry and Laboratory Medicine*, vol. 53, no. 7, pp. 1099–1108, 2015.

[266] J. Balasch, "The role of FSH and LH in ovulation induction: current concepts," in *Textbook of assisted reproductive techniques*, 4th ed., D. K. Gardner, A. Weissman, C. M. Howles, and Z. Shoham, Eds. CRC Press, 2012, pp. 75–98.

[267] T. I. Korevaar, E. A. Steegers, Y. B. de Rijke, S. Schalekamp-Timmermans, W. E. Visser, A. Hofman *et al.*, "Reference ranges and determinants of total hCG levels during pregnancy: the Generation R Study," *European Journal*

of Epidemiology, vol. 30, no. 9, pp. 1057–1066, 2015.

[268] D. Tulchinsky, "Endocrine assessments of fetal-placental well-being," *Clinics in Perinatology*, vol. 10, no. 3, pp. 763–776, 1983.

[269] M. J. Zinaman, S. Johnson, and G. Warren, "hCG and FSH Levels During Early Pregnancy and Reproductive Aging [24A]," *Obstetrics & Gynecology*, vol. 135, p. 14S, 2020.

[270] J. Y. Goh, S. He, J. C. Allen, R. Malhotra, and T. C. Tan, "Maternal obesity is associated with a low serum progesterone level in early pregnancy," *Hormone Molecular Biology and Clinical Investigation*, vol. 27, no. 3, pp. 97–100, 2016.

[271] D. Rizopoulos. (2018) Multivariate joint models. [Online]. Available: https://www.drizopoulos.com/vignettes/multivariate%20joint%20models

[272] K. E. Dillon, V. D. Sioulas, M. D. Sammel, K. Chung, P. Takacs, A. Shaunik, and K. T. Barnhart, "How and when human chorionic gonadotropin curves in women with an ectopic pregnancy mimic other outcomes: differences by race and ethnicity," *Fertility and Sterility*, vol. 98, no. 4, pp. 911–916, 2012.

[273] A. Póvoa, P. Xavier, A. Matias, and I. Blickstein, "First trimester β-hcg and estradiol levels in singleton and twin pregnancies after assisted reproduction," *Journal of Perinatal Medicine*, vol. 46, no. 8, pp. 853–856, 2018.

[274] J. Rogers and G. W. Mitchell Jr, "The relation of obesity to menstrual disturbances," *New England Journal of Medicine*, vol. 247, no. 2, pp. 53–55, 1952.

[275] T. Brodin, T. Bergh, L. Berglund, N. Hadziosmanovic, and J. Holte, "Menstrual cycle length is an age-independent marker of female fertility: results from 6271 treatment cycles of in vitro fertilization," *Fertility and Sterility*,

vol. 90, no. 5, pp. 1656–1661, 2008.

[276] A. M. Lower and J. L. Yovich, "The value of serum levels of oestradiol, progesterone and β-human chorionic gonadotrophin in the prediction of early pregnancy loss," *Human Reproduction*, vol. 7, no. 5, pp. 711–717, 1992.

[277] M. K. Lim, C. W. Ku, T. C. Tan, Y. H. J. Lee, J. C. Allen, and N. S. Tan, "Characterisation of serum progesterone and progesterone-induced blocking factor (PIBF) levels across trimesters in healthy pregnant women," *Scientific Reports*, vol. 10, no. 1, pp. 1–9, 2020.

[278] P. G. Whittaker, C. A. Schreiber, and M. D. Sammel, "Gestational hormone trajectories and early pregnancy failure: a reassessment," *Reproductive Biology and Endocrinology*, vol. 16, no. 1, pp. 1–6, 2018.

[279] J. Verhaegen, I. D. Gallos, N. M. Van Mello, M. Abdel-Aziz, Y. Takwoingi, H. Harb, J. J. Deeks, B. W. Mol, and A. Coomarasamy, "Accuracy of single progesterone test to predict early pregnancy outcome in women with pain or bleeding: meta-analysis of cohort studies," *BMJ*, vol. 345, 2012.

[280] M. G. Phipps, J. W. Hogan, J. F. Peipert, G. M. Lambert-Messerlian, J. A. Canick, and D. B. Seifer, "Progesterone, inhibin, and hCG multiple marker strategy to differentiate viable from nonviable pregnancies," *Obstetrics & Gynecology*, vol. 95, no. 2, pp. 227–231, 2000.

[281] G. Jasienska and M. Jasienski, "Interpopulation, interindividual, intercycle, and intracycle natural variation in progesterone levels: a quantitative assessment and implications for population studies," *American Journal of Human Biology*, vol. 20, no. 1, pp. 35–42, 2008.

[282] J. Roos, S. Johnson, S. Weddell, E. Godehardt, J. Schiffner, G. Freundl, and C. Gnoth, "Monitoring the menstrual cycle: comparison of urinary and

serum reproductive hormones referenced to true ovulation," *The European Journal of Contraception & Reproductive Health Care*, vol. 20, no. 6, pp. 438–450, 2015.

[283] A. Eskild, P. Fedorcsak, L. Mørkrid, and T. G. Tanbo, "Maternal body mass index and serum concentrations of human chorionic gonadotropin in very early pregnancy," *Fertility and Sterility*, vol. 98, no. 4, pp. 905–910, 2012.

[284] A. M. Gronowski and D. G. Grenache, "Characterization of the hCG variants recognized by different hCG immunoassays: an important step toward standardization of hCG measurements," 2009.

[285] G. Vranken, T. Reynolds, and J. Van Nueten, "Medians for second-trimester maternal serum markers: geographical differences and variation caused by median multiples-of-median equations," *Journal of Clinical Pathology*, vol. 59, no. 6, pp. 639–644, 2006.

[286] S. Dukhovny, C. Zera, S. E. Little, T. McElrath, and L. Wilkins-Haug, "Eliminating first trimester markers: will replacing PAPP-A and β hCG miss women at risk for small for gestational age?" *The Journal of Maternal-Fetal & Neonatal Medicine*, vol. 27, no. 17, pp. 1761–1764, 2014.

[287] T. Leung, L. Chan, T. Leung, T. Fung, D. Sahota, and T. Lau, "First-trimester maternal serum levels of placental hormones are independent predictors of second-trimester fetal growth parameters," *Ultrasound in Obstetrics and Gynecology*, vol. 27, no. 2, pp. 156–161, 2006.

[288] M. R. Chernick and R. A. LaBudde, *An introduction to bootstrap methods with applications to R.* John Wiley & Sons, 2014.

[289] M. E. Alfaro, S. Zoller, and F. Lutzoni, "Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo

sampling and bootstrapping in assessing phylogenetic confidence," *Molecular Biology and Evolution*, vol. 20, no. 2, pp. 255–266, 2003.

[290] R. Kolamunnage-Dona and A. N. Kamarudin, "Adjustment for the measurement error in evaluating biomarker performances at baseline for future survival outcomes: time-dependent receiver operating characteristic curve within a joint modelling framework," *Research Methods in Medicine & Health Sciences*, vol. 2, no. 2, pp. 51–60, 2021.

[291] D. M. Powers, "What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes," *arXiv preprint arXiv:1503.06410*, 2015.

[292] M. Rota and L. Antolini, "Finding the optimal cut-point for Gaussian and Gamma distributed biomarkers," *Computational Statistics & Data Analysis*, vol. 69, pp. 1–14, 2014.

[293] J. Perkins Neil and F. Schisterman Enrique, "The inconsistency of "optimal" cut-points using two ROC based criteria," *American Journal of Epidemiology*, vol. 163, no. 7, pp. 670–675, 2006.

[294] X. Liu, "Classification accuracy and cut point selection," *Statistics in Medicine*, vol. 31, no. 23, pp. 2676–2686, 2012.

[295] S. Dijkland, I. Retel Helmrich, and E. Steyerberg, "Validation of prognostic models: challenges and opportunities," *Journal of Emergency and Critical Care Medicine*, vol. 2, no. 91, pp. 1–4, 2018.

[296] D. G. Altman, Y. Vergouwe, P. Royston, and K. G. Moons, "Prognosis and prognostic research: validating a prognostic model," *BMJ*, vol. 338, 2009.

[297] J. Oates, I. Casikar, A. Campain, S. Müller, J. Yang, S. Reid, and G. Condous, "A prediction model for viability at the end of the first trimester after a single early pregnancy evaluation," *Australian and New Zealand Journal of Obstetrics and Gynaecology*, vol. 53, no. 1, pp. 51–57, 2013.

[298] I. Soumpasis, B. Grace, and S. Johnson, "Real-life insights on menstrual cycles and ovulation using big data," *Human Reproduction Open*, vol. 2020, no. 2, 2020.

[299] J. B. Stanford, G. L. White Jr, and H. Hatasaka, "Timing intercourse to achieve pregnancy: current evidence," *Obstetrics & Gynecology*, vol. 100, no. 6, pp. 1333–1341, 2002.

[300] D. Dunson, C. Weinberg, D. Baird, J. Kesner, and A. Wilcox, "Assessing human fertility using several markers of ovulation," *Statistics in Medicine*, vol. 20, no. 6, pp. 965–978, 2001.

[301] J. C. Barrett and J. Marshall, "The risk of conception on different days of the menstrual cycle," *Population Studies*, vol. 23, no. 3, pp. 455–461, 1969.

[302] D. Schwartz, P. MacDonald, and V. Heuchel, "Fecundability, coital frequency and the viability of ova," *Population Studies*, vol. 34, no. 2, pp. 397–400, 1980.

[303] B. Colombo, G. Masarotto, and M. C. F. S. Group, "Daily fecundability: first results from a new data base," *Demographic Research*, vol. 3, 2000.

[304] J. P. Royston, "Basal body temperature, ovulation and the risk of conception, with special reference to the lifetimes of sperm and egg," *Biometrics*, vol. 38, no. 2, pp. 397–406, 1982.

[305] J. E. Robinson, M. Wakelin, and J. E. Ellis, "Increased pregnancy rate with use of the Clearblue Easy Fertility Monitor," *Fertility and Sterility*, vol. 87, no. 2, pp. 329–334, 2007.

[306] S. Johnson, J. B. Stanford, G. Warren, S. Bond, S. Bench-Capon, and M. J. Zinaman, "Increased likelihood of pregnancy using an app-connected ovulation test system: a randomized controlled trial," *Journal of Women's Health*, vol. 29, no. 1, pp. 84–90, 2020.

[307] P. A. Carpentier, J. B. Stanford, and P. C. Boyle, "Progesterone in women with recurrent miscarriages," *New England Journal of Medicine*, vol. 374, no. 9, p. 894, 2016.

[308] A. Moscrop, "Can sex during pregnancy cause a miscarriage? a concise history of not knowing," *British Journal of General Practice*, vol. 62, no. 597, pp. e308–e310, 2012.

[309] C. Fox, H. Wolff, and J. Baker, "Measurement of intra-vaginal and intra-uterine pressures during human coitus by radio-telemetry," *Reproduction*, vol. 22, no. 2, pp. 243–251, 1970.

[310] S. A. Robertson, J. R. Prins, D. J. Sharkey, and L. M. Moldenhauer, "Seminal fluid and the generation of regulatory t cells for embryo implantation," *American journal of reproductive immunology*, vol. 69, no. 4, pp. 315–330, 2013.

[311] J. B. Stanford, J. L. Hansen, S. K. Willis, N. Hu, and A. Thomas, "Periimplantation intercourse does not lower fecundability," *Human Reproduction*, vol. 35, no. 9, pp. 2107–2112, 2020.

[312] A. Adolfsson, P.-G. Larsson, B. Wijma, and C. Bertero, "Guilt and emptiness: women's experiences of miscarriage," *Health Care for Women International*, vol. 25, no. 6, pp. 543–560, 2004.

[313] J. L. Simpson, R. H. Gray, J. T. Queenan, R. T. Kambic, A. Pérez, P. Mena, M. Barbato, F. Pardo, G. Tagliabue, A. Bitto *et al.*, "Fetal outcome among pregnancies in natural family planning acceptors: an international cohort study," *American Journal of Obstetrics and Gynecology*, vol. 165, no. 6, pp. 1981–1982, 1991.

[314] J. Simpson, R. H. Gray, J. Queenan, P. Mena, A. Perez, R. Kambic, G. Tagliabue, F. Pardo, W. Stevenson, M. Barbato *et al.*, "Pregnancy outcome associated with natural family planning (NFP): scientific basis and experimental design for an international cohort study," *Advances in Contraception*, vol. 4, no. 4, pp. 247–264, 1988.

[315] A. J. Wilcox, C. R. Weinberg, and D. D. Baird, "Timing of sexual intercourse in relation to ovulation—effects on the probability of conception, survival of the pregnancy, and sex of the baby," *New England Journal of Medicine*, vol. 333, no. 23, pp. 1517–1521, 1995.

[316] C. G. Petersen, A. L. Mauri, L. D. Vagnini, A. Renzi, B. Petersen, M. Mattila, V. Comar, J. Ricci, F. Dieamant, J. B. A. Oliveira *et al.*, "The effects of male age on sperm DNA damage: an evaluation of 2,178 semen samples," *JBRA Assisted Reproduction*, vol. 22, no. 4, p. 323, 2018.

[317] D. B. McQueen, J. Zhang, and J. C. Robins, "Sperm DNA fragmentation and recurrent pregnancy loss: a systematic review and meta-analysis," *Fertility and Sterility*, vol. 112, no. 1, pp. 54–60, 2019.

[318] D. A. Driscoll, S. J. Kilpatrick, E. R. M. Jauniaux, H. L. Galan, V. Berghella, M. B. Landon, W. A. Grobman, and A. G. Cahill, *Obstetrics: Normal and Problem Pregnancies E-Book*. Saunders, 2020.

[319] C. Jones, C. Chan, and D. Farine, "Sex in pregnancy," *Canadian Medical Association Journal*, vol. 183, no. 7, pp. 815–818, 2011.

[320] E. Bartellas, J. M. Crane, M. Daley, K. A. Bennett, and D. Hutchens, "Sexuality and sexual activity in pregnancy," *BJOG*, vol. 107, no. 8, pp. 964–968, 2000.

[321] S. Oruç, A. Esen, S. Laçin, H. Adigüzel, Y. Uyar, and F. Koyuncu, "Sexual behaviour during pregnancy," *Australian and New Zealand journal of*

obstetrics and gynaecology, vol. 39, no. 1, pp. 48–50, 1999.

[322] A. Jukic, C. Weinberg, D. Baird, and A. Wilcox, "The association of maternal factors with delayed implantation and the initial rise of urinary human chorionic gonadotrophin," *Human Reproduction*, vol. 26, no. 4, pp. 920–926, 2011.

[323] A. R. Baerwald, G. P. Adams, and R. A. Pierson, "Form and function of the corpus luteum during the human menstrual cycle," *Ultrasound in Obstetrics and Gynecology*, vol. 25, no. 5, pp. 498–507, 2005.

[324] D. M. Haas and P. S. Ramsey, "Progestogen for preventing miscarriage in women with recurrent miscarriage of unclear etiology," *Cochrane Database of Systematic Reviews*, no. 11, 2019.

[325] L. Robinson, I. D. Gallos, S. J. Conner, M. Rajkhowa, D. Miller, S. Lewis, J. Kirkman-Brown, and A. Coomarasamy, "The effect of sperm DNA fragmentation on miscarriage rates: a systematic review and meta-analysis," *Human Reproduction*, vol. 27, no. 10, pp. 2908–2917, 2012.

[326] I. Tur-Kaspa, Y. Maor, D. Levran, M. Yonish, S. Mashiach, and J. Dor, "How often should infertile men have intercourse to achieve conception?" *Fertility and Sterility*, vol. 62, no. 2, pp. 370–375, 1994.

[327] National Institue for Health and Care Excellence, "Clinical guideline [CG156] Fertility problems: assessment and treatment," NICE, Sep 2017. [Online]. Available: https://www.nice.org.uk/guidance/cg156

[328] E. O. Ogundimu, D. G. Altman, and G. S. Collins, "Adequate sample size for developing prediction models is not simply related to events per variable," *Journal of Clinical Epidemiology*, vol. 76, pp. 175–182, 2016.

[329] P. Peduzzi, J. Concato, A. R. Feinstein, and T. R. Holford, "Importance of events per independent variable in proportional hazards regression analysis ii. accuracy and precision of regression estimates," *Journal of Clinical Epidemiology*, vol. 48, no. 12, pp. 1503–1510, 1995.

[330] E. Vittinghoff and C. E. McCulloch, "Relaxing the rule of ten events per variable in logistic and cox regression," *American Journal of Epidemiology*, vol. 165, no. 6, pp. 710–718, 2007.

[331] H.-W. Su, Y.-C. Yi, T.-Y. Wei, T.-C. Chang, and C.-M. Cheng, "Detection of ovulation, a review of currently available methods," *Bioengineering & Translational Medicine*, vol. 2, no. 3, pp. 238–246, 2017.

[332] J. R. Bull, S. P. Rowland, E. B. Scherwitzl, R. Scherwitzl, K. G. Danielsson, and J. Harper, "Real-world menstrual cycle characteristics of more than 600,000 menstrual cycles," *NPJ Digital Medicine*, vol. 2, no. 1, pp. 1–8, 2019.

[333] V. Pino, A. Sanz, N. Valdés, J. Crosby, and A. Mackenna, "The effects of aging on semen parameters and sperm DNA fragmentation," *JBRA Assisted Reproduction*, vol. 24, no. 1, p. 82, 2020.

[334] M. Perera, C. Tsokos *et al.*, "A statistical model with non-linear effects and non-proportional hazards for breast cancer survival analysis," *Advances in Breast Cancer Research*, vol. 7, no. 1, p. 65, 2018.

[335] W. Sauerbrei, P. Royston, H. Bojar, C. Schmoor, and M. Schumacher, "Modelling the effects of standard prognostic factors in node-positive breast cancer," *British Journal of Cancer*, vol. 79, no. 11, pp. 1752–1760, 1999.

[336] M. Laxy, R. Stark, A. Peters, H. Hauner, R. Holle, and C. M. Teuner, "The non-linear relationship between bmi and health care costs and the resulting cost fraction attributable to obesity," *International Journal of*

Environmental Research and Public Health, vol. 14, no. 9, p. 984, 2017.

[337] F. Zaccardi, N. N. Dhalwani, D. Papamargaritis, D. R. Webb, G. J. Murphy, M. J. Davies, and K. Khunti, "Nonlinear association of bmi with all-cause and cardiovascular mortality in type 2 diabetes mellitus: a systematic review and meta-analysis of 414,587 participants in prospective studies," *Diabetologia*, vol. 60, no. 2, pp. 240–248, 2017.

[338] M. J. Crowther, P. C. Lambert, and K. R. Abrams, "Adjusting for measurement error in baseline prognostic biomarkers included in a time-to-event analysis: a joint modelling approach," *BMC Medical Research Methodology*, vol. 13, no. 1, pp. 1–8, 2013.

[339] A. Sayers, J. Heron, A. D. Smith, C. Macdonald-Wallis, M. Gilthorpe, F. Steele, and K. Tilling, "Joint modelling compared with two stage methods for analysing longitudinal data and prospective outcomes: a simulation study of childhood growth and BP," *Statistical Methods in Medical Research*, vol. 26, no. 1, pp. 437–452, 2017.

[340] J. Hutton and P. Solomon, "Parameter orthogonality in mixed regression models for survival data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 59, no. 1, pp. 125–136, 1997.

[341] G. Kwong and J. Hutton, "Choice of parametric models in survival analysis: applications to monotherapy for epilepsy and cerebral palsy," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 52, no. 2, pp. 153–168, 2003.

[342] X. Song, M. Davidian, and A. A. Tsiatis, "A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data," *Biometrics*, vol. 58, no. 4, pp. 742–753, 2002.

[343] E.-R. Andrinopoulou, M. O. Harhay, S. J. Ratcliffe, and D. Rizopoulos, "Reflections on modern methods: Dynamic prediction using joint models of longitudinal and time-to-event data," *International Journal of Epidemiology*, 2021.

[344] P. A. Murtaugh, E. R. Dickson, G. M. Van Dam, M. Malinchoc, P. M. Grambsch, A. L. Langworthy, and C. H. Gips, "Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits," *Hepatology*, vol. 20, no. 1, pp. 126–134, 1994.

[345] M. J. Crowther and P. C. Lambert, "Simulating complex survival data," *The Stata Journal*, vol. 12, no. 4, pp. 674–687, 2012.

[346] ——, "Simulating biologically plausible complex survival data," *Statistics in Medicine*, vol. 32, no. 23, pp. 4118–4134, 2013.

[347] R. Bender, T. Augustin, and M. Blettner, "Generating survival times to simulate Cox proportional hazards models," *Statistics in Medicine*, vol. 24, no. 11, pp. 1713–1723, 2005.

[348] T. P. Morris, I. R. White, and M. J. Crowther, "Using simulation studies to evaluate statistical methods," *Statistics in Medicine*, vol. 38, no. 11, pp. 2074–2102, 2019.

[349] J. Bretagnolle and C. Huber-Carol, "Effects of omitting covariates in Cox's model for survival data," *Scandinavian Journal of Statistics*, vol. 15, no. 2, pp. 125–138, 1988.

[350] A. J. Wilcox, C. R. Weinberg, R. E. Wehmann, E. G. Armstrong, R. E. Canfield, and B. C. Nisula, "Measuring early pregnancy loss: laboratory and field methods," *Fertility and Sterility*, vol. 44, no. 3, pp. 366–374, 1985.

[351] N. Kuhadiya, A. Karunakara, M. Garg, R. Katkar, and P. Mason, "A case report of elevated hcg levels in menopause—a clinical dilemma," *Endocrinol*

Metab Int J, vol. 4, no. 2, pp. 46–48, 2017.

[352] S. Khattri, A. Vivekanandarajah, S. Varma, and F. Kong, "Secretion of beta-human chorionic gonadotropin by non-small cell lung cancer: a case report," *Journal of Medical Case Reports*, vol. 5, no. 1, pp. 1–4, 2011.

[353] D. Li, X. Wen, L. Ghali, F. Al-Shalabi, S. M. Docherty, P. Purkis, and R. K. Iles, "hcgβ expression by cervical squamous carcinoma–in vivo histological association with tumour invasion and apoptosis," *Histopathology*, vol. 53, no. 2, pp. 147–155, 2008.

[354] L. M. Chen, J. G. Ibrahim, and H. Chu, "Sample size and power determination in joint modeling of longitudinal and survival data," *Statistics in Medicine*, vol. 30, no. 18, pp. 2295–2309, 2011.

[355] P. Kumar and N. Magon, "Hormones in pregnancy," *Nigerian Medical Journal*, vol. 53, no. 4, p. 179, 2012.

[356] S. Guha, F. Ayim, J. Ludlow, A. Sayasneh, G. Condous, E. Kirk, C. Stalder, D. Timmerman, T. Bourne, and B. V. Calster, "Triaging pregnancies of unknown location: the performance of protocols based on single serum progesterone or repeated serum hcg levels," *Human Reproduction*, vol. 29, no. 5, pp. 938–945, 2014.

[357] V. K. Kadam, S. Agrawal, P. Saxena, and P. Laul, "Predictive value of single serum progesterone level for viability in threatened miscarriage," *The Journal of Obstetrics and Gynecology of India*, vol. 69, no. 5, pp. 431–435, 2019.

[358] D. Rizopoulos, J. M. Taylor, J. Van Rosmalen, E. W. Steyerberg, and J. J. Takkenberg, "Personalized screening intervals for biomarkers using joint models for longitudinal and survival data," *Biostatistics*, vol. 17, no. 1, pp. 149–164, 2016.

[359] R. Gueorguieva, R. Rosenheck, and H. Lin, "Joint modelling of longitudinal outcome and interval-censored competing risk dropout in a schizophrenia clinical trial," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 175, no. 2, pp. 417–433, 2012.

[360] S.-C. Weng, Y.-C. Chang, and C.-M. Chen, "Joint analysis of longitudinal and interval-censored failure time data," *Communications in Statistics-Simulation and Computation*, pp. 1–17, 2020.

[361] S. L. Brilleman, M. J. Crowther, M. T. May, M. Gompels, and K. R. Abrams, "Joint longitudinal hurdle and time-to-event models: an application related to viral load and duration of the first treatment regimen in patients with hiv initiating therapy," *Statistics in Medicine*, vol. 35, no. 20, pp. 3583–3594, 2016.