JOINT MODELING OF HIERARCHICAL DATA WITH APPLICATION TO PROSPECTIVE PREGNANCY STUDIES

by

Kirsten J. Lum

A dissertation submitted to Johns Hopkins University
in conformity with the requirements for the degree of
Doctor of Philosophy

Baltimore, Maryland
October, 2014

Abstract

Although fecundity has been studied for decades, the heterogeneity among couples, both biologically and behaviorally, is not well understood. The length of the woman's menstrual cycle has been shown to play an important role; however, a complete assessment of this role requires a model that accounts for both male and female risk factors and the couple's intercourse pattern. We develop and implement a Bayesian joint modeling approach to estimate the woman's underlying distribution of cycle length and assess its relation with couple fecundity while accounting for risk factors of both partners and intercourse frequency and timing relative to ovulation. We apply our approach to prospective pregnancy studies in which couples may enroll when they learn of the study as opposed to waiting for the start of a new menstrual cycle. Due to length-bias, the enrollment cycle will be stochastically larger than the general run of cycles, a typical property of prevalent cohort studies. We develop and evaluate an approach for unbiased estimation of the cycle length distribution for the study population that accounts for length-bias and selection effects, where the probability of enrollment may depend on the time since the last menstrual period. We find that shorter and longer cycle lengths are negatively associated with fecundity even with adjustment for semen quality, age, smoking, and intercourse pattern. This finding motivates investigation of environmental chemicals, in particular perfluoroalkyl surfactants (PFASs), and their potential role in cycle length and fecundity. We extend the joint model to include exposure to PFASs and find that 2-(N-methyl-perfluorooctane sulfonamido) acetate (MeFOSAA) and perfluorooctanoate (PFOA) are associated with shorter cycles while perfluorodecanoate (PFDA) is associated with longer cycles. Further, we find perfluorononanoate (PFNA) and perfluorooctane sulfonamide (PFOSA) are adversely associated with fecundity.

Thesis Advisor: Thomas A. Louis

Committee: Peter Lees, Alvaro Muñoz, and Mei-Cheng Wang

Alternates: Roger D. Peng and Saifuddin Ahmed

Acknowledgments

I would like to thank my advisor Dr. Tom Louis for his mentorship, teaching, and guidance throughout my doctoral training. I would also like to thank my thesis committee for their thought provoking questions and suggestions: Drs. Mei-Cheng Wang, Peter Lees, and Alvaro Muñoz and alternates: Drs. Roger Peng and Saifuddin Ahmed. I am grateful to my teaching mentors, Drs. Karen Bandeen Roche and Marie Diener-West, for showing me how to lead and engage a class in statistics and providing me with many opportunities to develop my communication skills. I also appreciate the assistance of the Biostatistics staff who were always available and happy to help, in particular: Mary Joy Argo, Patty Hubbard, Ashley Gilliam, and Debra Moffitt. To my cohort: Tom Prior, Jeongyong Kim, Finbarr Leacy, Dr. Jenna Krall, Dr. Paige Maas, Dr. Haochang Shou, Dr. Sarah Khasawinah, Dr. Zhenke Wu, and Dr. Yingying Wei, I am thankful for your friendship and support as we made our way through the rigors of obtaining a PhD. And to the Department of Biostatistics faculty, thank you for setting fine examples of scholarship. I will always think of you as family and remember my time spent in this department fondly.

I am grateful to my predoctoral mentors and collaborators at the Eunice Kennedy Shriver National Institute for Child Health and Human Development (NICHD), especially Dr. Raji Sundaram for her insightful comments, ideas, and suggestions for this body of work. I am also forever indebted to her for her constant support, and encouraging me to pursue a PhD. I am thankful to Dr. Germaine Buck Louis for her invaluable advice both in collaborative work and in career training. I would like to thank my branch chief, Dr. Paul Albert, for implementing bag lunch talks and tea time, leading retreats, and encouraging participation in conferences and career training. I would also like to acknowledge the members of the Biostatistics and Bioinformatics Branch at NICHD for their friendship and conversations.

I am also thankful to Dr. Germaine Buck Louis and the LIFE Study team members for

use of the LIFE Study data, an integral and fascinating component of this research. I am grateful as well to the couples who participated in the LIFE Study. I acknowledge that this research was supported by the Intramural Research Program of the U.S., NIH, NICHD [Contracts N01-HD-3-3355, N01-HD-3-3356, N01-HD-3-3358] and the National Institute of Environmental Health Sciences [Training Grant 2T32ES012871]. I also acknowledge that this research utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Maryland (`http://biowulf.nih.gov`).

To my mom, dad, step dad, in-laws, thank you for your loving support and words of encouragement. I am grateful that I could always count on you. I am also thankful to Audrey Lum, Dr. Michelle Folsom Elder, Kate Saporetti, Dr. Rebecca Guglielmo, Alexis Bernstein Wien, Gwen Houldsworth, Krista Brewin, Kate Sapra and Dr. Uba Backonja for your friendship. To my husband, Leon, I am especially thankful for your love, friendship, and neverending support and encouragement for me and my career, without which this degree would not be possible. And to my son, Brendan, for many hours of joyful diversion playing trains, reading books, singing songs, and lots of hugs and kisses.

Table of Contents

List of Abbreviations and Acronyms

AF	Acceleration factor
AIC	Akaike information criterion
BC_a	Bias-corrected and accelerated
BMI	Body mass index
BUGS	Bayesian inference Using Gibbs Sampling
CDF	Cumulative distribution function
DIC	Deviance information criterion
EtFOSAA	2-(N-ethyl-perfluorooctane sulfonamido) acetate
GM	Geometric mean
hCG	Human chorionic gonadotropin
HPD	Highest posterior density
IQR	Interquartile range
LH	Luteinizing hormone
LIFE	Longitudinal Investigation of Fertility and the Environment
LMP	Last menstrual period
LOD	Limit of detection
MCMC	Markov chain Monte Carlo
MeFOSAA	2-(N-methyl-perfluorooctane sulfonamido) acetate
MLE	Maximum likelihood estimate

MSE	Mean squared error
NICHD	Eunice Kennedy National Institute for Child Health and Human Development
NIEHS	National Institute of Environmental Health Sciences
NIH	National Institutes of Health
NPMLE	Nonparametric maximum likelihood estimate
OR	Odds ratio
PC	Principal component
PFAS	Perfluoroalkyl surfactants
PFDA	Perfluorodecanoate
PFNA	Perfluorononanoate
PFOA	Perfluorooctanoate
PFOSA	Perfluorooctane sulfonamide
PFOS	Perfluorooctane sulfonate
SAB	Spontaneous abortion
SD	Standard deviation
SE	Standard error
TTP	Time-to-pregnancy
US	United States

List of Tables

List of Figures

Chapter 1

Introduction

Interest is building in cross-disciplinary approaches to scientific research as evidenced in a speech by Dr. Alan Guttmacher, director of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. Dr. Guttmacher called for the advancement of research in nine key areas including environment, behavior, reproduction, pregnancy and pregnancy outcomes (*Eunice Kennedy Shriver* National Institute of Child Health and Human Development 2011). Each of these themes are rich in challenges that require use and enhancement of methodologies that cut across disciplines. One important example is the investigation of determinants and correlates of human fecundity. Fecundity impairments affect men and women of all countries, races, and socioeconomic levels, and there is evidence to suggest even further health implications later in life (Cooper and Sandler 1997; Whelan et al. 1994). However, relatively few risk factors have been identified. Several studies have shown evidence to suggest that a woman's menstrual cycle characteristics may play an important role in fecundity (Small et al. 2006, 2010; McLain et al. 2012). Scientifically valid and statistically efficient evaluations of the role of menstrual cycle characteristics and fecundity require use of modern statistical approaches and enhancements thereof. A complete investigation requires accounting for the role of biological, behavioral, and environmental risk factors (e.g., age, intercourse timing, and environmental chemicals).

This dissertation addresses the measurement and evaluation of the role of menstrual cycle length in human fecundity. In Chapter 2, we focus on the length of the menstrual cycle in

which the couple enrolls and develop and evaluate an estimation approach that accounts for features of the sampling plan, in particular length-bias and selection effects. This serves as the building block for Chapter 3 in which we incorporate the post-enrollment cycle lengths and jointly model both menstrual cycle length and fecundity accounting for biological and behavioral risk factors. Lastly in Chapter 4, we assess the relation between environmental chemicals, specifically perfluoroalkyl surfactants, and menstrual cycle length and fecundity. We conclude with a discussion of future work and an appendix containing R and OpenBugs code for implementing these methods.

In the remainder of this introduction we provide an overview of menstrual cycle length, fecundity, and perfluoroalkyl surfactants. We also describe key details of the data pertaining to its collection and sampling plan. Finally, we briefly summarize the three main chapters of this dissertation.

1.1. Scientific Background

1.1.1 Fecundity

Fecundity is defined as the biologic capacity of males and females for reproduction irrespective of pregnancy intentions (Buck Louis 2011). Measurement of fecundity is facilitated through one or more study endpoints which include, but are not limited to, onset and progression of puberty, menses and ovulation, hormone profile and libido, and conception and implantation on the female side; and onset and progression of puberty, semen quality, and sexual libido on the male side. This list emphasizes the role of both the male and female partner and the complexity of factors that contribute to a couple becoming pregnant, yet few statistical models have been developed that model more than one of these endpoints. In this dissertation, we focus on two key endpoints: menstrual cycle length and the probability of pregnancy in a menstrual cycle conditional on no pregnancy in previous cycles.

Pregnancy is typically detected by sonography or by home pregnancy tests which measure levels of human chorionic gonadotrophin (hCG) hormone. It should be distinguished

from conception because not only is conception currently unmeasurable (except in assisted reproductive technology), but also, unknown losses can occur between fertilization and implantation. The probability of a recognized pregnancy per menstrual cycle is very low in humans, perhaps emphasizing the many failures that can occur at the stages of ovulation, fertilization, and implantation and the impacts of early pregnancy losses (Norwitz et al. 2001).

Although a large number of studies have been conducted to investigate potential determinants of fecundity (see, e.g., Zaadstra et al. 1993; Jensen et al. 1999a; Kolstad et al. 1999; Small et al. 2006; Ramlau-Hansen et al. 2007; Small et al. 2010; McLain et al. 2012), there is still much work to be done. In 2006, Axmon et al. (2006) found that only 14% of the variation in fecundity as measured by time-to-pregnancy (TTP) was explained by use of oral contraceptives prior to attempting conception, menstrual cycle length, age, and parity. In investigating the role of a potential risk factor or exposure and fecundity, it is important to account for occurrences of intercourse (Tingen et al. 2004). A menstrual cycle is not 'at-risk' for pregnancy if the couple does not have intercourse at least once in the fertile window, a short window of time in a menstrual cycle in which conception is possible. The window's exact length and location are still a topic of research and are thought to potentially vary both within and between couples.

1.1.2 The Menstrual Cycle

The menstrual cycle consists of three phases of changes to the endometrium: bleeding, proliferative, and secretory. The timing and lengths of these phases are controlled by complex hormonal mechanisms. During the proliferative phase, follicular stimulating hormone (FSH) causes the ovum to grow and produce estrogen, which in turn causes endometrial growth. The rise in estrogen at the end of the proliferative phase is thought to trigger a rapid increase in leutinizing hormone (LH). This surge of LH is detectable using fertility monitors and can potentially be used as a biomarker for ovulation. At ovulation, there is a dip in estrogen

levels, which levels out as the corpus luteum (ruptured follicle) takes over production of estrogen and progesterone, resulting in maturation of the endometrium. If pregnancy does not occur, then the corpus luteum decays resulting in a decrease in estrogen and progesterone that leads to bleeding. On the other hand, if pregnancy does occur, increased levels of progesterone preclude menstrual bleeding in preparation for implantation of the blastocyst.

Researchers and clinicians define the beginning of the menstrual cycle with menstruation, designated as the first day of bleeding followed within one day by at least two additional days bleeding. The length of the menstrual cycle is then defined as time (in days) from the first day of menstrual bleeding to the day preceding menstrual bleeding of the next cycle. Prospective longitudinal studies have shown heterogeneity in menstrual cycles both within and across women, with the greatest variability occurring at extremes of age, 2-5 years after menarche and before menopause (Treloar et al. 1967; Vollman 1977; Fehring et al. 2006).

1.1.3 Perfluoroalkyl Surfactants

Perfluoroalkyl surfactants (PFASs) are synthetic chemicals that have been used since the 1950s in the United States and globally in a variety of consumer products (e.g., surfactants and surface protectors in carpets, leather, paper, packaging, fabric and upholstery) (Giesy and Kannan 2002). They are examples of persistent organic pollutants (POPs), which are so named because they resist metabolism and thus persist in the environment. PFASs bio-accumulate in human tissue and have been found in varying levels in human populations from several countries (Kannan et al. 2004; Kato et al. 2011; Olsen et al. 2012; Wang et al. 2011). There is considerable interest in studying the role of PFASs and other POPs in human fecundity. In Chapter 4 we investigate the relations between PFASs, menstrual cycle length, and the probability of pregnancy.

1.2. Epidemiologic Pregnancy Studies

The most popular study designs for collection of fecundity data are retrospective and prospective cohort designs; however, a variety of other designs are possible such as historic prospective, current duration, and case-cohort designs. In retrospective designs, couples are contacted during pregnancy or some time after delivery and asked to recall TTP. Such studies are common because of their ease and low cost, but validity of retrospective TTP has yet to be extensively tested and a number of biases should be carefully considered including behavior modification, exposure time trends, planning, wantedness, pregnancy recognition, medical intervention and the unhealthy-worker effect (Weinberg et al. 1994). Moreover, study samples should be expanded to include couples who did not achieve pregnancy. Alternatively, prospective study designs recruit couples at the start of the pregnancy attempt and prospectively follow them for 6 or 12 menstrual cycles at-risk for pregnancy. To ensure high validity, some prospective designs ask couples to keep daily journals of menstruation, intercourse frequency, contraceptive use, and lifestyle behaviors. Prospective design challenges include intensive data collection demands on the couples, potential selection bias, and extensive recruitment since approximately only 1% of the population is planning pregnancy.

For this research, we use prospective pregnancy data from the Longitudinal Investigation of Fertility and the Environment (LIFE) Study (Buck Louis et al. 2011). The LIFE study sampled couples planning pregnancy who were likely to consume large amounts of fish caught in contaminated waters, and followed 501 couples until delivery, pregnancy loss, or 12 cycles at risk for pregnancy. Couples were accepted if they fulfilled the following: married or in a committed relationship, females aged 18-40 and males over age 18, English or Spanish speaking, self-reported menstrual cycle lengths within 21-42 days, and no hormonal birth control injections in the past 12 months. Couples were excluded if they had been attempting pregnancy for more than 2 months prior to screening. In order to capture preconception concentrations, exposure to PFASs was quantified from non-fasting blood collected on the day of enrollment. An interview was conducted at enrollment to confirm that the couple

was not already pregnant and to obtain baseline data including demographics, reproductive history, and date of last menstrual period (LMP), regularity of menstrual cycles. For the length of the pregnancy attempt, both the male and female kept daily journals of intercourse frequency, contraceptive use, servings of fish or shellfish eaten, multivitamin use, alcohol and caffeinated drinks consumed, cigarettes smoked, and menstruation (female diary only). Male journals also included number of ejaculations and excessive heat exposure at work. In addition, women used the Clearblue®Easy fertility monitor to measure daily levels of oestrone-3-glucuronide, a metabolite of oestradiol, and luteinizing hormone (LH). Pregnancy was recognized using Clearblue®Easy pregnancy tests that detect levels of hCG over 25 mIU/mL and are administered at home on the day menstruation is expected.

The sampling plan for the LIFE Study is a prevalent sampling plan wherein enrollment occurs after some initiating event but prior to a terminating event. Couples were enrolled in the study at any point in the woman's menstrual cycle, as opposed to the first day of menstrual bleeding (the initiating event). Assuming a couple's awareness of the study is equally likely on any day and their propensity to enroll does not depend on the day, then a couple is more likely to enroll during a menstrual cycle that is generally longer than the woman's typical cycle. The observed lengths of the cycle in which the couple enrolls, which we will refer to as the enrollment cycle, are said to be length-biased. This problem is well known in the area of disease screening where those living with detectable disease for longer periods of time are more likely to be screened compared to those with fast moving disease. Furthermore, there may be selection bias if the couple's decision to enroll is based upon the length of time between the woman's LMP and the date of enrollment (a backward recurrence time). Sampling weights may be used to account for the fact that the enrollment cycle lengths are not random observations from the same inherent woman-specific distribution as the follow-up cycle lengths; however, the weights are a function of the unknown probability of enrollment.

1.3. Overview of Chapters

In Chapter 2, we propose an approach for estimating the distribution of menstrual cycle length that accounts for features of the sampling plan, in particular for length bias and selection bias on the backward recurrence time. For this chapter only, we restrict our attention to the enrollment cycle (i.e., the cycle in which the couple enrolls), as the distribution estimated from these observations can be influenced by the way in which these lengths are ascertained. We propose a two-stage approach wherein we first estimate the probability of enrollment as a function of the backward recurrence time and then use it in the full likelihood with sampling weights to account for length-bias and selection effects. To broaden the applicability of this approach, we incorporate a time-independent covariate and woman-specific random effect using a mixed effects accelerated failure time model. We conduct simulation studies to evaluate the performance of the algorithm for two scenarios of the probability of enrollment. Lastly, we demonstrate our approach using enrollment cycle data from the LIFE Study to estimate the woman's menstrual cycle length distribution and the association between age group and cycle length. Our approach yields an estimate of the probability of enrollment as a function of time since LMP which may inform the design of future studies.

In Chapter 3, we develop and implement a Bayesian hierarchical joint model to assess the association between the woman-specific distribution of menstrual cycle length and couple fecundity, while accounting for the timing and frequency of intercourse and risk factors of both the male and female partner. For menstrual cycle length, we build on the methods developed in Chapter 2 and model all cycles of follow-up while addressing issues such as skewness, intra- and inter-woman variability, and right-censoring of the cycle in which the woman becomes pregnant. For fecundity, we develop a hierarchical model for the conditional probability of pregnancy in a menstrual cycle given no pregnancy in previous cycles of trying and the couples' history of intercourse acts. We incorporate a flexible spline function of intercourse timing and consider a broad window of days at risk for pregnancy around the estimated day of ovulation. Based on an analysis of the LIFE Study data, we find evidence

that a woman's age is inversely associated with menstrual cycle length (for ages 19 to 40) and that there is a quadratic relation between cycle length and fecundity wherein shorter and longer cycle lengths are negatively associated with couple fecundity even with adjustment for male semen quality, age, smoking status and intercourse pattern.

In Chapter 4, we investigate the role of female exposure to PFASs with (i) menstrual cycle length and (ii) fecundity in the context of menstrual cycle length. We incorporate female concentration levels (in ng/mL) of seven PFASs in the menstrual cycle length and pregnancy models of the joint model developed in Chapter 3. We fit this model using data from the LIFE Study with preconception quantification of PFASs. As some of the PFASs have values below the limit of detection (LOD), we consider both the machine observed values (continuous) and also categorical forms of the exposures in which the values below LOD are in the category of lowest concentration. We find evidence that both 2-(N-methyl-perfluorooctane sulfonamido) acetate and perfluorooctanoate are associated with shorter menstrual cycle lengths while perfluorodecanoate is associated with longer length. In the context of menstrual cycle length, we find evidence that perfluorononanoate (PFNA) and perfluorooctane sulfonamide are negatively associated with the probability of pregnancy in a two stage model. In the joint model which accounts for uncertainty in estimating menstrual cycle length, PFNA is again directly associated with the probability of pregnancy.

Chapter 2

Accounting for Length-bias and Selection Effects in Estimating the Distribution of Menstrual Cycle Length

2.1. Introduction

We consider the setting of longitudinal prospective studies of cyclic events forming a renewal process, where the primary objective is to estimate the distribution of interarrival times. A necessary consideration when making inferences about the study population based on the observed times is the sampling plan; in particular, one in which enrollment can occur at any time in a gap interval including its start or end, and the enrollment date may differ among participants. After enrollment, generally participants are followed with cycle renewals recorded until censoring at the end of the study, after a fixed number of cycles, loss to follow-up, or occurrence of a competing event. Also, the gap time can be truncated.

For example, the process underlying the periodic shedding of the lining of the uterus known as menstruation is a stationary renewal process in that the occurrence of bleeding events is cyclic over time. The interarrival time of menstrual bleeding events, or menstrual cycle length, is an important determinant of risk for several chronic diseases afflicting females including heart disease and breast cancer, and it is also an important component in assessing female fecundity (see, e.g., Jensen et al. 1999b). It is typically defined as the number of days

from the first day of menstrual bleeding (the day marking the last menstrual period, LMP) to the day preceding the next menstrual cycle, with care taken to distinguish menstrual bleeding from episodic bleeding. Cycle length is commonly measured in time-to-pregnancy studies either via retrospective recall or more reliably in prospective designs using daily diaries of intensity of bleeding along with hormonal monitoring. In this paper, we will analyze cycle lengths from a comprehensive prospective pregnancy study, the Longitudinal Investigation of Fertility and the Environment (LIFE) Study (Buck Louis et al. 2011) in which repeated measurements of menstrual cycle length and couple's preconception exposure to persistent environmental chemicals were collected. Research goals include estimating the distribution of menstrual cycle length accounting for covariates and residual variation, and using the estimated distribution to predict time-to-pregnancy. An initial step is to estimate the distribution of menstrual cycle length, both in a population of women and for an individual woman.

A complicating feature of the LIFE and other such studies of renewal processes is that in order to efficiently sample couples attempting to become pregnant, enrollments occur between renewals as opposed to waiting for the next bleeding event. Furthermore, the interarrival time is potentially right censored by pregnancy or exit from the study resulting in four different combinations of truncation and censoring relevant to the observed enrollment menstrual cycle lengths (more generally, the gap times) as described by Gill and Keiding (2010). In modeling the distribution of (potentially censored) enrollment lengths, we account for two features of this sampling process: length-bias and selection effects. Length-bias relates to the process by which an announcement of a study has a probability of "intersecting" a menstrual cycle length that is proportional to its length. Due to length-biased sampling, the distribution of enrollment lengths is biased in the positive direction compared with the population distribution. Selection relates to the possible dependence of the probability that the couple will decide in favor of enrollment on the time from LMP until they learn of the study (the backward recurrence time), resulting in effects that could be positive or negative;

thus, adding to or canceling out the length-bias.

A common and possibly unbiased approach for estimating the menstrual time distribution is to circumvent these challenges by excluding the enrollment cycle lengths and analyzing only post-enrollment cycle lengths (see, e.g., Murphy, Bentley, and O'Hanesian 1995), conceding some loss of information. However, in some applications the additional information provided by the enrollment data can be important in detecting subtle covariate effects and the selection effects may in themselves be of interest. Moreover, in prospective pregnancy studies, the most fecund women may have very few cycles under observation, if any, post-enrollment.

Several authors have developed methods for estimating the unbiased distribution of survival times from length-biased or left-truncated data (see, e.g., Asgharian, M'Lan, and Wolfson 2002; Cox 1969; Huang and Qin 2011; Vardi 1989; Wang 1989, 1991; Wang, Jewell, and Tsai 1986; Winter and Foldes 1988), which would allow for inclusion of the enrollment menstrual cycle lengths. Building upon this body of work, we propose an approach for estimating both the probability of enrollment as a function of time from LMP and the distribution of menstrual cycle length for the study population, adjusting for length-bias and selection effects. In Sections 2.2-2.5, we formalize the issue and develop our estimation approach with a focus on enrollment data. In Sections 2.6 and 2.7, we present results from a simulation study and analysis of the distribution of menstrual cycle length for the LIFE study population, respectively. We conclude with a discussion of remaining challenges and impact on analysis and design of longitudinal studies of recurrent events.

2.2. Estimating the Distribution of Menstrual Cycle Length

Our aim is to estimate the distribution of menstrual cycle length for a population focusing on the role of the enrollment cycle data, as the distribution estimated from observations of this cycle can be influenced by the way in which these cycles are ascertained. We assume that the enrollment cycle lengths for different individuals are independent and identically distributed. The sampling (enrollment process) induces a distribution for the initial observed

cycle length that could potentially be different from that for the population. Specifically, we account for length-bias and selection effects on the enrollment cycle length distribution so that inferences relate to a reference population. We first consider the uncensored case and then generalize.

2.2.1 Notation, Sampling Plan and Consequences

Let $\tilde{Y} \sim f(\cdot), 0 < \tilde{Y} < \infty$ with associated survivor function $S(y) = \int_y^\infty f(u)du$ and cumulative distribution function (CDF) $F(y) = \int_0^y f(u)du$. Without loss of generality, assume $\tilde{Y} < V < \infty$, where V is a finite upper bound. Assume that the sequence $(\tilde{Y}_1, \tilde{Y}_2, \ldots)$, indexed by m, starts at some origin (T_0) and continues over time with $T_m = T_0 + \tilde{Y}_1 + \tilde{Y}_2 + \cdots + \tilde{Y}_m$. Consider the following sampling plan for a longitudinal study:

- An announcement of study availability is issued at calendar time t^* and it "intersects" cycle length $\tilde{Y}_{m(t^*)}$ with $m(t^*) = \{m : T_{m-1} < t^* \leq T_m\}$ at $\tilde{B}$, the backwards recurrence time. That is, $\tilde{B}$ is the time from T_{m-1} to t^* and can take on values in the range $0 \leq \tilde{B} < V$. In addition let,
  - $\lambda(\tilde{Y}) = \text{pr}(\text{the announcement "intersects" a cycle of length } \tilde{Y})$.
  - $\tilde{A}$ be the residual time from enrollment to the next event, so $\tilde{Y}_{m(t^*)} = \tilde{B} + \tilde{A}$. We consider the case where both $\tilde{B}$ and $\tilde{A}$ are known. As examples, Keiding et al. (2002) and Keiding et al. (2012) studied estimation of the distribution of time-to-pregnancy in the current-duration design when only $\tilde{B}$ was known; and Brookmeyer and Gail (1987) studied bias in estimation of the distribution of time to disease when $\tilde{B}$ was unknown, but $\tilde{A}$ was known.
- The potential participant (the couple) decides on enrolling in the study. Let $R = 1$ if the participant enrolls and $R = 0$ otherwise. If the participant enrolls, the enrollment cycle length is $\tilde{Y}_{m(t^*)}$.
  - Assume that t^* is stochastically independent of the T_m sequence. As the underlying process $\tilde{Y}$ is a renewal process, using the theorem proved by Winter (1989) and

details included in Section 2.A.1.1, $[\tilde{B} \mid \tilde{Y}_{m(t^*)}] \sim \mathcal{U}(0, \tilde{Y}_{m(t^*)}), 0 \leq \tilde{B} < \tilde{Y} < V$.

- A potential participant knows only the $\tilde{B}$ value and not $\tilde{Y}_{m(t^*)}$, and so define the probability of enrollment, $\pi(\tilde{B}) = pr(R = 1 \mid \tilde{Y}_{m(t^*)}, \tilde{B}) = pr(R = 1 \mid \tilde{B})$.

A $\lambda(\tilde{Y}) \propto \tilde{Y}$ produces standard length-bias analogous to the waiting time paradox (see, Feller 1971) or, for example, in Aalen and Husebye (1991) pertaining to the cyclic movements in the small bowel during the fasting state. Just as longer bus waiting times are more likely to be intersected by an individual arriving at a bus stop, longer cycle lengths are more likely to be intersected by the announcement of the study than are shorter. This length-biased sampling results in a distribution of enrollment cycle lengths that is stochastically larger than the population distribution. Further, a non-constant $\pi(\cdot)$ induces additional selection effects and, depending on the shape of $\pi(\cdot)$, can induce a stochastically larger or smaller distribution for an enrollment cycle.

In modeling the enrollment cycle length, one must account for the aforementioned length-bias and selection effects on the first observation of the underlying renewal process. Our focus is on estimating the distribution of $\tilde{Y}$ using a likelihood that adjusts for the enrollment sampling process and henceforth we consider only enrollment lengths.

2.2.2 Modeling

If $\lambda(\tilde{Y}) \equiv \bar{\lambda}$ and $\pi(\tilde{B}) \equiv \bar{\pi}$, then neither length-bias nor selection effects operate and standard modeling is valid. If $\lambda(\tilde{Y}) \propto \tilde{Y}$ and $\pi(\tilde{B}) \equiv \bar{\pi}$, then $\tilde{Y}_{m(t^*)}$ is distributed according to the length-biased distribution: $f(y)y\mu^{-1}$, where $\mu = E(\tilde{Y})$ (see Cox 1962). If, in addition, $\pi(\tilde{B})$ depends on $\tilde{B}$, then $\tilde{Y}_{m(t^*)}$ is distributed as a more generally weighted version of $f(\cdot)$. Consequently, statistical models must account for the enrollment sampling plan by conditioning on $R = 1$. To simplify notation, henceforth we drop the " ˜ " notation and subscript $m(t^*)$, let $[Y, B] \stackrel{\mathcal{D}}{=} [\tilde{Y}_{m(t^*)}, \tilde{B} \mid R = 1]$ and denote this enrollment distribution by g. We have

the joint distribution

$$g_{Y,B}(y,b) \;=\; f_{\tilde{Y},\tilde{B}|R}(y,b \mid R=1) = \frac{f(y)\lambda(y)\pi(b)\frac{1}{y}}{\int_0^\infty f(u)\lambda(u)\frac{1}{u}\Big\{\int_0^u \pi(t)dt\Big\}du},\ 0 \le b < y < V < \infty.$$

The marginal distribution of Y is

$$g_Y(y) = \frac{f(y)\lambda(y)\frac{1}{y}\int_0^y \pi(t)dt}{\int_0^\infty f(u)\lambda(u)\frac{1}{u}\Big\{\int_0^u \pi(t)dt\Big\}du},\ 0 < y < V < \infty; \tag{2.1}$$

a weighted distribution with weight $\omega(\lambda, y, \pi) \propto \lambda(y)\frac{1}{y}\int_0^y \pi(t)dt$ and CDF $G_Y(y) = \int_0^y g_Y(u)du$. The marginal distribution for B is

$$g_B(b) = \pi(b)\frac{\int_b^\infty f(u)\lambda(u)\frac{1}{u}du}{\int_0^\infty f(u)\lambda(u)\frac{1}{u}\Big\{\int_0^u \pi(t)dt\Big\}du},\ 0 \le b < V < \infty; \tag{2.2}$$

which is proportional to $\pi(b)$, and if $\lambda(Y) \propto Y$, to $S(b)$.

2.3. Parameter Estimation

Assume that $f(\cdot) = f(\cdot \mid \theta)$, and that recruitment for the longitudinal study continues until n participants are enrolled. At enrollment, participants report the LMP of the current cycle, so observed data for participant i are (b_i, y_i). The full likelihood is the product of the conditional likelihood ($[y \mid b]$) and the marginal likelihood for B. Specifically, with $f(\cdot) = f(\cdot \mid \theta)$, assuming that the functions $\lambda(\cdot)$ and $\pi(\cdot)$ are known,

$$\mathcal{L}(\theta; \mathbf{y} \mid \mathbf{b}, \lambda) \;=\; \mathcal{L}_c(\theta; \mathbf{y} \mid \mathbf{b}, \lambda) \times \mathcal{L}_m(\theta; \mathbf{b}, \lambda) \tag{2.3}$$

$$\mathcal{L}_c(\theta; \mathbf{y} \mid \mathbf{b}, \lambda) \;=\; \prod_{i=1}^n g_{Y|B}(y_i \mid b_i) = \prod_{i=1}^n \frac{f(y_i)\lambda(y_i)\frac{1}{y_i}}{\int_{b_i}^\infty f(u)\lambda(u)\frac{1}{u}du}, \tag{2.4}$$

and from (2.2), the marginal likelihood of B is

$$\mathcal{L}_m(\theta \mid \mathbf{b}, \pi, \lambda) = \prod_{i=1}^n \frac{\pi(b_i)\int_{b_i}^\infty f(u)\lambda(u)\frac{1}{u}du}{\int_0^\infty f(u)\lambda(u)\frac{1}{u}\Big\{\int_0^u \pi(t)dt\Big\}du}. \tag{2.5}$$

Note that $\mathcal{L}_m$ depends only on the shape of $\pi(\cdot)$ and not on its magnitude (is invariant to multiplication of $\pi(\cdot)$ by a positive constant).

We focus on estimating the survivor (tail) function $S(y) = \int_y^\infty f(u)du$ or in the parametric case $S(y \mid \theta) = \int_y^\infty f(u \mid \theta)du$. Under the renewal process sampling plan in Section 2.2.1, we henceforth assume that $\lambda(y) \propto y$. If $\pi(\cdot)$ can be estimated, it is straightforward to use the full likelihood. The conditional likelihood ($\mathcal{L}_c$) does not depend on $\pi(\cdot)$ and so if this selection function is not of interest, inferences can be based on (2.4). Including the marginal likelihood ($\mathcal{L}_m$) may provide a notable increment in information on θ, but it does depend on $\pi(\cdot)$. As (2.2) shows, $\pi(\cdot)$ and f interact, and a recursive approach is effective.

2.3.1 Recursive Estimation

The following recursive algorithm requires g_B (equivalently the CDF G_B). Using the observed backward recurrence times, g_B can be estimated by the empirical probability mass function (equivalently G_B by the empirical distribution function). Alternatively, a smoothed estimate of g_B, can be obtained using adaptive bandwidth kernel density estimation (see, e.g. Wand and Jones 1994). A two-stage algorithm for estimating $\pi(\cdot)$ and $S(\cdot)$ proceeds as follows:

1. Conditioning on B, let $\hat{S}^{(0)}(\cdot)$ be a non-parametric estimate of $S(\cdot)$ based on the conditional likelihood (see, e.g. Winter and Foldes 1988). For parametric estimation, equivalently let $\hat{\theta}^{(0)}$ be the estimate of θ based on maximizing $\mathcal{L}_c(\theta; \mathbf{y} \mid \mathbf{b}, \lambda)$ in (2.4).
2. Using the relation between $\pi(\cdot)$ and g_B in (2.2), obtain $\hat{\pi}^{(1)}(\cdot)$ (only the shape is needed), with $S(\cdot) = \hat{S}^{(0)}(\cdot)$.
3. Let $\hat{S}^{(\nu)}(\cdot)$ be the NPMLE using the full likelihood, with $\pi(\cdot) = \hat{\pi}^{(\nu)}(\cdot)$. Equivalently, for parametric estimation let $\hat{\theta}^{(\nu)}$ be the MLE using $\mathcal{L}(\theta; \mathbf{y} \mid \mathbf{b}, \lambda)$ in (2.3).
4. Using relation (2.2), update $\hat{\pi}^{(\nu+1)}(\cdot)$ (up to a constant of proportionality) with $S(\cdot) = \hat{S}^{(\nu)}(\cdot)$.
5. Iterate between steps 3 and 4 until convergence criterion is met.

In Section 2.A.1.2 we give a proof of the existence and uniqueness of the solution for π.

2.3.2 Nonparametric Estimation of $S(y)$

Assuming $\lambda(y) \propto y$ and $\pi(b) \equiv \bar{\pi}$, the inverse probability weighted empirical estimate of $S(y)$ derived by Cox (1969) for uncensored lengths $(y_i; i = 1, \ldots, n)$ sampled from G_Y is

$$\hat{S}(y) = \sum_{i=1}^{n} w_i \mathbb{1}\{y_i \geq y\};\ y > 0,\ \text{where } w_i = \frac{1/y_i}{\sum_{j=1}^{n} 1/y_j}. \tag{2.6}$$

Under the sampling plan in Section 2.2.1, we have a single sample of uncensored enrollment cycle lengths and backward recurrence times $\{(y_i, b_i); i = 1, \ldots, n\}$ and we allow for a non-constant $\pi(b)$. In this case, Y has the marginal density, $g_Y(\cdot)$, given in (2.1) and it is straightforward to derive the relation:

$$E\Big[\Big\{\lambda(Y_i)\frac{1}{Y_i}\int_0^{Y_i}\pi(t)dt\Big\}^{-1}\mathbb{1}\{Y_i \geq y\}\Big] = \frac{S(y)}{\int_0^{\infty} f(u)\lambda(u)\frac{1}{u}\Big\{\int_0^{u}\pi(t)dt\Big\}du}.$$

Since this expectation is proportional to $S(y)$, we estimate $S(y)$ as in (2.6), but with modified weights:

$$w_i = \frac{\Big\{\lambda(y_i)\frac{1}{y_i}\int_0^{y_i}\pi(t)dt\Big\}^{-1}}{\sum_{j=1}^{n}\Big\{\lambda(y_j)\frac{1}{y_j}\int_0^{y_j}\pi(t)dt\Big\}^{-1}}. \tag{2.7}$$

The weights in (2.7) depend on $\pi(\cdot)$ up to a multiplicative constant; therefore, the recursive algorithm presented in Section 2.3.1 can be amended to include an additional step prior to step 3 updating $w_i^{(\nu)}$ using (2.7) with $\pi(\cdot) = \hat{\pi}^{(\nu)}(\cdot)$.

For right-censored data, assuming $\pi(\cdot) \equiv \bar{\pi}$, several non-parametric approaches for estimating $S(\cdot)$ accounting for length-bias or left-truncation have been developed (Asgharian, M'Lan, and Wolfson 2002; Huang and Qin 2011; Vardi 1989). In order to relax the assumption on $\pi(\cdot)$, we note that conditioning on B the likelihood does not depend on $\pi(\cdot)$. Methods for estimating $S(\cdot)$ based on $\mathcal{L}_c$ have been developed by many including Keiding and Gill (1990); Wang (1991), Wang, Jewell, and Tsai (1986), and Winter and Foldes (1988). When estimation of $\pi(\cdot)$ is not of interest, these methods allow for non-parametric estimation of $S(\cdot)$, but may lead to a loss of efficiency when compared with using the full likelihood. In the simulation study (see Section 2.6), we do not find evidence of an efficiency loss.

2.3.3 Semi-parametric Estimation of $S(y \mid \theta)$

In the semi-parametric approach, we assume $\tilde{Y} \sim f(\cdot;\theta)$; however, we do not make any parametric assumptions about $\pi(\cdot)$. The recursive algorithm given in Section 2.3.1 may be used to estimate $\pi(\cdot)$ and $S(y \mid \theta) = \int_y^\infty f(u \mid \theta)du$ with the adaptation in step 3 of producing $\hat{\theta}^{(\nu)}$ from the full likelihood written in terms of the current estimate of $\pi(\cdot) = \hat{\pi}^{(\nu)}(\cdot)$. As we will show in Section 2.4.1, the semi-parametric approach can accommodate censoring. Bootstrap bias-corrected and accelerated (BC_a) confidence intervals (Efron 1987) can be constructed for θ that account for the uncertainty in estimating $\pi(\cdot)$.

2.4. Censoring

To accommodate censoring, let $\tilde{C}$ denote the censoring time defined from the most recent event (LMP) to the censoring date. Under the sampling plan detailed in Section 2.2.1, length-bias operates on the enrollment cycle length and also on the enrollment cycle censoring time, when defined from LMP. Consequently, length-bias operates on the available length: $\tilde{Y} \wedge \tilde{C}$. For the enrollment cycle, we eliminate the " ˜ " notation and write $C = B + V$, where V is the residual time from enrollment date to censoring date, and write $X = Y \wedge C$ as the length from LMP to the next menstrual period or censoring, with event indicator $\delta = I(Y \wedge C)$. As Asgharian, M'Lan, and Wolfson (2002), Huang and Qin (2011), and Wang (1991) among others noted, B is the initial segment of both Y and C, so in general Y and C are dependent. We assume the standard relations for left truncated, right-censored random variables, that $[A \perp V \mid B]$ (conditional independence) and the conditional distribution of V given B is non-informative regarding f, and so $f_{V|B}$ does not depend on θ.

As before recruitment continues until n couples are enrolled, with enrollment data (b_i, x_i, δ_i);

$i = 1, \ldots, n$. Under the conditional independence and non-informative assumptions, the conditional likelihood of X given B is (leaving out θ on the right-hand side)

$$\mathcal{L}_c(\theta \mid \mathbf{b}, \lambda, \mathbf{x}, \delta) \propto \prod_{i=1}^{n} \frac{\left\{ f(x_i)\lambda(x_i)\frac{1}{x_i} \right\}^{\delta_i} \left\{ \int_{x_i}^{\infty} f(u)\lambda(u)\frac{1}{u}du \right\}^{(1-\delta_i)}}{\int_{b_i}^{\infty} f(u)\lambda(u)\frac{1}{u}du}. \tag{2.8}$$

The marginal likelihood retains the form from (2.5), and so the full likelihood is

$$\mathcal{L}(\theta \mid \pi, \lambda, \mathbf{b}, \mathbf{x}, \boldsymbol{\delta}) \propto \prod_{i=1}^{n} \frac{\left\{ f(x_i) \right\}^{\delta_i} \left\{ \int_{x_i}^{\infty} f(u)\lambda(u)\frac{1}{u}du \right\}^{(1-\delta_i)}}{\int_{0}^{\infty} f(u)\lambda(u)\frac{1}{u}\left\{ \int_{0}^{u} \pi(t)dt \right\}du}. \tag{2.9}$$

2.4.1 Semi-parametric Estimation of $S(y \mid \theta)$ in the Censored Case

We find the MLE $\hat{\theta}$ and use $S(y \mid \hat{\theta}) = \int_y^{\infty} f(u; \hat{\theta})du, y > 0$. For known $\pi(\cdot)$, maximum likelihood methods can be used to estimate θ, $S(\cdot \mid \theta)$ and other functions of it, and the likelihood ratio can be used to calculate confidence regions. However, $\pi(\cdot)$ is usually not known and a recursive algorithm similar to that in Section 2.3.1 is necessary. Starting with an estimate of $g_B(b)$, proceed as follows: 1) let $\hat{\theta}^{(0)}$ be the MLE using the conditional likelihood in (2.8) so, $\hat{S}^{(0)} = S(y \mid \hat{\theta}^{(0)}) = \int_y^{\infty} f(u \mid \hat{\theta}^{(0)})du$; 2) estimate the shape of $\pi(b)$ by $\hat{\pi}^{(1)}(b) \propto \frac{\hat{g}_B(b)}{S(b|\hat{\theta}^{(0)})}$; 3) let $\hat{\theta}^{(\nu)}$ be the MLE using the full likelihood in (2.9) with $\pi(\cdot) = \hat{\pi}^{(\nu)}(\cdot)$; 4) update $\hat{\pi}^{(\nu+1)}(b) \propto \frac{\hat{g}_B(b)}{S(b|\hat{\theta}^{(\nu)})}$; and 5) iterate between steps (3) and (4) until convergence criterion is met. Use bootstrap BC_a confidence intervals (Efron 1987) to account for uncertainty in estimating $\pi(\cdot)$.

2.5. Incorporating a Random Effect and a Covariate

In many applications it is desirable to account for observed and unobserved residual heterogeneity; therefore, we propose the following model for the observed menstrual cycle length for couple i:

$$\tilde{Y}_i = \tilde{Y}_{0i} W_i e^{\beta z_i}, \tag{2.10}$$

where W_i is an unobserved couple-specific random effect with support $w_i \in (0, \infty)$ and CDF $H(w_i; \alpha), \alpha \in \mathcal{A}$; z_i is an observed time-independent covariate; β is a fixed effect

parameter; and $\tilde{Y}_{0i}$ is the baseline cycle length for a couple with $W_i = 1$ and $z_i = 0$ with CDF $F_0(y_{0i};\theta), \theta \in \Theta$. We assume the distribution of Z does not depend on θ and that conditional on $Z_i = z_i$ and $W_i = w_i$, the cycle lengths within a woman are independent and identically distributed with CDF $F(y_i \mid w_i, z_i; \theta, \beta) = F_0(y_i w_i^{-1} e^{-\beta z_i}; \theta, \beta)$ and pdf $f(y_i \mid w_i, z_i; \theta, \beta) = w_i^{-1} e^{-\beta z_i} f_0(y_i w_i^{-1} e^{-\beta z_i}; \theta, \beta)$.

Our goal is to use the enrollment cycle data to estimate the marginal distribution of menstrual cycle length for the reference population with covariate $Z = z$. For the enrollment cycle, we eliminate the " ˜ " notation and assume as in Section 2.4 that length-bias operates on $X_i = Y_i \wedge C_i$ with Y_i modeled as in (2.10). Since the random effects are not observed, we assume $\pi(\cdot)$ does not depend on W_i, and further we restrict to the case where $\pi(\cdot)$ does not explicitly depend on z_i. Incorporating the covariate, the observed enrollment data consist of $(b_i, x_i, \delta_i, z_i); i = 1, \ldots, n$. We assume non-informative and independent censoring conditional on $Z_i = z_i$ and $W_i = w_i$. Under these assumptions, the observed data conditional likelihood of X given (B, Z) is

$$\mathcal{L}_c(\theta, \alpha, \beta \mid \mathbf{b}, \mathbf{z}, \lambda, \mathbf{x}, \delta) \propto \prod_{i=1}^{n} \frac{\left\{ f(x_i \mid z_i; \theta, \alpha, \beta)\lambda(x_i)\frac{1}{x_i} \right\}^{\delta_i} \left\{ \int_{x_i}^{\infty} f(u \mid z_i; \theta, \alpha, \beta)\lambda(u)\frac{1}{u} du \right\}^{(1-\delta_i)}}{\int_{b_i}^{\infty} f(u \mid z_i; \theta, \alpha, \beta)\lambda(u)\frac{1}{u} du}. \tag{2.11}$$

Here, $f(\cdot \mid z_i; \theta, \alpha, \beta) = \int_W f(\cdot \mid w, z_i; \theta, \beta) dH(w; \alpha)$, where we use Gaussian quadrature integration to integrate over the unobserved random effect; however, one could alternatively derive the complete data likelihood and use an Expectation Maximization approach. Extending the full likelihood in (2.9) to incorporate the covariate and random effect, we have

$$\mathcal{L}(\theta, \alpha, \beta \mid \pi, \lambda, \mathbf{b}, \mathbf{x}, \mathbf{z}, \boldsymbol{\delta}) \propto \prod_{i=1}^{n} \frac{\left\{ f(x_i \mid z_i; \theta, \alpha, \beta) \right\}^{\delta_i} \left\{ \int_{x_i}^{\infty} f(u \mid z_i; \theta, \alpha, \beta)\lambda(u)\frac{1}{u} du \right\}^{(1-\delta_i)}}{\int_0^{\infty} f(u \mid z_i; \theta, \alpha, \beta)\lambda(u)\frac{1}{u} \left\{ \int_0^u \pi(t \mid z_i) dt \right\} du}. \tag{2.12}$$

As before, note that $\mathcal{L}_c$ in (2.11) does not depend on $\pi(\cdot)$. Therefore, one could estimate (θ, α, β) using a maximum likelihood approach based on it, and we compare the estimates and standard errors obtained using $\mathcal{L}_c$ to those based on $\mathcal{L}$ in Sections 2.6 and 2.7.

2.5.1 Semi-parametric Estimation of $S(y \mid z; \theta, \alpha, \beta)$

To extend the semi-parametric approach to incorporate a covariate and random effect, we assume a parametric distribution for $\tilde{Y}_0$; but as before, we do not make any parametric assumptions about the form of $\pi(\cdot)$. Starting with an estimate of $g_B(b \mid z)$ (as shown in Section 2.6, the empirical probability mass function works well), proceed as follows:

1. Let $(\hat{\theta}^{(0)}, \hat{\alpha}^{(0)}, \hat{\beta}^{(0)})$ be the MLE using the conditional likelihood in (2.11) so,

$$S^{(0)}(y \mid z; \hat{\theta}^{(0)}, \hat{\alpha}^{(0)}, \hat{\beta}^{(0)}) = \int_y^\infty f(u \mid z; \hat{\theta}^{(0)}, \hat{\alpha}^{(0)}, \hat{\beta}^{(0)}) du.$$

2. $\hat{\pi}^{(1)}(b \mid z) \propto \dfrac{\hat{g}_{B|Z}(b \mid z)}{S^{(0)}(y \mid z; \hat{\theta}^{(0)}, \hat{\alpha}^{(0)}, \hat{\beta}^{(0)})}$.

3. Let $(\hat{\theta}^{(\nu)}, \hat{\alpha}^{(\nu)}, \hat{\beta}^{(\nu)})$ be the MLE using the full likelihood in (2.12) with $\pi(\cdot \mid z) = \hat{\pi}^{(\nu)}(\cdot \mid z)$.

4. $\hat{\pi}^{(\nu+1)}(b \mid z) \propto \dfrac{\hat{g}_{B|Z}(b \mid z)}{S^{(\nu)}(y \mid z; \hat{\theta}^{(\nu)}, \hat{\alpha}^{(\nu)}, \hat{\beta}^{(\nu)})}$.

5. Iterate between steps (3) and (4) until convergence criterion is met.

In Sections 2.6 and 2.7, we consider the case where Z is a binary covariate, which is useful when comparing the distribution of menstrual cycle length for two populations.

2.6. Simulation Studies

We describe and report a simulation study examining the semi-parametric estimation approach described in Section 2.5 in the setting in which we model the observed menstrual cycle length dependent on an unobserved random effect W and an observed time-independent binary covariate z (see model 2.10). Our primary interest is in the estimation of the marginal means and standard deviations: $E(\tilde{Y} \mid z = 0)$, $E(\tilde{Y} \mid z = 1)$, $SD(\tilde{Y} \mid z = 0)$, and $SD(\tilde{Y} \mid z = 1)$, which are functions of the parameters (θ, α, β). We evaluate performance of maximum likelihood estimators when proper account is taken of length-bias and selection effects. We focus on censored data using the full likelihood in (2.12).

Using the data generation scheme described in Appendix A.1, we generate the simulated data consisting of 2000 samples, each with 500 participants enrolled using quota sampling, a sampling scheme in which participants are enrolled until a quota is reached. The first 250 participants enrolled of each sample compose a small sample size case. In the data generation scheme, the choice of parametric family and the values of the parameters mimic the LIFE data analysis in Section 2.7.

We study the performance of the proposed recursive semi-parametric algorithm for empirical estimation of $\pi(b \mid z)$ and parametric estimation of the distribution of $\tilde{Y}$, under model (2.10) in which $\tilde{Y}_0$ is assumed to be a normal-Gumbel mixture and $W \sim Gamma(\alpha, 1/\alpha)$ so $E(W) = 1$ and $Var(W) = 1/\alpha$. In particular, we are interested in two data generating models for π, constant as found in the data analysis, and decreasing with b. The bias, standard errors, mean squared error and coverage probabilities for the estimators of $E(\tilde{Y} \mid z = 0)$, $E(\tilde{Y} \mid z = 1)$, $SD(\tilde{Y} \mid z = 0)$, $SD(\tilde{Y} \mid z = 1)$, β, and $SD(W)$ are summarized in Table 2.1. For both cases of π, the bias of each of the marginal parameters is approximately 0. To account for the uncertainty in estimating both π and the model parameters, we calculated bootstrap BC_a intervals with 1000 bootstrap replicates. The coverage probability of the nuisance parameter, $SD(W)$, is less than the nominal 95%; however, the coverage probability for each of the marginal parameters of interest is very good even for small samples. Using the same simulated data, we also studied the performance of maximum likelihood estimators based on the conditional likelihood, $\mathcal{L}_c$ in (2.11) which does not require estimation of π (see Table 2.2). Under both generating models for π and for each sample size, we did not observe a notable difference between the two approaches in any of the performance measures. Computer code in R for this simulation is given in Appendix A.2.

Table 2.1: Performance of semi-parametric algorithm for estimating $S(y \mid z; \theta, \alpha, \beta)$ using the full likelihood under the following model assumptions: $\tilde{Y} = \tilde{Y}_0 W e^{\beta z}$, z an observed binary covariate, $W \sim Gamma(\alpha, 1/\alpha)$, $\tilde{Y}_0 \sim q f_{normal}(\mu_N, \sigma_N^2) + (1-q) f_{Gumbel}(\mu_G, \sigma_G^2)$ and $\lambda(y) \propto y$. Data generated using algorithm enumerated in Appendix A.1 with 2 cases for $\pi(b)$: constant ($\zeta = 0$) and decreasing ($\zeta = 5.5$) with b. Number of simulations = 2000. Standard errors based on simulation estimates (Sim SE). Bootstrap BC_a CI constructed using 1000 bootstrap replicates. Two sample sizes: 250 and 500.

Setting	Parameter	True	Bias	Sim SE	MSE	BC_a CP
Const $\pi(b)$	$E(Y \mid z=0)$	31.95	0.03	0.81	0.66	92.80
(n=250)	$E(Y \mid z=1)$	30.39	-0.04	0.83	0.70	93.05
	$SD(Y \mid z=0)$	9.70	0.02	1.08	1.17	94.45
	$SD(Y \mid z=1)$	9.23	0.00	1.04	1.08	94.35
	β	-0.05	0.00	0.03	0.00	93.05
	$SD(W)$	0.10	0.00	0.00	0.00	78.70
Setting	Parameter	True	Bias	Sim SE	MSE	BC_a CP
Const $\pi(b)$	$E(Y \mid z=0)$	31.95	0.01	0.55	0.31	94.90
(n=500)	$E(Y \mid z=1)$	30.39	-0.04	0.56	0.31	91.50
	$SD(Y \mid z=0)$	9.70	0.02	0.72	0.52	94.75
	$SD(Y \mid z=1)$	9.23	0.00	0.69	0.48	94.65
	β	-0.05	0.00	0.02	0.00	90.70
	$SD(W)$	0.10	0.00	0.00	0.00	75.35
Setting	Parameter	True	Bias	Sim SE	MSE	BC_a CP
Decr $\pi(b)$	$E(Y \mid z=0)$	31.95	0.04	0.76	0.58	94.75
(n=250)	$E(Y \mid z=1)$	30.39	-0.03	0.76	0.58	94.90
	$SD(Y \mid z=0)$	9.70	0.02	1.10	1.21	95.70
	$SD(Y \mid z=1)$	9.23	0.00	1.05	1.11	95.35
	β	-0.05	0.00	0.02	0.00	93.35
	$SD(W)$	0.10	0.00	0.00	0.00	81.15
Setting	Parameter	True	Bias	Sim SE	MSE	BC_a CP
Decr $\pi(b)$	$E(Y \mid z=0)$	31.95	0.01	0.53	0.29	95.45
(n=500)	$E(Y \mid z=1)$	30.39	-0.03	0.52	0.27	92.50
	$SD(Y \mid z=0)$	9.70	0.01	0.77	0.59	95.40
	$SD(Y \mid z=1)$	9.23	0.00	0.73	0.54	95.25
	β	-0.05	0.00	0.02	0.00	92.70
	$SD(W)$	0.10	0.00	0.00	0.00	73.45

Table 2.2: Performance of estimators of $S(y \mid z; \theta, \alpha, \beta)$ using the conditional likelihood under the following model assumptions: $\tilde{Y} = \tilde{Y}_0 W e^{\beta z}$, z an observed binary covariate, $W \sim Gamma(\alpha, 1/\alpha)$, $\tilde{Y}_0 \sim q f_{normal}(\mu_N, \sigma_N^2) + (1-q) f_{Gumbel}(\mu_G, \sigma_G^2)$ and $\lambda(y) \propto y$. Data generated using algorithm enumerated in Appendix A.1 with 2 cases for $\pi(b)$: constant ($\zeta = 0$) and decreasing ($\zeta = 5.5$) with b. Number of simulations = 2000. Standard errors based on simulation estimates (Sim SE). Bootstrap BC_a CI constructed using 1000 bootstrap replicates. Two sample sizes: 250 and 500.

Setting	Parameter	True	Bias	Sim SE	MSE	BC_a CP
Const $\pi(b)$	$E(Y \mid z=0)$	31.95	0.04	0.81	0.66	92.85
(n=250)	$E(Y \mid z=1)$	30.39	0.01	0.81	0.66	93.25
	$SD(Y \mid z=0)$	9.70	0.02	1.08	1.17	94.50
	$SD(Y \mid z=1)$	9.23	0.02	1.04	1.08	94.30
	β	-0.05	0.00	0.03	0.00	95.25
	$SD(W)$	0.10	0.00	0.00	0.00	78.95
Setting	Parameter	True	Bias	Sim SE	MSE	BC_a CP
Const $\pi(b)$	$E(Y \mid z=0)$	31.95	0.01	0.55	0.31	95.00
(n=500)	$E(Y \mid z=1)$	30.39	-0.01	0.54	0.29	95.50
	$SD(Y \mid z=0)$	9.70	0.02	0.72	0.52	94.70
	$SD(Y \mid z=1)$	9.23	0.01	0.69	0.47	95.05
	β	-0.05	0.00	0.02	0.00	96.20
	$SD(W)$	0.10	0.00	0.00	0.00	72.25
Setting	Parameter	True	Bias	Sim SE	MSE	BC_a CP
Decr $\pi(b)$	$E(Y \mid z=0)$	31.95	0.04	0.76	0.58	94.75
(n=250)	$E(Y \mid z=1)$	30.39	-0.01	0.76	0.58	95.10
	$SD(Y \mid z=0)$	9.70	0.02	1.10	1.21	95.80
	$SD(Y \mid z=1)$	9.23	0.00	1.06	1.12	95.35
	β	-0.05	0.00	0.02	0.00	95.00
	$SD(W)$	0.10	0.00	0.00	0.00	81.05
Setting	Parameter	True	Bias	Sim SE	MSE	BC_a CP
Decr $\pi(b)$	$E(Y \mid z=0)$	31.95	0.01	0.53	0.29	95.65
(n=500)	$E(Y \mid z=1)$	30.39	-0.01	0.52	0.27	95.85
	$SD(Y \mid z=0)$	9.70	0.01	0.77	0.59	95.25
	$SD(Y \mid z=1)$	9.23	0.00	0.73	0.54	95.15
	β	-0.05	0.00	0.00	0.00	96.50
	$SD(W)$	0.10	0.00	0.02	0.00	72.25

2.7. Application to the LIFE Study

We apply the proposed semi-parametric approach to estimate the enrollment probability function and the distribution of menstrual cycle length for the population studied in the LIFE Study, focusing on enrollment cycle data with adjustment for length-bias and selection features of the sampling plan. For this prospective pregnancy study, recruitment took place in the U.S. in Michigan and Texas in 2005-2009 with minimal criteria for eligibility listed as follows: female ages 18-44 years and male ages 18+ years; in a committed relationship; ability to communicate in English or Spanish; menstrual cycle length between 21 and 42 days; no hormonal contraception injections during past year; and no sterilization procedures or physician diagnosed infertility. Prior to conception, 501 couples satisfying the eligibility criteria and interested in becoming pregnant were enrolled on a calendar date irrespective of the day of the woman's menstrual cycle and followed until pregnancy, exit from the study, or censored at the completion of 12 menstrual cycles at-risk for pregnancy. Under these inclusion criteria and protocol, the sampling plan is independent of the participants' cycle lengths. For further details, see the complete study protocol (Buck Louis et al. 2011).

The enrollment menstrual cycle length was observed as the sum of two segments: the backward recurrence time, which was recalled on the enrollment date as the time since LMP; and the residual time, which was prospectively observed from enrollment until the next menstrual bleeding event determined from the woman's daily recorded level of bleeding or spotting (0 = none to 4 = heavy), using an algorithm to distinguish menstrual bleeding from non-menstrual bleeding. The residual time was censored if the couple exited the study for one or more of the following reasons: no longer wishing to participate, relocation, diagnosis of sterility, and non-compliance with study protocol; or if pregnancy was detected using Clearblue®Easy pregnancy tests which detect levels of hCG over 25 mIU/mL. For couples becoming pregnant in the enrollment cycle, we would have ideally censored the residual time on the day on which the next menstrual bleeding is precluded by the absence of a decline in estrogen and progesterone. In the absence of knowledge of this date, we censored on the

date of the first positive pregnancy date.

We consider the enrollment cycles of 467 couples for which both the backward recurrence time and residual time (or censoring time) could be determined, with total observed length in the range of (8, 90) days and backward recurrence time less than 50 days. Since a menstrual cycle length of 90 days is more than twice the study's inclusion criteria of 42 days, we have excluded lengths outside this range as they are likely an artifact of a missing cycle stop/start. Furthermore, we have excluded 4 women because the onset date of the enrollment menstrual cycle could not be determined due to breastfeeding, miscarriage, or discontinuation of hormonal birth control. This restriction excludes women who recently were pregnant which may be important if cycle length is associated with pregnancy, but this is a very small number of exclusions. We excluded 6 women who exited the study before recording any bleeding information due to non-compliance or no longer wishing to participate. This exclusion is based on the assumption that these exits were not associated with menstrual cycle length.

We estimate the marginal distribution of menstrual cycle length using model (2.10) for the observed menstrual cycle length, with $z = 0$ for the women aged < 30 years and $z = 1$ for the women aged ≥ 30 years, where 30 is the median age; and we assume $W \sim Gamma(\alpha, 1/\alpha)$ so $E(W) = 1$ and $Var(W) = 1/\alpha$. Motivated by the statistical model for menstrual cycle length implemented by McLain, Lum, and Sundaram (2012), we assume $\tilde{Y}_0$ has a normal-Gumbel mixture distribution with unknown parameter vector: $\theta = (q, \mu_N, \sigma_N, \mu_G, \sigma_G)$, where q is the proportion of normally distributed lengths and (μ_N, σ_N) and (μ_G, σ_G) are the mean and standard deviation of the normal and Gumbel distributions, respectively. As described in Section 2.A.2, we find this parametric choice for $\tilde{Y}_0$ fits the menstrual cycle length data better than the log-normal model. We assume $\lambda(y) \propto y$ and provide a check of this assumption as suggested by a reviewer in Section 2.A.3. Rather than assuming a model for $\pi(\cdot \mid z)$ in the full likelihood (see 2.12), we use the semi-parametric approach proposed in Section 2.5.1 to estimate the shape of $\pi(\cdot \mid z)$ and (θ, β, α).

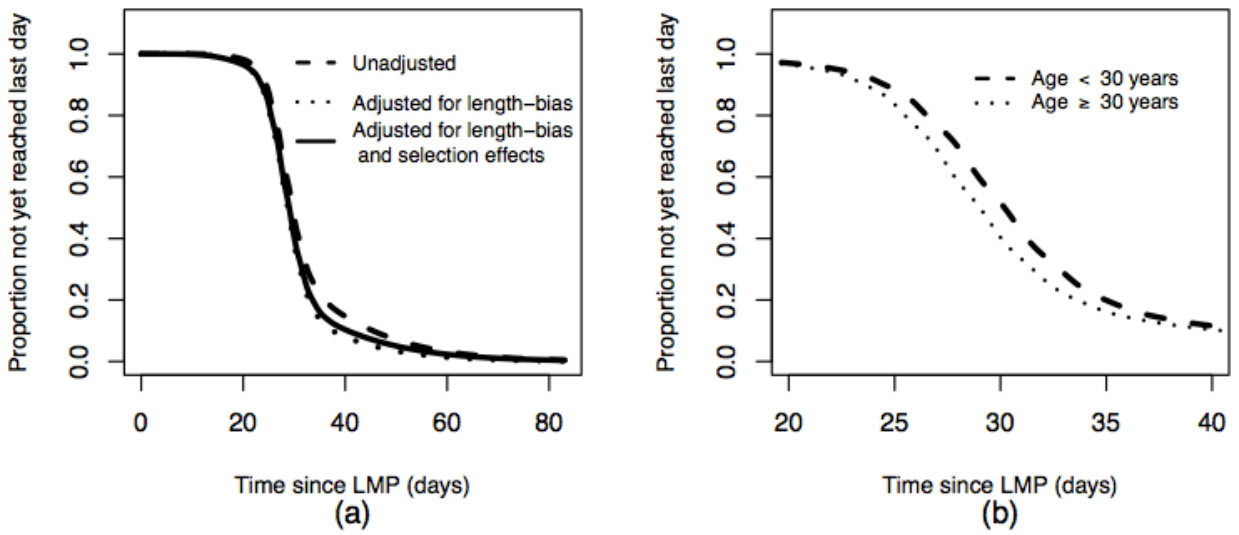


Figure 2.1: Parametric estimates of marginal survivor function of menstrual cycle length (a) for women aged < 30 years based on enrollment cycle data, with adjustments for both length-bias and selection effects (solid), length-bias only (dotted), and without adjustments (dashed); and (b) for women aged < 30 years (dashed) compared with women aged ≥ 30 years (dotted) based on enrollment cycle data with adjustments for both length-bias and selection effects, zoomed in to a range of 20-40 days. Restricted to $n = 467$ enrollment cycles of known observed length within (8, 90) days and with backward recurrence time less than 50 days.

In order to illustrate the effects of adjusting for the sampling plan, we show the marginal survival curves (Fig. 2.1a) for women aged < 30 years, point estimates and 95% confidence intervals (Table 2.3) with and without adjustment for length-bias and selection effects. The estimates of the unadjusted marginal survivor function and marginal means and standard deviations are located above those which have been adjusted, due to the over-representation of long cycles. The marginal survivor function that adjusts for both length-bias and selection effects lies in between the other curves, suggesting some cancellation of the length-bias by the selection effects. The same pattern was observed for the women aged ≥ 30 years (not shown). The point estimate and confidence interval of the standard deviation of the random effect are small indicating that the residual variation is small. Of biological interest is that the estimated marginal survivor curve (see Fig. 2.1b) and marginal mean and standard deviation

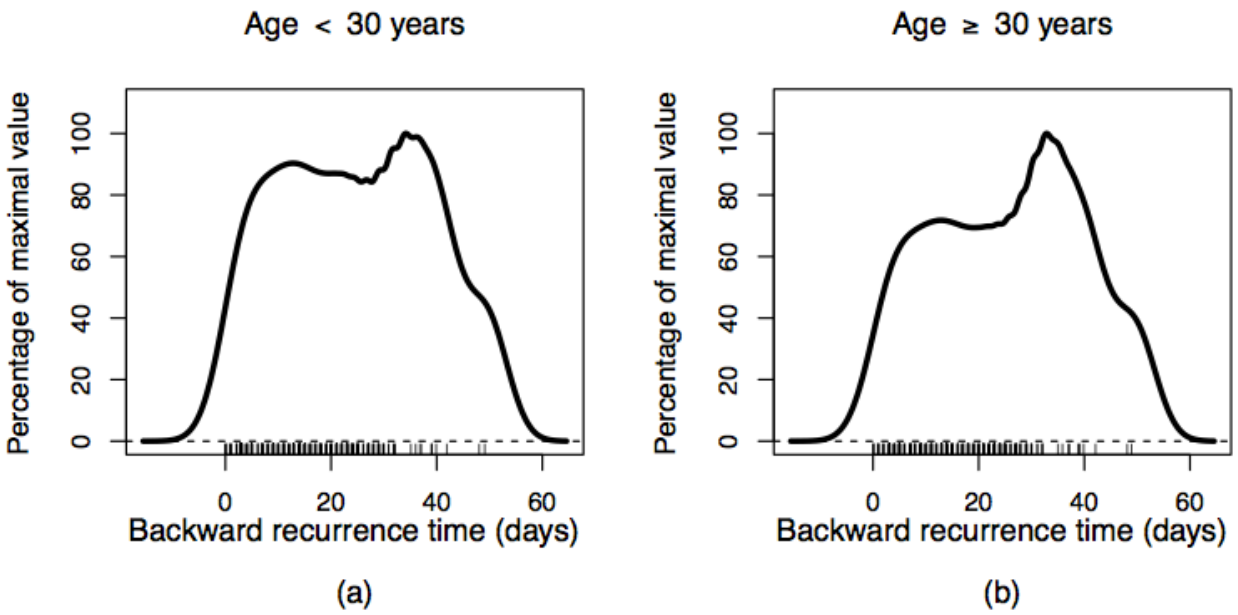


Figure 2.2: Smoothed shape of empirical estimate of $\pi(b \mid z)$ for LIFE Study for women aged < 30 years (a) and women aged ≥ 30 years (b), scaled so that maximum value is 100%. Rug shows backward recurrence times. Note that a non-smoothed estimate of the shape of $\pi(b \mid z)$ was used to calculate the estimates shown in column 3 of Table 2.3. Restricted to $n = 467$ enrollment cycles of known observed length within (8, 90) days and with backward recurrence time less than 50 days.

Table 2.3: Estimates of parameters of menstrual cycle distribution for LIFE Study population based on enrollment cycle data with and without adjustments for length-bias (LB) and selection effects (S). Model assumptions: $\tilde{Y} = \tilde{Y}_0 W e^{\beta z}$, $z = 0$ (age < 30), $z = 1$ (age ≥ 30), $W \sim Gamma(\alpha, 1/\alpha)$, $\tilde{Y}_0 \sim q f_{normal}(\mu_N, \sigma_N^2) + (1-q) f_{Gumbel}(\mu_G, \sigma_G^2)$. Adjustments for LB and S done using semi-parametric algorithm to estimate model parameters and $\pi(\cdot \mid z)$ assuming $\lambda(y) \propto y$. Point estimate with BC_a 95% confidence limits displayed as: ${}_{Lo}Est_{Up}$ (Louis and Zeger 2009). BC_a confidence intervals were calculated using non-parametric bootstrapping with 1000 replicates. Restricted to $n = 467$ enrollment cycles of known observed length within (8, 90) days and with backward recurrence time less than 50 days.

Parameter	Unadjusted	Adj. for LB	Adj. for LB and S
$E(\tilde{Y} \mid age < 30)$	${}_{32.55}33.76_{34.77}$	${}_{30.42}31.40_{32.31}$	${}_{31.01}32.18_{33.39}$
$E(\tilde{Y} \mid age \geq 30)$	${}_{31.30}32.51_{33.71}$	${}_{29.12}30.13_{31.00}$	${}_{29.91}31.07_{32.22}$
$SD(\tilde{Y} \mid age < 30)$	${}_{8.70}10.92_{12.30}$	${}_{7.59}8.56_{9.84}$	${}_{8.54}10.04_{11.98}$
$SD(\tilde{Y} \mid age \geq 30)$	${}_{8.50}10.51_{11.99}$	${}_{7.31}8.21_{9.42}$	${}_{8.21}9.69_{11.66}$
$\beta : age$	${}_{-0.08}-0.04_{-0.01}$	${}_{-0.08}-0.04_{-0.01}$	${}_{-0.07}-0.04_{-0.01}$
α	${}_{82.69}85.24_{88.00}$	${}_{73.28}80.94_{84.91}$	${}_{71.41}82.88_{85.85}$
$SD(W)$	${}_{0.11}0.11_{0.11}$	${}_{0.11}0.11_{0.12}$	${}_{0.11}0.11_{0.12}$
q	${}_{0.54}0.64_{0.70}$	${}_{0.62}0.71_{0.78}$	${}_{0.60}0.70_{0.77}$
μ_N	${}_{28.93}29.83_{31.11}$	${}_{29.00}29.72_{30.88}$	${}_{28.98}29.73_{30.70}$
σ_N	${}_{0.65}0.76_{1.35}$	${}_{0.70}0.83_{1.69}$	${}_{0.69}0.81_{1.72}$
μ_G	${}_{37.54}40.72_{42.87}$	${}_{32.72}35.47_{38.66}$	${}_{34.13}37.78_{41.56}$
σ_G	${}_{10.31}14.59_{16.74}$	${}_{11.15}13.45_{16.79}$	${}_{12.39}15.49_{19.99}$

for the women aged ≥ 30 years is below that of women aged < 30 years which agrees with the findings of Harlow, Lin, and Ho (2000). If estimation of $\pi(\cdot \mid z)$ is not of interest, we can alternatively estimate the parameters by maximizing the censored form of the conditional likelihood, $\mathcal{L}_c$, in (2.11). For these data, the point estimates and standard errors for the full and conditional approaches are approximately identical (not shown); however, this is not always the case as seen in the simulation results in Table 2.2. Additionally, in Section 2.A.4, we compare the unadjusted and adjusted estimated distribution of $\tilde{Y}$ based on enrollment cycle data to the estimated distribution based on first post-enrollment cycle data.

Although the estimates shown in Fig. 2.1 and Table 2.3 are based on an empirical estimate of $\pi(b \mid z)$, for the purpose of visualizing the enrollment probability function we display a smoothed version of the estimate for each age group in Fig. 2.2 which was obtained using a kernel density smoothing approach with oversmoothed bandwidth selection (Wand and Jones 1994), implemented in `R` via the function `bkde`. The magnitude of the estimate is scaled by its maximum so that the values on the y-axis are the percent of maximum value. For both groups, the shape of the smoothed estimate is approximately uniform for most values of b; therefore, we do not see evidence of selection effects with regard to the length of the backward recurrence time.

2.8. Discussion

We have considered the problem of estimating the distribution of menstrual cycle length from enrollment cycle data for which the distribution is potentially weighted by features of the sampling process, in particular length-bias and selection effects. Addressing these features in the likelihood, we developed a two-stage semi-parametric approach for estimating the population distribution along with the shape of the enrollment probability. Using the empirical probability mass function to estimate the distribution of backward recurrence times, we obtained approximately unbiased estimates of the parameters of the distribution and traded very little variance for estimating the shape of the enrollment probability.

In the analysis of the LIFE Study data, we showed that adjustment for both length-bias and selection effects leads to visible differences in parameter and survivor curves estimates. These differences may be especially important when making risk predictions for TTP or disease, or when studying covariate effects in studies where an effect size on the order of 1-3 days is not uncommon such as those investigating the effects of environmental pollutants on menstrual cycle length (see, e.g. Buck Louis et al. 2011a).

This approach is particularly useful in settings where there is a desire to model the first gap time or an interest in the characteristics of the selection process. For example, one could compare the estimated enrollment probability across multiple studies or recruitment strategies and, if a significant difference is found, then account for selection effects to make inference on a reference population.

As is done in many studies involving a renewal process, the backward recurrence time was ascertained using self-report at enrollment. Since the inclusion criteria restricted to women with regular cycles, we expect the recall bias and uncertainty to be small. As is common with retrospective recalled information, we do observe some digit preference, in particular for an LMP of 7, 14, and 21 days before enrollment. In a more comprehensive analysis of these data, we could conduct a sensitivity analysis on digit preference generating distributions around those digits.

We investigated the possibility that an undetected spontaneous abortion (SAB) could make the enrollment cycle different from post-enrollment cycles in addition to length-bias and selection. Based on the eligibility screening, the couples must have been off contraception for 0, 1, or 2 cycles prior to enrollment. Seventy percent of the couples who enrolled had been off contraception for 0 cycles. A pregnancy test was conducted on day of enrollment using the Clearblue®Easy pregnancy test which has been shown to be sensitive for detecting 25 mIU/mL of hCG (Cole 2011); therefore the probability that a SAB would not have been detected on enrollment day is low. Couples who had been trying for 1-2 cycles and who had a positive test prior to enrollment would not have enrolled. In the case of a couple who did not

test for pregnancy prior to enrollment, if an SAB occurred and the woman mistook bleeding for menses, the woman's hCG level would have had to return to less than 25mIU/mL by the time of the pregnancy test on the day of enrollment, an unlikely combination of events. Furthermore all bleeding, including spotting, was recorded as a categorical variable indicating intensity in the female daily journal starting on the day of enrollment. The possibility of SAB after enrollment was investigated when the data was completed and it was concluded that the majority of the long cycles could not be contributed to SAB.

Focusing solely on enrollment cycle data, we were able to obtain relatively small confidence intervals for all parameters of the menstrual cycle length distribution, suggesting potential gain for inclusion of these data even when post-enrollment data are available. Our future work will focus on modeling the full history of menstrual cycle data from enrollment to pregnancy along with the time-to-pregnancy process. In modeling the pregnancy cycles for this work, we used time-to-positive test date as a proxy for the date on which the menstrual cycle is paused; however, the true date is some time before the positive test date. Improvements can be made using data on intercourse patterns, ovulation, and pregnancy tests and by conditioning on the date for which the menstrual cycle is paused being earlier than the positive pregnancy test.

This work impacts the analysis of existing longitudinal data by providing an approach to estimate $\pi(\cdot)$ for a particular study and to use this estimate to adjust inferences to a reference population. The estimate of $\pi(\cdot)$ can also be useful in designing recruitment for a new study. Additionally, it might be of interest to collect data on the number of participants who decline enrollment in order to estimate the magnitude of $\pi(\cdot)$.

Appendices

2.A.1. Proofs

2.A.1.1 The Conditional Distribution of Backward Recurrence Times

In the sampling plan of the renewal process described in Section 2.2.1 of the manuscript, we assume that the study announcement, t^*, is stochastically independent of the sequence: $T_m = T_0+\tilde{Y}_1+\tilde{Y}_2+\cdots+\tilde{Y}_m$. Here, we show how our subsequent claim that $[\tilde{B} \mid \tilde{Y}_{m(t^*)}] \sim \mathcal{U}(0, \tilde{Y}_{m(t^*)})$ follows from a theorem proven by Winter (1989). To simplify notation, henceforth we drop the " ˜ " notation and subscript $m(t^*)$ and show that this follows from Winter (1989)'s proof that $B = UY$ where $U \sim \mathcal{U}(0,1)$ and $Y \sim g_Y(\cdot)$.

First, by the transformation, $B(U,Y) = UY$ and $W(U,Y) = Y$, the joint distribution of (B,W) is:

$$\begin{aligned} f_{B,W}(b,w) =& \frac{f_{U,Y}(\frac{b}{y},w)}{\left|\begin{pmatrix} y & u \\ 0 & 1 \end{pmatrix}\right|} \\ =& \frac{f_{U,Y}(\frac{b}{y},w)}{y} \\ =& \frac{1}{y}\mathbb{1}\{0 < b/y < 1\} g_Y(w) \text{ by } U \perp\!\!\!\perp Y \\ =& \frac{1}{y}\mathbb{1}\{0 < b < y\} g_Y(w). \end{aligned}$$

Therefore, the joint distribution of (B,Y) is:

$$f_{B,Y}(b,y) = \frac{1}{y}\mathbb{1}\{0 < b < y\} g_Y(y).$$

Then, the conditional distribution of B given Y is:

$$\begin{aligned} f_{B|Y}(b \mid y) =& \frac{f_{B,Y}(b,y)}{g_Y(y)} \\ =& \frac{1}{y}\frac{\mathbb{1}\{0 < b < y\} g_Y(y)}{g_Y(y)} \\ =& \frac{1}{y}\mathbb{1}\{0 < b < y\}. \end{aligned}$$

Thus, $[\tilde{B} \mid \tilde{Y}_{m(t^*)}] \sim \mathcal{U}(0, \tilde{Y}_{m(t^*)})$.

This concludes the proof.

2.A.1.2 Solution to the Integral Equation

Define

$$\beta(u) := \int_0^u \frac{dG_B(z)}{\int_z^\infty f(t)\frac{\lambda(t)}{t}\,dt} = \int_0^u \frac{g_B(z)dz}{\int_z^\infty f(t)\frac{\lambda(t)}{t}\,dt}, \tag{13}$$

where $g_B(z)$ is the marginal density of B. The following observation is crucial, namely,

$$\int_0^\infty \beta(u)f(u)\frac{\lambda(u)}{u}\,du = 1. \tag{14}$$

To see this, note that

$$\begin{aligned}
\int_0^\infty \beta(u)f(u)\frac{\lambda(u)}{u}\,du &= \int_0^\infty f(u)\frac{\lambda(u)}{u}\left(\int_0^u \frac{g_B(z)dz}{\int_z^\infty f(t)\frac{\lambda(t)}{t}\,dt}\right)du \\
&= \int_0^\infty \frac{g_B(z)}{\int_z^\infty f(t)\frac{\lambda(t)}{t}\,dt}\left(\int_z^\infty f(u)\frac{\lambda(u)}{u}\,du\right)dz \qquad (15) \\
&= \int_0^\infty g_B(z)\,dz = 1,
\end{aligned}$$

where in (15), we used Tonelli's theorem to change the order of integration as the integrand is non-negative.

Recall now that as $B < V, a.s.$ and the support of f and g_B are included in $[0, V]$, all integrals can be truncated to the interval $[0, V]$ and we are interested in $\pi(b)$ for $b \leq V$. Define $\gamma(u) = \int_0^u \pi(t)dt$. Consider now the integral equation in γ given by

$$\gamma(u) = \beta(u)\left(\int_0^\infty f(t)\frac{\lambda(t)}{t}\gamma(t)\,dt\right) = \beta(u)\left(\int_0^V f(t)\frac{\lambda(t)}{t}\gamma(t)\,dt\right), 0 \leq u \leq V. \tag{16}$$

In order that the integral equation is meaningful, we will need to restrict our attention to those γ such that

$$\|\gamma\|_1 := \int_0^V f(t)\frac{\lambda(t)}{t}|\gamma(t)|\,dt < \infty.$$

Let $X = \{\gamma : [0, V] \to \mathbb{R}, \|\gamma\|_1 < \infty\}$. Clearly, X is a Banach space.

Consider the mapping $\gamma \to T\gamma$ given by

$$(T\gamma)(u) = \beta(u) \left(\int_0^V f(t) \frac{\lambda(t)}{t} \gamma(t)\, dt \right).$$

Note that f, λ, β etc are non-negative functions and therefore,

$$\begin{aligned} \|T\gamma\|_1 &= \int \beta(u) f(u) \frac{\lambda(u)}{u} du \left| \int_0^V f(t) \frac{\lambda(t)}{t} \gamma(t)\, dt \right| \\ &\leq \int \beta(u) f(u) \frac{\lambda(u)}{u} du \int_0^V f(t) \frac{\lambda(t)}{t} |\gamma(t)|\, dt \\ &= \|\gamma\|_1, \end{aligned}$$

where we used (14) and the definition of $\|\gamma\|_1$. This shows that T maps the unit ball B in X into the unit ball B, i.e., T is a self map of the unit ball. Moreover, $T(B) = B \cap \{c\beta : c \in \mathbb{R}\}$, which immediately implies that the image $T(B)$ is a compact set in B (as it is the intersection of a one dimensional subspace and the unit ball B in the Banach space X). Applying Schauder's fixed point theorem (Conway 1990), we infer that there exists a fixed point of the map T, i.e., a solution of the integral equation (16).

Concerning uniqueness, it is easy to see that if γ is a solution of (16), then so is $c\gamma$ for any $c \in \mathbb{R}$. Thus uniqueness cannot be expected in the strict sense. However, the solution is unique modulo a proportionality constant. To see this, for $\gamma \in X$, define the constant (which depends on γ) $c_\gamma = \int_0^V f(t) \frac{\lambda(t)}{t} \gamma(t)\, dt$. If γ_1, γ_2 satisfy (16), then $\gamma_1(u) = c_{\gamma_1} \beta(u)$ and $\gamma_2(u) = c_{\gamma_2} \beta(u)$. If both $c_{\gamma_1} = c_{\gamma_2} = 0$, then clearly $\gamma_1 = \gamma_2 = 0$. Now assuming $c_{\gamma_1} \neq 0$, we may divide and conclude $\frac{\gamma_1(u)}{\gamma_2(u)} = \frac{c_{\gamma_2}}{c_{\gamma_1}}$. The right hand side is a constant independent of u. This concludes the proof.

2.A.1.3 Alternative Method for Solving for π

Using (14), it is immediate that $\gamma_1(u) = \beta(u)$ solves (16). Differentiating (and using the fundamental theorem of calculus), we get $\pi_1(u) = \frac{g_b(u)}{\int_u^\infty f(t) \frac{\lambda(t)}{t}\, dt}$. Note that $\pi_1(u) \geq 0$. However, note that π is a probability and therefore $\pi(u) \leq 1$ for all $0 \leq u \leq V$. In case

$\|\pi_1\|_\infty := \sup_{0\le u\le V} \pi_1(u) < \infty$, we may simply take $\pi(u) = \frac{\pi_1(u)}{\|\pi_1\|_\infty}$. Since we know π_1 solves (16) and so does any scalar multiple of it, π will be a solution, which, as proven in the previous section, is unique up to a constant of proportionality.

2.A.2. Comparison of Model Fit: Normal-Gumbel vs Log-normal

In this section we compare the fit of the normal-Gumbel model for $\tilde{Y}_0$ to that of a log-normal model motivated by the skewness of the menstrual cycle length distribution. Using model (2.10) for the observed cycle length, we incorporate a gamma random effect W with mean 1 and variance $1/\alpha$ and a binary covariate that indicates age$\geq$30 years. We first consider the subset of 467 couples with known enrollment cycle lengths in the range (8,90) days and known backward recurrence times < 50 days. Estimates of the parameters adjusted for length-bias and selection effects using the approach in Section 2.5.1 are shown in Table 2.A.1. Comparing the AIC of each model, the normal-Gumbel mixture model is a better fit to the data. As a second comparison, we consider the first cycle post-enrollment for 304 couples who neither exited the study nor became pregnant in the first post-enrollment cycle and with known cycle length in the range of (8, 90) days and known backward recurrence time less than 50 days. Figure 2.A.1 compares the kernel density estimates for the two age groups. Table 2.A.2 displays the parameter estimates and Figures 2.A.2-2.A.5 show the estimated density and survival function curves comparing to non-parametric kernel density and Kaplan-Meier estimates, respectively. The normal-Gumbel model has an overall better fit to the data.

Table 2.A.1: Estimates of parameters of menstrual cycle distribution for LIFE Study population based on enrollment cycle data assuming a normal-Gumbel distribution for Y_0 versus assuming a log-normal distribution. Estimates calculated using semi-parametric algorithm to estimate model parameters and $\pi(\cdot \mid z)$ assuming $\lambda(y) \propto y$, $\tilde{Y} = \tilde{Y}_0 W e^{\beta z}$, $z = 0$ (age < 30), $z = 1$ (age ≥ 30), and $W \sim Gamma(\alpha, 1/\alpha)$. Point estimate with BC_a 95% confidence limits displayed as: ${}_{Lo}Est_{\,Up}$ (Louis and Zeger 2009). BC_a confidence intervals were calculated using non-parametric bootstrapping with 1000 replicates. Restricted to $n = 467$ couples with known observed length within (8, 90) days and with backward recurrence time less than 50 days.

Parameter	Normal-Gumbel	Log-normal
$E(\tilde{Y} \mid age < 30)$	${}_{31.01}32.18_{\,33.39}$	${}_{30.45}32.00_{\,33.28}$
$E(\tilde{Y} \mid age \geq 30)$	${}_{29.91}31.07_{\,32.22}$	${}_{28.72}30.09_{\,31.35}$
$SD(\tilde{Y} \mid age < 30)$	${}_{8.54}10.04_{\,11.98}$	${}_{7.37}8.50_{\,9.92}$
$SD(\tilde{Y} \mid age \geq 30)$	${}_{8.21}9.69_{\,11.66}$	${}_{6.97}7.99_{\,9.41}$
$\beta : age$	${}_{-0.07}-0.04_{\,-0.01}$	${}_{-0.12}-0.06_{\,-0.01}$
$SD(W)$	${}_{0.11}0.11_{\,0.12}$	${}_{0.01}0.11_{\,0.11}$
AIC	2799	2945

2.A.3. A Check of the Assumption: $\lambda(y) \propto y$

Beginning in Section 2.3 of the manuscript, we make the assumption $\lambda(y) \propto y$. Here we provide a check of this assumption for the menstrual cycle data from the LIFE Study as suggested by a reviewer.

In Section 2.2.2, the marginal distribution of Y in (2.1) is a weighted version of the reference distribution f with weight proportional to $\lambda(y)\frac{1}{y}\int_0^y \pi(t)dt$. If we assume

$$\frac{1}{y}\int_0^y \pi(t)dt \approx \bar{\pi},$$

then the ratio of $g_Y(y)$ to $f(y)$ is approximately equal to $C\lambda(y)$ where C is a constant of proportionality. If longitudinal data are available, the form of λ can be approximated by

Figure 2.A.1: Nonparametric kernel density estimates of menstrual cycle length density for women aged less than 30 years (red) versus women aged 30 years or older (blue) based on first post-enrollment cycle. Restricted to $n = 304$ couples who did not exit the study or become pregnant in the enrollment cycle and in the first post-enrollment cycle and for which these cycles are of known observed length within (8, 90) days and with backward recurrence time less than 50 days.

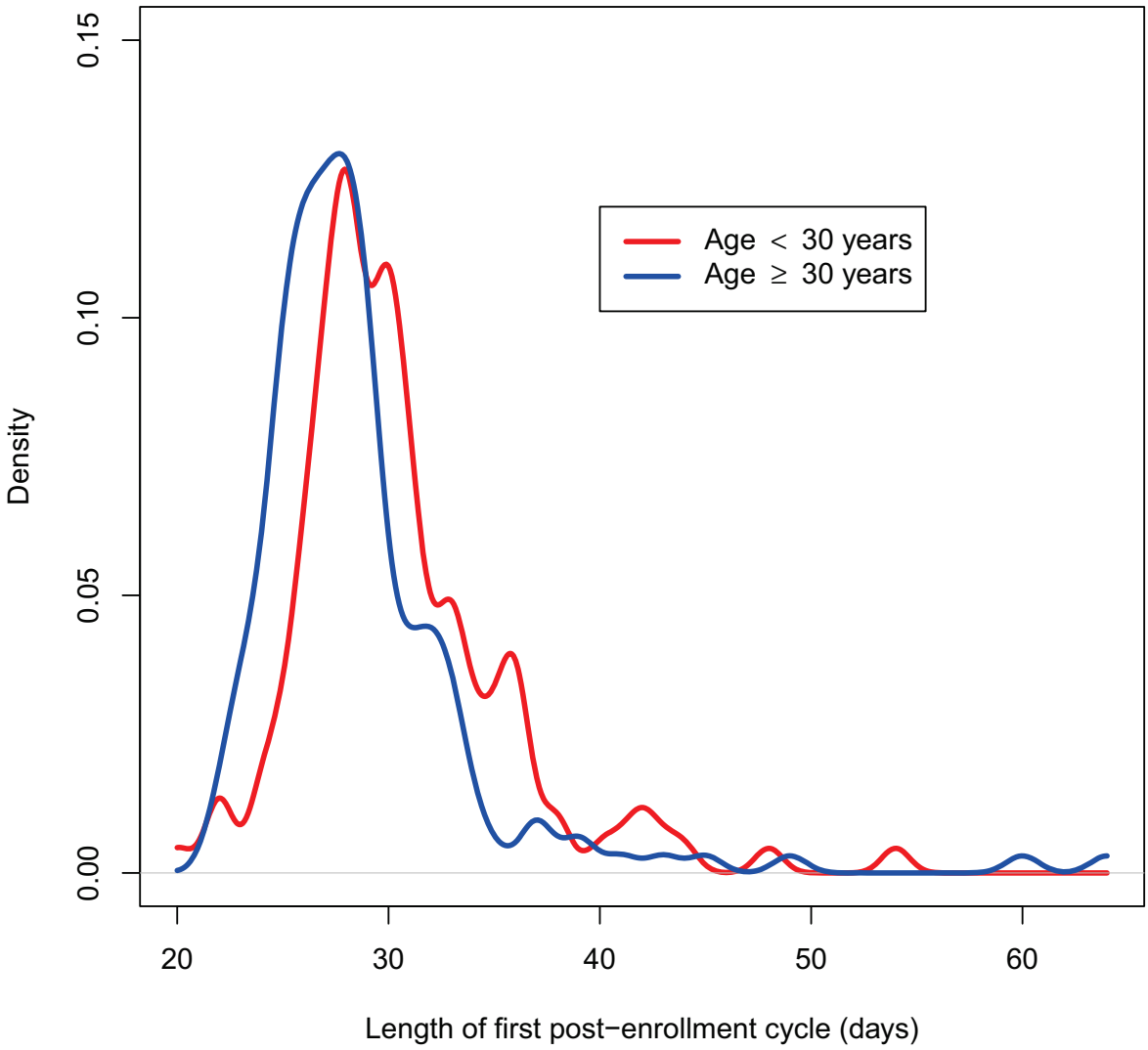


Figure 2.A.2: Parametric estimates of marginal density function of menstrual cycle length for women aged less than 30 years based on first post-enrollment cycle assuming a normal-Gumbel (red) vs assuming a log-normal (blue) model for Y_0 vs a non-parametric kernel density estimate (black). Restricted to $n = 304$ couples who did not exit the study or become pregnant in the enrollment cycle and in the first post-enrollment cycle and for which these cycles are of known observed length within (8, 90) days and with backward recurrence time less than 50 days.

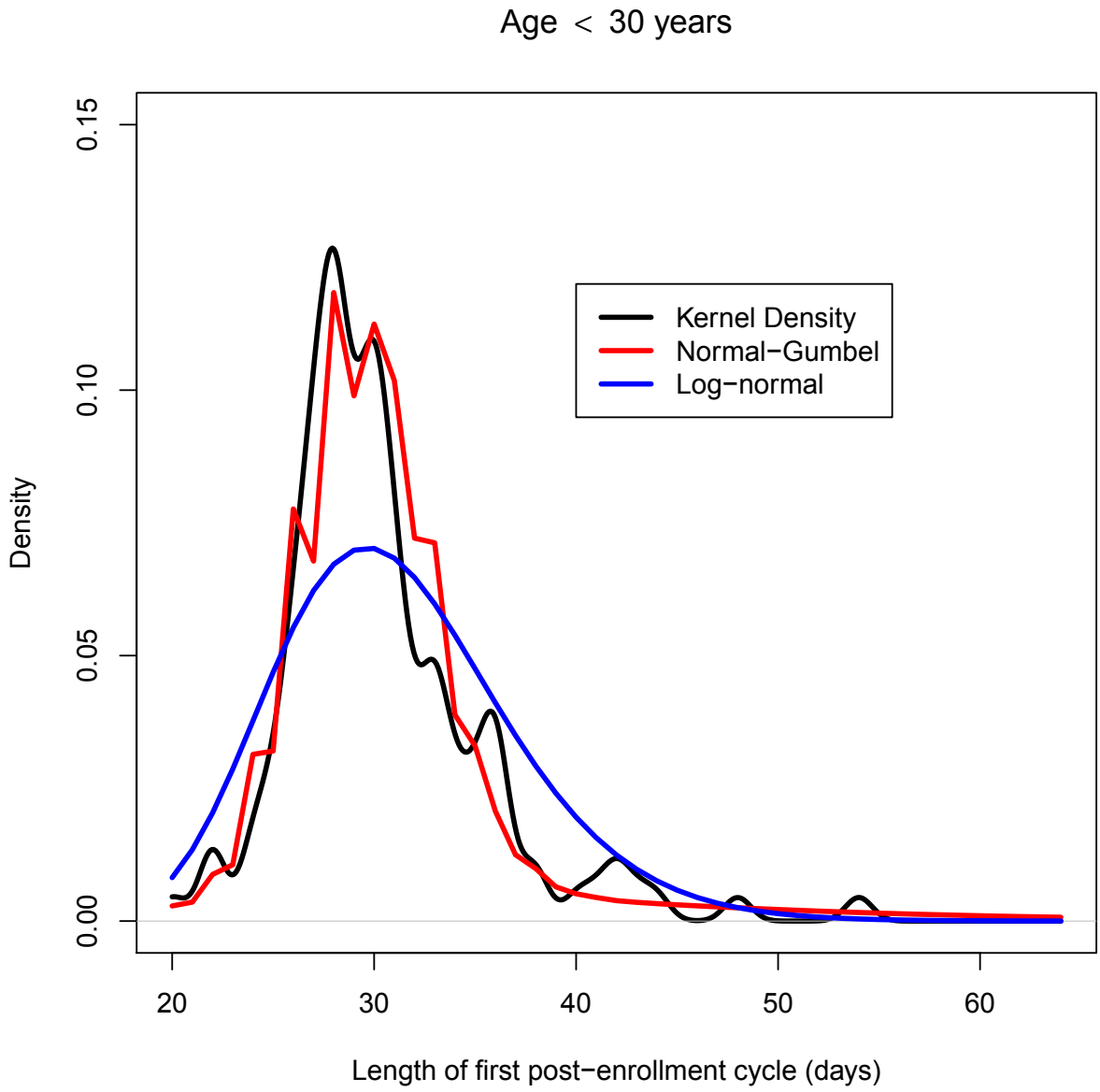


Figure 2.A.3: Parametric estimates of marginal density function of menstrual cycle length for women aged 30 years or older based on first post-enrollment cycle assuming a normal-Gumbel (red) vs assuming a log-normal (blue) model for Y_0 vs a non-parametric kernel density estimate (black). Restricted to $n = 304$ couples who did not exit the study or become pregnant in the enrollment cycle and in the first post-enrollment cycle and for which these cycles are of known observed length within (8, 90) days and with backward recurrence time less than 50 days.

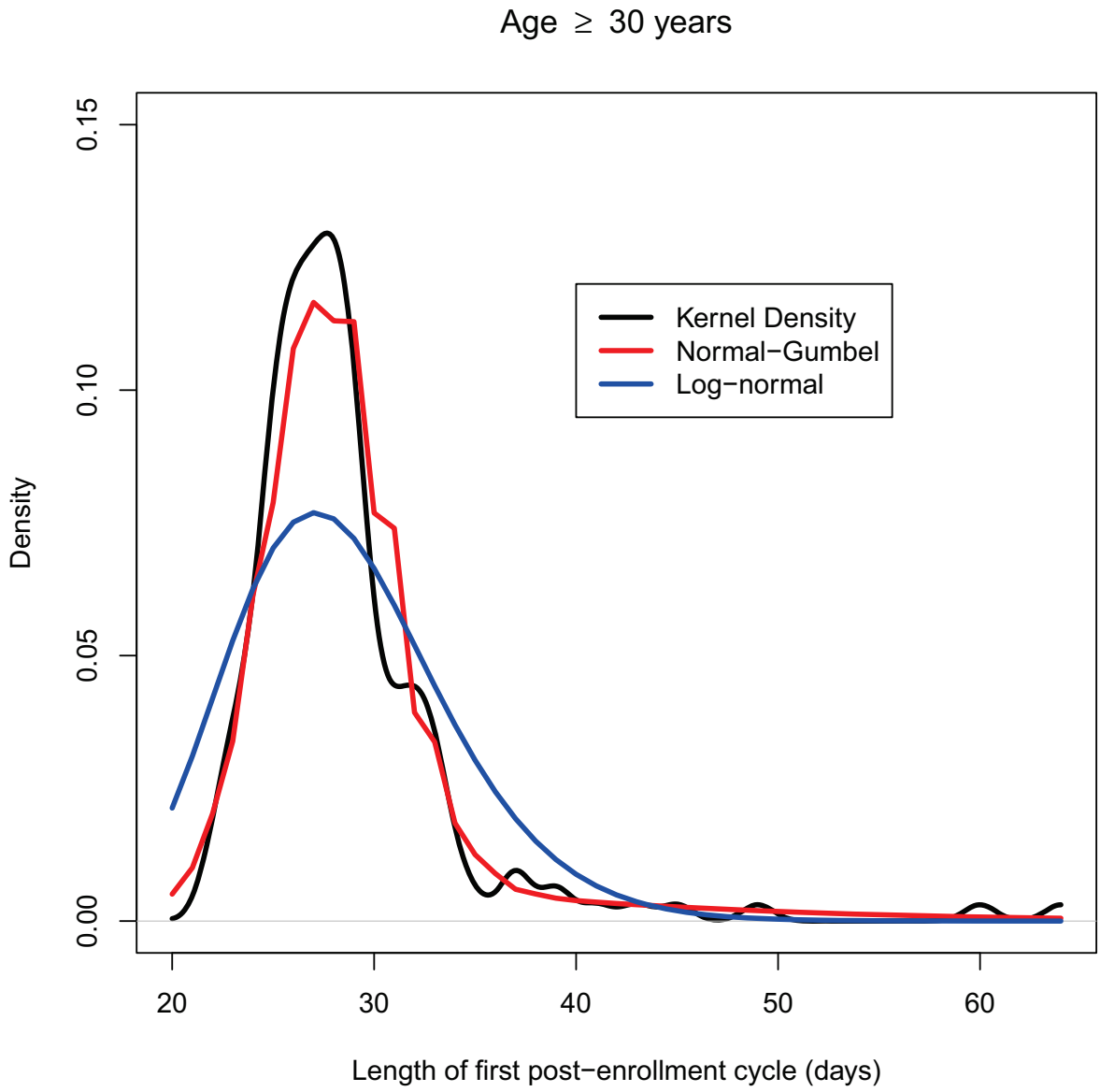


Figure 2.A.4: Parametric estimates of marginal survival function of menstrual cycle length for women aged less than 30 years based on first post-enrollment cycle assuming a normal-Gumbel (red) vs assuming a log-normal (blue) model for Y_0 vs a non-parametric Kaplan Meier (black). Restricted to $n = 304$ couples who did not exit the study or become pregnant in the enrollment cycle and in the first post-enrollment cycle and for which these cycles are of known observed length within (8, 90) days and with backward recurrence time less than 50 days.

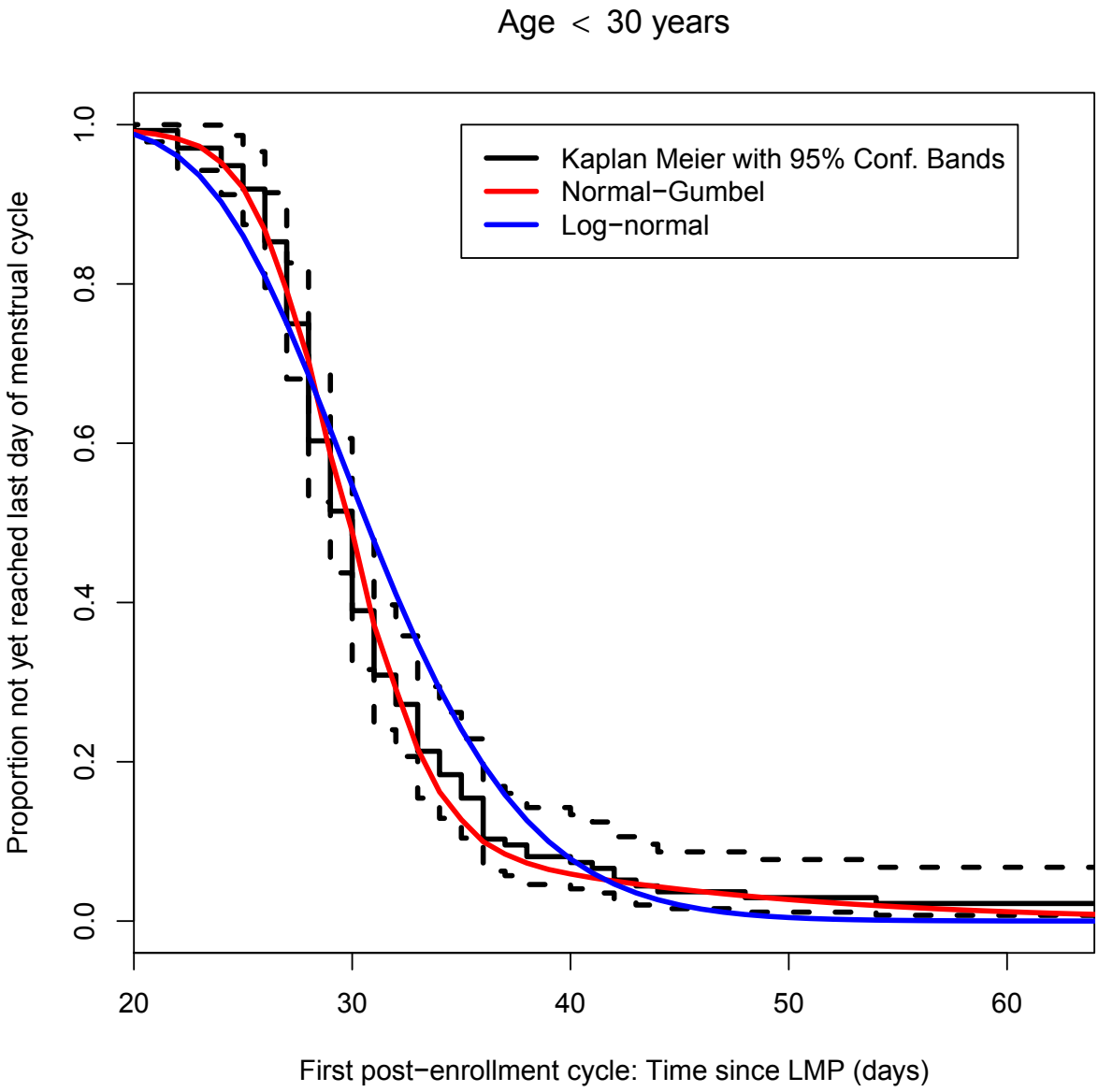


Figure 2.A.5: Parametric estimates of marginal survival function of menstrual cycle length for women aged 30 years or older based on first post-enrollment cycle assuming a normal-Gumbel (red) vs assuming a log-normal (blue) model for Y_0 vs a non-parametric Kaplan Meier estimate (black). Restricted to $n = 304$ couples who did not exit the study or become pregnant in the enrollment cycle and in the first post-enrollment cycle and for which these cycles are of known observed length within (8, 90) days and with backward recurrence time less than 50 days.

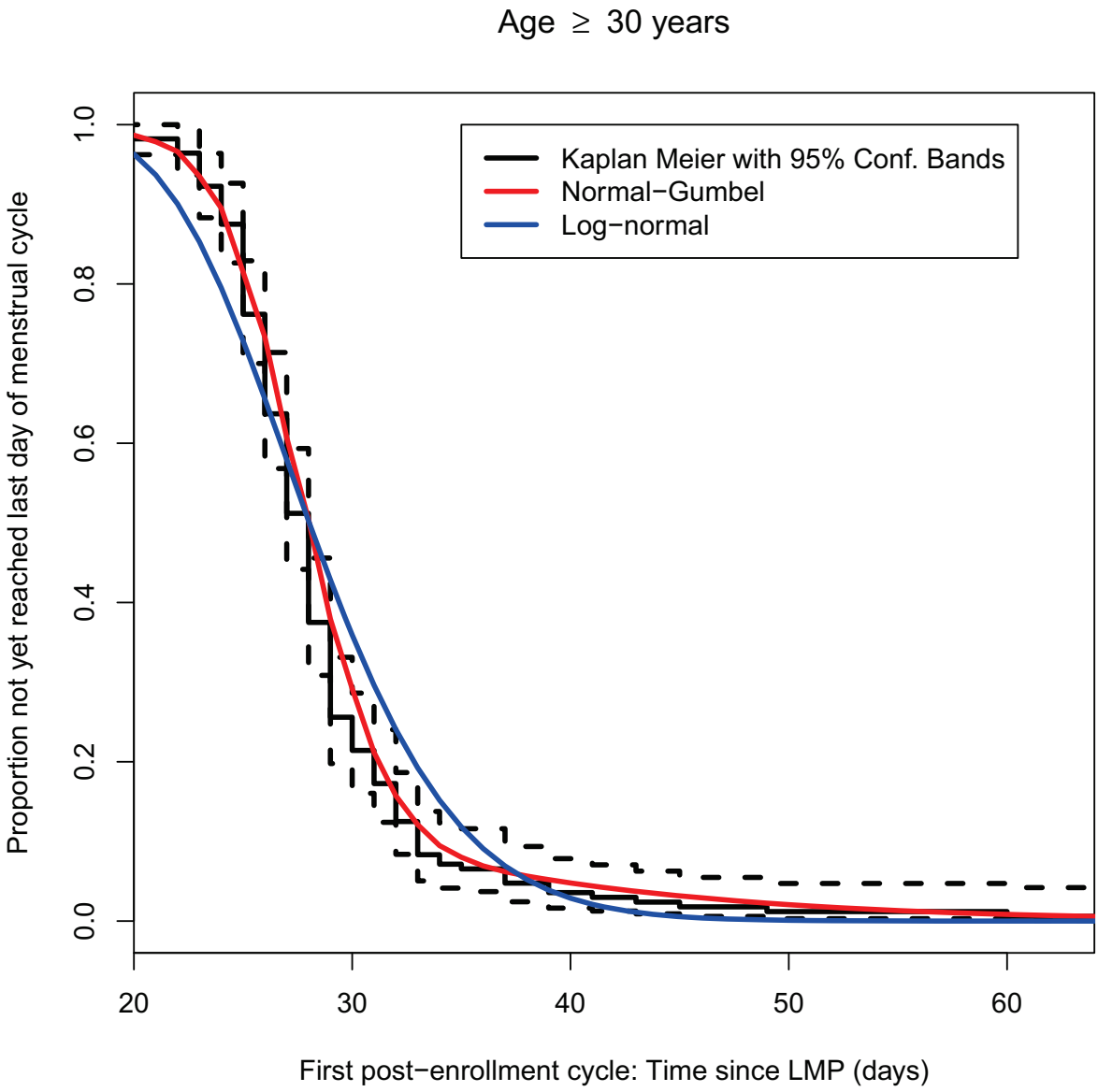


Table 2.A.2: Estimates of parameters of menstrual cycle distribution for LIFE Study population based on first post-enrollment cycle data assuming a normal-Gumbel distribution for Y_0 versus assuming a log-normal distribution. Estimates calculated assuming $\tilde{Y} = \tilde{Y}_0 W e^{\beta z}$, $z = 0$ (age < 30), $z = 1$ (age ≥ 30), and $W \sim Gamma(\alpha, 1/\alpha)$. Point estimate with BC_a 95% confidence limits displayed as: ${}_{Lo}Est_{\,Up}$ (Louis and Zeger 2009). BC_a confidence intervals were calculated using non-parametric bootstrapping with 1000 replicates. Restricted to $n = 304$ couples who did not exit the study or become pregnant in the enrollment cycle and in the first post-enrollment cycle and for which these cycles are of known observed length within (8, 90) days and with backward recurrence time less than 50 days.

Parameter	Normal-Gumbel	Log-normal
$E(\tilde{Y} \mid age < 30)$	${}_{30.11}30.98_{\,31.85}$	${}_{30.12}31.20_{\,32.45}$
$E(\tilde{Y} \mid age \geq 30)$	${}_{28.10}29.04_{\,30.17}$	${}_{27.74}28.52_{\,29.44}$
$SD(\tilde{Y} \mid age < 30)$	${}_{5.56}7.04_{\,9.10}$	${}_{4.86}5.91_{\,7.20}$
$SD(\tilde{Y} \mid age \geq 30)$	${}_{5.22}6.61_{\,8.56}$	${}_{4.52}5.40_{\,6.56}$
$\beta : age$	${}_{-0.10}-0.06_{\,-0.03}$	${}_{-0.13}-0.09_{\,-0.05}$
$SD(W)$	${}_{0.10}0.10_{\,0.11}$	${}_{0.01}0.10_{\,0.10}$
AIC	1780	1901

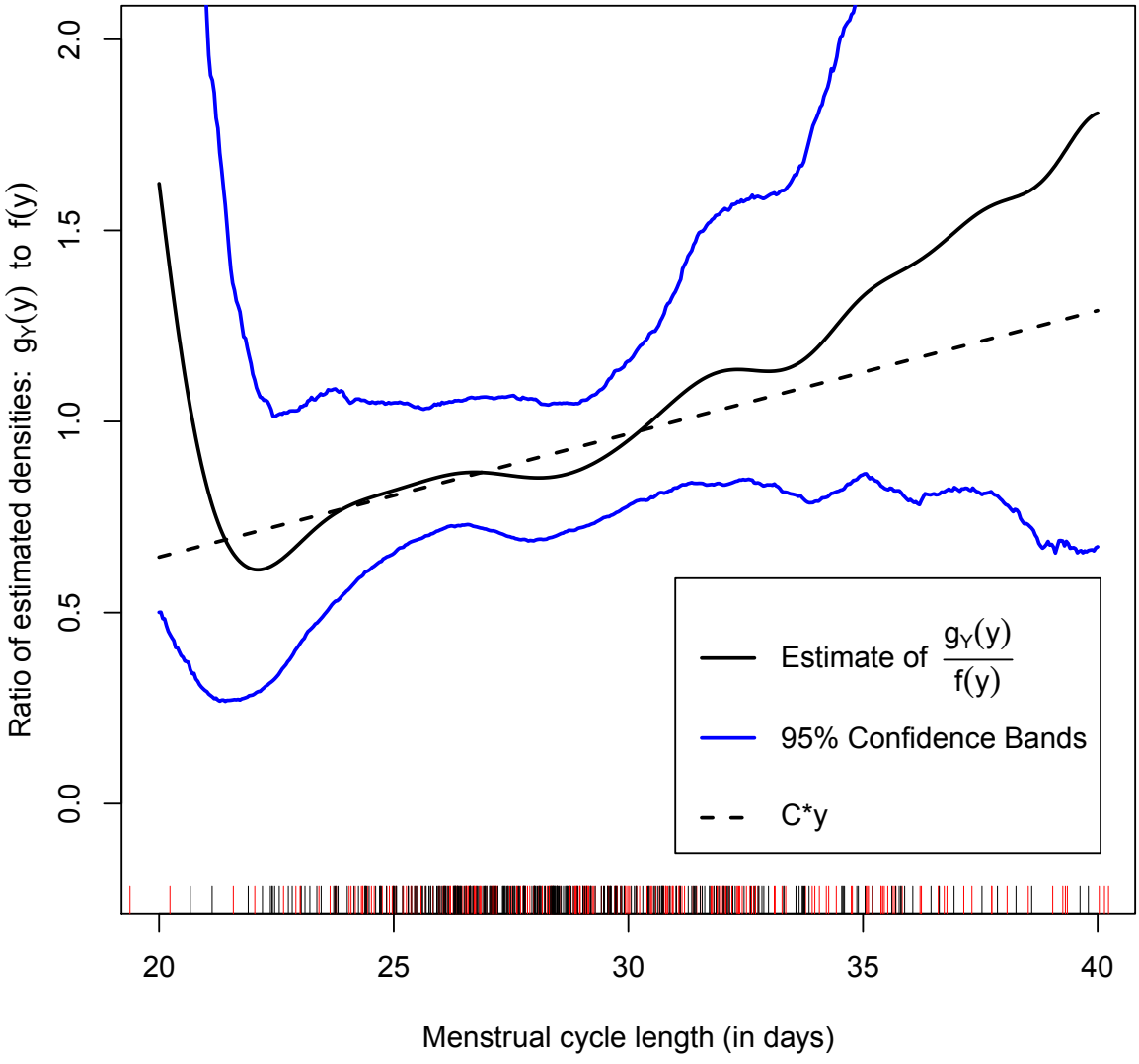
plotting an estimate of the ratio $\frac{g_Y(y)}{f(y)}$.

We focus on a subset of 304 couples from the LIFE Study with at least one cycle post-enrollment. For these couples, we use adaptive bandwidth kernel density estimation (see, e.g. Wand and Jones 1994) to estimate $g_Y(y)$ from the enrollment cycle data (without adjustment for length-bias and selection effects) and $f(y)$ from the first cycle post-enrollment; and use their ratio to estimate $\frac{g_Y(y)}{f(y)}$ for a grid of y values. In order to obtain an estimate of the uncertainty for this ratio estimator, we use the bootstrapping algorithm below to estimate 95% pointwise confidence bands. Figure 2.A.6 shows a plot of the estimated ratio curve (solid, black) and the 95% pointwise confidence bands versus y (solid, blue). A reference line Cy (dashed, black) is given. We also display a rug of the jittered data from the enrollment cycles (red) and post-enrollment cycles (black). In the range of 24-33 days, in which the uncertainty is the smallest, it seems reasonable that $\lambda(y) \propto y$.

Bootstrapping algorithm:

1. Sample n=304 ids with replacement from the couple ids.

Figure 2.A.6: Estimate of $\frac{g_y(y)}{f(y)}$ (solid, black) with 95% pointwise confidence bands (solid, blue) compared to reference line Cy (dashed, black) where C is a constant. Rug shows jittered data from the enrollment cycles (red) and post-enrollment cycles (black). Restricted to $n = 304$ couples who did not exit the study or become pregnant in the enrollment cycle and in the first post-enrollment cycle and for which these cycles are of known observed length within (8, 90) days and with backward recurrence time less than 50 days.



2. For the sampled ids, using a fixed bandwidth, estimate the density of the enrollment cycle and of the first post-enrollment cycle, and then calculate the ratio for each value of length y.

3. Repeat for a total of 1000 iterations.

4. For each value of length, y, find the 95% confidence interval. Together, these intervals form the 95% pointwise confidence bands.

2.A.4. A Comparison to Post-enrollment Cycle Length Distribution

For longitudinal studies of menstrual cycle length, we can compare the distribution of enrollment cycle data, adjusting for length-bias and selection effects, with the distribution of first post-enrollment cycle data. Assuming the sampling plan does not influence the distribution of post-enrollment cycle lengths, we expect these distributions to be similar. In this section, we make this comparison using data from the Longitudinal Investigation of Fertility and the Environment (LIFE) Study, detailed in Buck Louis et al. (2011). We consider a subset of the study population consisting of 304 couples who neither exited the study nor became pregnant in the enrollment and in the first post-enrollment cycles and for which both cycle lengths could be determined, with cycle length in the range of (8, 90) days and backward recurrence time less than 50 days. The raw mean and standard deviation of menstrual cycle length for the enrollment cycle is 30.9 days and 7.6 days, respectively, versus 29.8 days and 6.9 days for the first post-enrollment cycle.

We assume the observed length follows model (2.10) of the manuscript and model $\tilde{Y}_0$ parametrically as a normal-Gumbel mixture. For the enrollment cycles, we assume $\lambda(y) \propto y$ and use the proposed semi-parametric approach in Section 2.5.1 of the manuscript to estimate the parameters of the distribution adjusting for length-bias and selection effects. For the first post-enrollment cycles, we assume the observed data are a random sample from the distribution of $\tilde{Y}$ and use maximum likelihood methods to estimate the parameters of

the distribution. In order to demonstrate the effect of the adjustment on the enrollment cycle length distribution, we show the estimated marginal survivor curves (Figs 2.A.7-2.A.8) and parameter point estimates and 95% confidence intervals (Table 2.A.3) based on the enrollment cycle data with and without adjustment versus the marginal survivor curve and parameter point estimates and 95% confidence intervals of the first post-enrollment cycle. The adjusted enrollment marginal survivor function and marginal mean and standard deviation estimates are closer to those of the first post-enrollment cycle than the unadjusted, suggesting the adjustments have improved the estimated menstrual cycle length distribution by moving it closer to the estimated distribution of the first post-enrollment cycle, which is an estimate of the distribution of $\tilde{Y}$. We note that there is some difference between the adjusted survivor curve and that of the post-enrollment cycle for cycles approximately 20 days in length, which make up $< 5\%$ of enrollment cycle lengths. This could be due to the fact that in order to contribute to the first post-enrollment cycle, the couple must not have become pregnant in the enrollment cycle and so will tend to be less fecund than others in the enrolled population.

Figure 2.A.7: Parametric estimates of marginal survivor function of menstrual cycle length for women aged less than 30 years based on first post-enrollment cycle (dotted, blue) versus based on enrollment cycle with (solid, black) and without (dashed, red) adjustments for length-bias and selection effects. Restricted to $n = 304$ couples who did not exit the study or become pregnant in the enrollment cycle and in the first post-enrollment cycle and for which these cycles are of known observed length within (8, 90) days and with backward recurrence time less than 50 days.

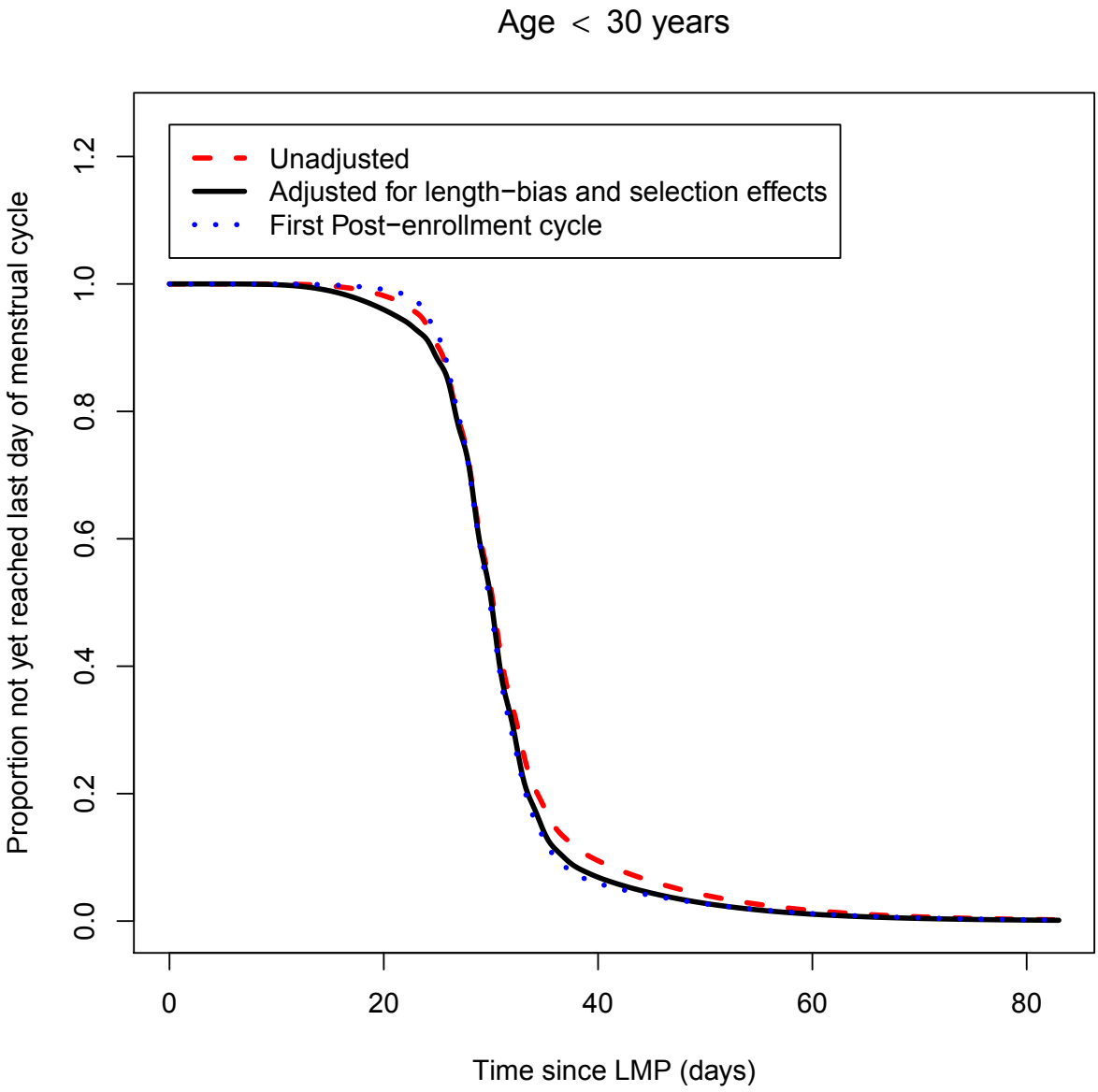


Figure 2.A.8: Parametric estimates of marginal survivor function of menstrual cycle length for women aged 30 years or older based on first post-enrollment cycle (dotted, blue) versus based on enrollment cycle with (solid, black) and without (dashed, red) adjustments for length-bias and selection effects. Restricted to $n = 304$ couples who did not exit the study or become pregnant in the enrollment cycle and in the first post-enrollment cycle and for which these cycles are of known observed length within (8, 90) days and with backward recurrence time less than 50 days.

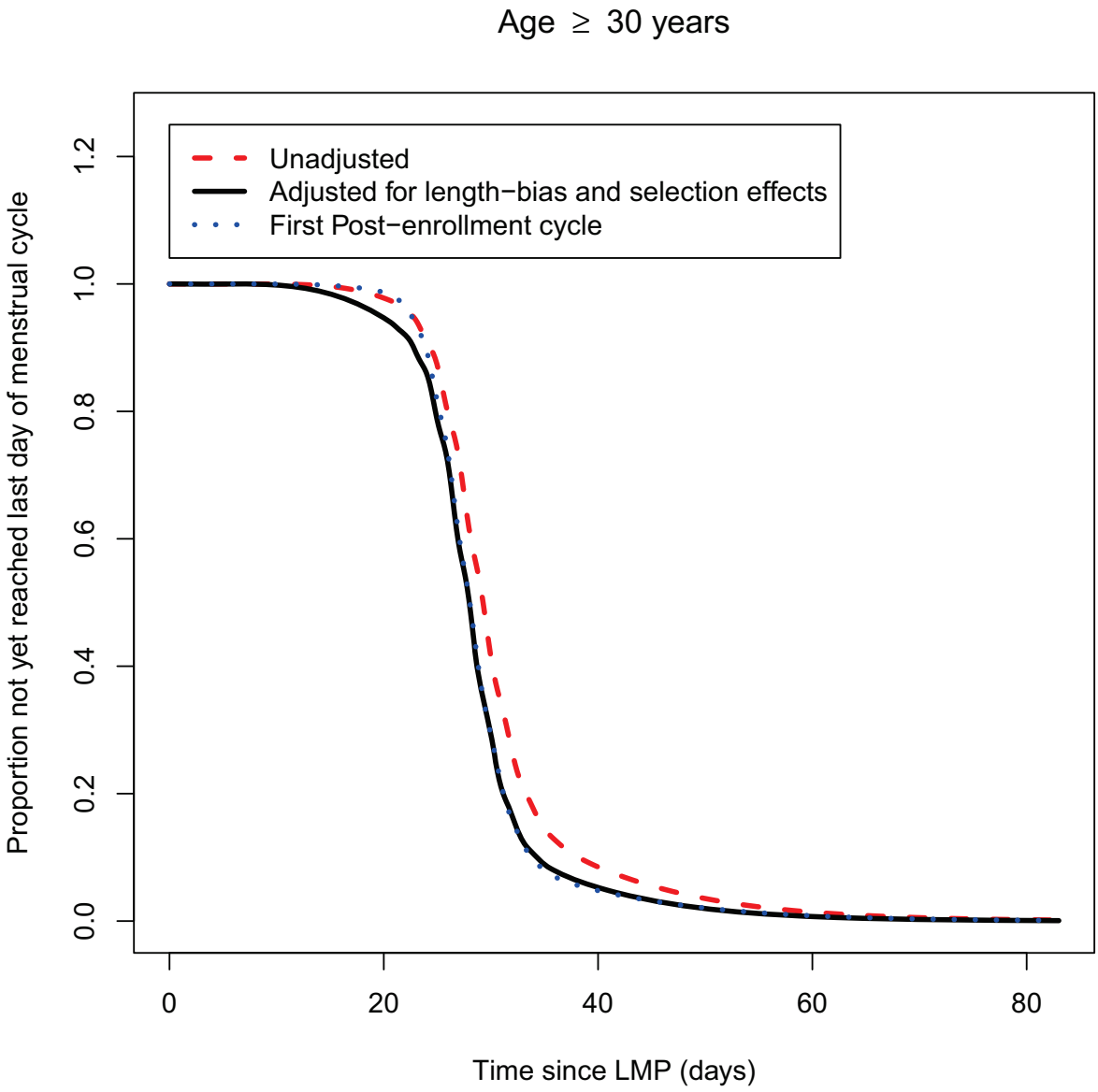


Table 2.A.3: Comparison of parameter estimates for menstrual cycle length distribution based on enrollment cycle data (unadjusted vs adjusted for length-bias (LB) and selection (S) effects) and based on first post-enrollment cycle (unadjusted). Model assumptions: $\tilde{Y} = \tilde{Y}_0 W e^{\beta z}$, $z = 0$ (age < 30), $z = 1$ (age ≥ 30), $W \sim Gamma(\alpha, 1/\alpha)$, $\tilde{Y}_0 \sim q f_{normal}(\mu_N, \sigma_N^2) + (1-q) f_{Gumbel}(\mu_G, \sigma_G^2)$. Adjustments for LB and S done using semi-parametric algorithm to estimate model parameters and $\pi(\cdot \mid z)$ assuming $\lambda(y) \propto y$. Point estimate with BC_a 95% confidence limits displayed as: ${}_{Lo}Est_{Up}$ (Louis and Zeger 2009). BC_a confidence intervals were calculated using non-parametric bootstrapping with 1000 replicates. Restricted to $n = 304$ couples who did not exit the study or become pregnant in the enrollment cycle and in the first post-enrollment cycle and for which these cycles are of known observed length within (8, 90) days and with backward recurrence time less than 50 days.

Parameter	Enrollment Cycle (Unadjusted)	Enrollment Cycle (Adjusted for LB and S)	Post-enrollment Cycle (Unadjusted)
$E(\tilde{Y} \mid age < 30)$	${}_{30.59}31.69_{32.64}$	${}_{29.56}30.72_{31.69}$	${}_{30.11}30.98_{31.85}$
$E(\tilde{Y} \mid age \geq 30)$	${}_{29.71}30.71_{32.01}$	${}_{26.54}28.00_{28.24}$	${}_{28.10}29.04_{30.17}$
$SD(\tilde{Y} \mid age < 30)$	${}_{6.89}8.25_{9.87}$	${}_{6.41}7.51_{9.02}$	${}_{5.56}7.04_{9.10}$
$SD(\tilde{Y} \mid age \geq 30)$	${}_{6.77}7.99_{9.65}$	${}_{5.53}6.85_{8.04}$	${}_{5.22}6.61_{8.56}$
$\beta : age$	${}_{-0.07}-0.03_{0.01}$	${}_{-0.12}-0.09_{-0.08}$	${}_{-0.10}-0.06_{-0.03}$
$SD(W)$	${}_{0.10}0.10_{0.11}$	${}_{0.09}0.10_{0.10}$	${}_{0.10}0.10_{0.11}$

Chapter 3

A Bayesian Approach to Joint Modeling of Menstrual Cycle Length and Fecundity

3.1. Introduction

Fecundity, or the biologic capacity of male and female for reproduction irrespective of pregnancy intentions (Buck Louis 2011), can be measured in terms of multiple endpoints such as pubertal onset and progression, menstrual cycle regularity and length, ovulation, semen quality, and pregnancy, among others. This list emphasizes the role of both the male and female partner and the complexity of factors that contribute to a couple becoming pregnant, yet few statistical models have been developed that model more than one of these endpoints. Until recently available study data on fecundity consisted mostly of retrospective pregnancy studies or prospective studies with fewer than six months follow-up, and mainly of the female partner. However, investigators of a recent prospective pregnancy study called the Longitudinal Investigation of Fertility and the Environment (LIFE) Study (Buck Louis et al. 2011) approached fecundity from a couple-based perspective, collecting measurements on exposures for both partners with the goal of assessing the effects of exposures on multiple fecundity related outcomes. We develop a joint model to investigate the relation between female menstrual cycle length and the couple's probability of pregnancy. We develop a framework that allows for the assessment of time-constant exposures on the probability of pregnancy as well

as those mediated through the female menstrual cycle length. Furthermore, we incorporate male factors which may be associated with the probability of pregnancy.

Statistical models of repeated measures of menstrual cycle length data have mainly focused on extensions of mixed effects models to account for covariates, large and heterogeneous within-female variability, and extreme cycle lengths. Harlow and Zeger (1991) categorized menstrual cycle lengths as standard and non-standard and developed separate random effects models for the mean of the standard lengths and the risk of having a non-standard length. Including only standard lengths, Lin et al. (1997) extended the linear mixed model to allow for heterogeneous within-female variability (see Laird and Ware 1982; Diggle et al. 2002; Verbeke and Molenberghs 2009; Carlin and Louis 2009, for details). Allowing for non-standard cycle lengths, Guo et al. (2006) proposed a marginal model with covariates for the population mean and variance, assuming a mixture of Gaussian and shifted Weibull error distributions.

Focusing on prediction, Bortot et al. (2010) used a state space approach to develop a predictive model of menstrual cycle length accounting for trend, autocorrelation and extremely long or short outlying lengths. McLain et al. (2012) proposed a parametric model assuming a mixture of Gaussian and Gumbel error distributions and incorporating random effects and covariates on the mean and variance parameters of the Gaussian component. Building on these work, in the menstrual cycle submodel of our joint model we propose a Bayesian hierarchical, accelerated failure time model with a mixture error distribution to allow for skewness in cycle length. In addition, we account for length-bias in modeling the length of the cycle in which the couple is enrolled and censoring of the length of the cycle in which the couple becomes pregnant.

Statistical modeling of pregnancy attempts has focused on assessing biological risk factors in the context of the couple's day-specific intercourse behavior (see Ecochard 2006, for a review). However, many models assume there exists a fixed, narrow (fertile) window of days around ovulation outside of which there is no risk for pregnancy. This restriction

may not be reasonable, as evidenced by studies which assessed the fertile window, including ovulation using the gold standard, i.e., serial vaginal ultrasound (Keulers et al. 2007). In contrast, Dominik et al. (2001) developed a method that incorporates all intercourse acts in a menstrual cycle, modeling the day-specific probability of pregnancy as a quadratic function of distance of the intercourse act from ovulation day (assumed to be day 14).

We generalize this approach to a Bayesian hierarchical model for the day-specific probabilities of pregnancy using natural, cubic splines to model the probability of pregnancy, incorporating cycle-specific information on the couple's ovulation day as measured using a fertility monitor, and using the joint model in Section 3.2 investigate the association of female menstrual cycle length with the probability of pregnancy in the context of male and female risk factors.

In addition to modeling the relation between menstrual cycle length and the probability of pregnancy, we include adjustment for the male contribution. For the LIFE Study, Buck Louis et al. (2014) found significant associations between fecundity and several semen quality parameters, when each parameter was entered into the model individually. To account for the male contribution, we incorporate male age, smoking status and multiple semen quality parameters in addition to female covariates in the model for the probability of pregnancy. Here, we focus on four World Health Organization (WHO) semen quality parameters (morphology (strict criteria), semen volume, sperm concentration, and total sperm count) which are common across many studies of semen quality (see e.g., Cooper et al. 2010).

3.2. Joint Model for Menstrual Cycle Length and Pregnancy

For the i^{th} couple ($i = 1, \ldots, n$), let j ($j = 1, \ldots, n_i$) index the female menstrual cycles with lengths Y_{ij} and ovulation dates O_{ij}. For the $k^{th}(k = 1, \ldots, Y_{ij})$ day of the j^{th} cycle, let $x_{ijk} \in \{0, 1\}$ be the observed day-specific intercourse indicator. For the j^{th} cycle, collect the intercourse indicators in the vector $\mathbf{x_{ij}} = (x_{ij1}, \ldots, x_{ijY_{ij}})$. Further, A_{ij} and δ_i are the cycle-specific and couple-specific pregnancy indicators, respectively.

3.2.1 Submodel for Longitudinal Menstrual Cycle Length

In the submodel for menstrual cycle length, we model the relation between baseline covariates (e.g., age, smoking status) and the expected menstrual cycle length, while accounting for within-female correlation in longitudinal cycle lengths and extremely short and long cycle lengths. To meet these challenges, we use a hierarchical accelerated failure time model (see, e.g., Kalbfleisch and Prentice 2011), which was adapted for modeling menstrual cycle length by Lum et al. (2014). For the j^{th} $(j = 2, \ldots, n_i)$ menstrual cycle length from the i^{th} $(i = 1, \ldots, n)$ female, we assume

$$
\begin{aligned}
&[\epsilon_{ij} \mid q, \mu_N, \sigma_N, \mu_G, \sigma_G] \\
&\sim q\text{Gaussian}(\mu_N, \sigma_N^2) + (1-q)\text{Gumbel}(\mu_G, \sigma_G^2),\ \mu_N > 0, \sigma_N > 0, \mu_G > 0, \sigma_G > 0; \\
&[W_i \mid \sigma_W] \sim \text{Gamma}(\text{shape}{=}\sigma_W^{-2}, \text{rate}{=}\sigma_W^{-2}),\ \sigma_W > 0; \\
&Y_{ij} \mid W_i, \exp(\mathbf{v_i}^\top \boldsymbol{\eta}), \epsilon_{ij} = W_i \exp(\mathbf{v_i}^\top \boldsymbol{\eta})\epsilon_{ij}. \qquad (3.1)
\end{aligned}
$$

Here, $\mathbf{v_i}$ is an r-dimensional vector of observed covariates with corresponding unknown parameter vector η and W_i is a couple-specific random effect with support of the distribution function $\in (0, \infty)$. We assume a Gaussian/Gumbel mixture distribution for the error variables to accommodate two groups of menstrual cycle lengths. We will use $T_i = 1$ to denote cycle lengths from a Gaussian distribution (group 1) and $T_i = 2$ to denote cycle lengths from a Gumbel distribution (group 2). As the mean of the error distribution is non-zero, we do not include an intercept in the fixed effects and we parameterize the random effect distribution such that $E[W] = 1$ and $V[W] = \sigma_W^2$. We assume the distribution of the covariates does not depend on $(q, \mu_N, \sigma_N^2, \mu_G, \sigma_G^2)$ and that the cycle lengths within a female are conditionally independent given $\mathbf{v_i}$ and $W_i = w_i$. Let $F_\epsilon(\epsilon; q, \mu_N, \sigma_N^2, \mu_G, \sigma_G^2)$ denote the Gaussian-Gumbel cumulative distribution function (CDF). By a transformation, the CDF of $Y_{ij}, i = 1, \ldots, n, j = 2, \ldots, n_i$ is $F(y_{ij} \mid w_i, \mathbf{v_i}; q, \mu_N, \sigma_N^2, \mu_G, \sigma_G^2, \eta) = F_\epsilon(y_{ij} w_i^{-1} \exp(-\mathbf{v_i}^\top \boldsymbol{\eta}); q, \mu_N, \sigma_N^2, \mu_G, \sigma_G^2, \eta)$ and the probability density function is

$$f(y_{ij} \mid w_i, \mathbf{v_i}; q, \mu_N, \sigma_N^2, \mu_G, \sigma_G^2, \eta) = w_i^{-1} \exp(-\mathbf{v_i}^\top \boldsymbol{\eta}) f_\epsilon(y_{ij} w_i^{-1} \exp(-\mathbf{v_i}^\top \boldsymbol{\eta}); q, \mu_N, \sigma_N^2, \mu_G, \sigma_G^2, \eta).$$

Modeling the Length of the Enrollment Cycle

In Lum et al. (2014), we proposed a method for estimating the population menstrual cycle length distribution using the enrollment cycle (i.e., the menstrual cycle during which the women enrolled), that accounts for features of the sampling plan of the LIFE Study, specifically length-bias and a selection process that is potentially a function of the time since the last menstrual period (LMP). For the LIFE study, we found that the estimated probability of enrollment was approximately constant with respect to time from LMP. Based on this finding, we account for the sampling plan by assuming a length-biased sampling distribution for Y_{i1}

$$f_{Y_1}(y_{i1} \mid w_i, \mathbf{v_i}; q, \mu_N, \sigma_N^2, \mu_G, \sigma_G^2, \eta) = \frac{y_{i1} f(y_{i1} \mid w_i, \mathbf{v_i}; q, \mu_N, \sigma_N^2, \mu_G, \sigma_G^2, \eta)}{E(Y_i \mid W_i, \exp(\mathbf{v_i}^\top \boldsymbol{\eta}), q, \mu_N, \mu_G)} \tag{3.2}$$

where $E(Y_i \mid W_i, \exp(\mathbf{v_i}^\top \boldsymbol{\eta}), q, \mu_N, \mu_G) = W_i \exp(\mathbf{v_i}^\top \boldsymbol{\eta})\{q\mu_N + (1-q)\mu_G\}$.

Modeling the Length of the Pregnancy Cycle

Menstrual cycle lengths were prospectively measured until the pregnancy cycle, during which increased levels of progesterone preclude menstrual bleeding in preparation for implantation of the blastocyst. Therefore, the length of the pregnancy cycle (i.e., the menstrual cycle during which the couple becomes pregnant) is right censored. Let τ_{in_i} be the time (in days) from the first day of the n_i^{th} cycle to the censoring day. We assume the censoring distribution is non-informative and conditionally independent of Y given $W = w$ and $\mathbf{v}$. For enrollment pregnancy cycles, the contribution to the likelihood is proportional to
$\{1 - F(\tau_{in_i} \mid w_i, \mathbf{v_i}; q, \mu_N, \sigma_N^2, \mu_G, \sigma_G^2, \eta)\} / E(Y_i \mid W_i, \exp(\mathbf{v_i}^\top \boldsymbol{\eta}), q, \mu_N, \mu_G)$; while for non-enrollment pregnancy cycles, the contribution is proportional to
$1 - F(\tau_{in_i} \mid w_i, \mathbf{v_i}; q, \mu_N, \sigma_N^2, \mu_G, \sigma_G^2, \eta)$.

Choice of Priors for Menstrual Cycle Length Submodel

Let $\boldsymbol{\varphi_Y} = (\boldsymbol{\eta}, \sigma_W, q, \mu_N, \sigma_N, \mu_G, \sigma_G)$ denote the unknown population parameters of the menstrual cycle length model. We assume the components of $\boldsymbol{\varphi}_Y$ are independent *a priori*, such that $[\boldsymbol{\varphi_Y}] = [\sigma_W][q][\mu_N][\sigma_N][\mu_G][\sigma_G] \prod_{r=1}^{R} [\eta_r]$. We complete the hierarchical model with the following specification of uniform ($\mathcal{U}$) priors:

$$
\begin{aligned}
[\eta_r \mid a_{\eta_r}, b_{\eta_r}] &\sim \mathcal{U}(a_{\eta_r}, b_{\eta_r}), r = 1, \ldots, R, \\
[\sigma_W] &\sim \mathcal{U}(0, b_{\sigma_W}), \\
[q] &\sim \mathcal{U}(0, 1), \\
[\mu_N] &\sim \mathcal{U}(a_{\mu_N}, b_{\mu_N}), \\
[\sigma_N] &\sim \mathcal{U}(0, b_{\sigma_N}), \\
[\mu_G] &\sim \mathcal{U}(a_{\mu_G}, b_{\mu_G}), \\
[\sigma_G] &\sim \mathcal{U}(0, b_{\sigma_G}),
\end{aligned}
\tag{3.3}
$$

where we choose values for the hyperparameters that are scaled to determine vague priors. In (3.3), we choose a uniform prior on the scale of the standard deviation of the female-specific parameter so as not to bias the prior away from 0 (Gelman 2006). If the posterior distribution of σ_W includes zero, for example, this would indicate no between-female variability in the mean beyond that explained by the fixed effects. We discuss alternative choices for the priors in Appendix A.3.

3.2.2 Submodel for the Probability of Pregnancy

The observed prospective pregnancy data include at the cycle level, pregnancy indicated by Clearblue®Easy pregnancy tests which detect levels of human chorionic gonadotropin over 25 mIU/mL; and within each cycle, the ovulation day and the pattern of intercourse acts.

We aim to estimate the association between the female-specific distribution of menstrual cycle length and the probability of pregnancy in a cycle, accounting for intercourse pattern in both the current and previous cycles, the preceding cycles of attempts ending without pregnancy, and male and female covariates.

For the j^{th} cycle, let k' ($k' = 1, \ldots, K'_{ij}$) index the intercourse acts in the cycle, $\nu_{ijk'}$ denote the (possibly unobserved) indicator of fertilization at the $(k')^{th}$ intercourse act, $d_{ijk'}$ denote the time difference (in days) of the intercourse day from ovulation; and let $\boldsymbol{\nu_{ij}} = (\nu_{ij1}, \ldots, \nu_{ijK'_{ij}})$ and $\mathbf{d_i} = (d_{i11}, \ldots, d_{ijK'_{ij}})$. We develop a model for the conditional probability of pregnancy in the j^{th} cycle given no pregnancy at each of the intercourse acts in previous cycles and timing of all intercourse, denoted
$Pr(A_{ij} = 1 \mid A_{i1} = \ldots = A_{i(j-1)} = 0, \boldsymbol{\nu_{i1}}, \ldots, \boldsymbol{\nu_{i(j-1)}}, K'_{ij}, \mathbf{d_{ij}}), j > 1.$

We assume intercourse cannot be successful if the ovum has already been fertilized; thus, $\nu_{ijk'}$ is dependent on $\{\nu_{i11}, \ldots, \nu_{ij(k'-1)}\}$. Let $\rho(d_{ijk'})$ be the conditional probability of fertilization at the $(k')^{th}$ intercourse given no prior fertilization. For the $(k')^{th}$ intercourse, $k' > 1$, we have

$$\text{pr}(\nu_{ijk'} = 1 \mid \nu_{i11} = \ldots = \nu_{ij(k'-1)} = 0) = \rho(d_{ijk'}),$$

$$\text{pr}(\nu_{ijk'} = 1 \mid \max(\nu_{i11}, \ldots, \nu_{ij(k'-1)}) = 1) = 0; \tag{3.4}$$

where (3.4) follows from the fact that prior fertilization prevents subsequent. For the cycles preceding the j^{th} cycle, we observe $\boldsymbol{\nu_{i1}} = \ldots = \boldsymbol{\nu_{i(j-1)}} = \mathbf{0}$. However, if pregnancy occurs in the j^{th} cycle, $\boldsymbol{\nu_{ij}}$ is unobserved for $K'_{ij} > 1$. Therefore, we express the conditional probability of pregnancy in the j^{th} cycle as the sum of the probabilities of all possible permutations of $(v_{ij1}, \ldots, v_{ijK'_{ij}})$ under the condition $\sum_{k'} v_{ijk'} = 1$. In the special case of one act of intercourse in the j^{th} cycle,

$$Pr(A_{ij} = 1 \mid A_{i1} = \ldots = A_{i(j-1)} = 0, \boldsymbol{\nu_{i1}}, \ldots, \boldsymbol{\nu_{i(j-1)}}, K'_{ij} = 1, \mathbf{d_i}) = \rho(d_{ij1}).$$

For the probability of pregnancy in the j^{th} cycle with two intercourse acts,

$$Pr(A_{ij} = 1 \mid A_{i1} = \ldots = A_{i(j-1)} = 0, \boldsymbol{\nu_{i1}}, \ldots, \boldsymbol{\nu_{i(j-1)}}, K'_{ij} = 2, \mathbf{d_i}) = \rho(d_{ij1}) + \{1 - \rho(d_{ij1})\}\rho(d_{ij2});$$

and in general, for the probability of pregnancy in the j^{th} cycle with K'_{ij} intercourse acts,

$$Pr(A_{ij} = 1 \mid A_{i1} = \ldots = A_{i(j-1)} = 0, \boldsymbol{\nu_{i1}}, \ldots, \boldsymbol{\nu_{i(j-1)}}, K'_{ij}, \mathbf{d_i}) =$$

$$\rho(d_{ij1}) + \Big\{ \sum_{k'=2}^{K'_{ij}} \Big([\prod_{l=1}^{k'-1} \{1 - \rho(d_{ijl})\}] \rho(d_{ijk'}) \Big) \Big\}^{K'_{ij} > 1}. \tag{3.5}$$

The conditional probability of no pregnancy in the j^{th} cycle with K'_{ij} intercourse acts is

$$Pr(A_{ij} = 0 \mid A_{i1} = \ldots = A_{i(j-1)} = 0; \boldsymbol{\nu_{i1}}, \ldots, \boldsymbol{\nu_{i(j-1)}}, K'_{ij}, \mathbf{d_i}) = \prod_{k'=1}^{K'_{ij}} \{1 - \rho(d_{ijk'})\}.$$

The expression in (3.5) is equivalent to

$$Pr(A_{ij} = 1 \mid A_{i1} = \ldots = A_{i(j-1)} = 0, \boldsymbol{\nu_{i1}}, \ldots, \boldsymbol{\nu_{i(j-1)}}, K'_{ij}, \mathbf{d_i}) = 1 - \prod_{k'=1}^{K'_{ij}} \{1 - \rho(d_{ijk'})\}.$$

Let k ($k = 1, \ldots, \ddot{Y}_{ij}$) index day of the menstrual cycle, where $\ddot{Y} = [Y]$ and $[\ldots]$ denotes the greatest integer function. Using x_{ijk} to indicate intercourse and $\mathbf{x_i}$ to denote the full history of intercourse acts, we equivalently model

$$Pr(A_{ij} = 1 \mid A_{i1} = \ldots = A_{i(j-1)} = 0, \mathbf{x_i}, \mathbf{d_i}) = 1 - \prod_{k=1}^{\ddot{Y}_{ij}} \{1 - \rho(d_{ijk})\}^{x_{ijk}}. \tag{3.6}$$

One decision that must be made in fitting this model is which days of the menstrual cycle to include. Various approaches have been suggested such as assuming the intercourse act closest to ovulation is the one that fertilizes the ovum (Royston and Ferreira 1999) or assuming an intercourse act has a nonzero probability of success only if it falls within a predetermined fixed window around the ovulation day (see, e.g., Weinberg et al. 1994; Dunson and Stanford 2005). We employ this second approach, but since a wide variety of fixed windows have been reported (Lynch et al. 2006), we use a very broad window, specifically excluding only intercourse acts more than 16 days before ovulation or 18 days after ovulation. Approximately 5% of all observed intercourse acts occurred outside this range. In a model of barrier

contraceptive efficacy, Dominik and Chen (2006) consider a broad window consisting of 13 days prior to and 16 days after an assumed ovulation on day 14 of the cycle.

In the second level of the hierarchical model, we model the probability of pregnancy by intercourse on the k^{th} day of the cycle as a function of the latent female-specific menstrual cycle length, denoted Y_{pi}, with adjustment for male and female risk factors and difference of day k from ovulation:

$$\text{logit}[\rho\{\mathbf{m_i}, \boldsymbol{\beta}, \mathbf{z_i}, \boldsymbol{\gamma}, \alpha_0, g(d_{ijk})\}] = \mathbf{m_i}^\top \boldsymbol{\beta} + \mathbf{z_i}^\top \boldsymbol{\gamma} + \alpha_0 + g(d_{ijk}). \tag{3.7}$$

In (3.7), $\mathbf{m_i}$ is a vector composed of linear and potentially quadratic terms of Y_{pi}, which we incorporate by mixing over the posterior predictive distribution conditional on the woman's observed menstrual cycle lengths, priors and hyperpriors. The corresponding unknown regression coefficient vector $\boldsymbol{\beta}$, links the menstrual cycle length and pregnancy submodels.

In separate models, we also consider the relation between the probability of pregnancy and the latent female-specific menstrual cycle length conditional on group 1 denoted $Y_{pi} \mid T_i = 1$, and the corresponding conditional means, $E(Y_{pi} \mid \mathbf{Y_i}, \varphi_Y, W_i, \mathbf{v_i})$ and $E(Y_{pi} \mid \mathbf{Y_i}, \varphi_Y, W_i, \mathbf{v_i}, T_i = 1)$. In Section 3.4.2, we present results from models with the

following choices for $\mathbf{m_i}^\top \boldsymbol{\beta}$:

$$\mathbf{m_i}^\top \boldsymbol{\beta} = \beta_1 Y_{pi} \tag{Lin}$$

$$\mathbf{m_i}^\top \boldsymbol{\beta} = \beta_1 Y_{pi} + \beta_2 (Y_{pi})^2 \tag{LinQuad}$$

$$\mathbf{m_i}^\top \boldsymbol{\beta} = \beta_1 (Y_{pi} \mid T_i = 1) \tag{LinT=1}$$

$$\mathbf{m_i}^\top \boldsymbol{\beta} = \beta_1 (Y_{pi} \mid T_i = 1) + \beta_2 (Y_{pi} \mid T_i = 1)^2 \tag{LinQuadT=1}$$

$$\mathbf{m_i}^\top \boldsymbol{\beta} = \beta_1 E(Y_{pi} \mid \mathbf{Y_i}, \varphi_Y, W_i, \mathbf{v_i}) \tag{LinCmean}$$

$$\mathbf{m_i}^\top \boldsymbol{\beta} = \beta_1 E(Y_{pi} \mid \mathbf{Y_i}, \varphi_Y, W_i, \mathbf{v_i}) + \beta_2 \{E(Y_{pi} \mid \mathbf{Y_i}, \varphi_Y, W_i, \mathbf{v_i})\}^2 \tag{LinQuadCmean}$$

$$\mathbf{m_i}^\top \boldsymbol{\beta} = \beta_1 E(Y_{pi} \mid \mathbf{Y_i}, \varphi_Y, W_i, \mathbf{v_i}, T_i = 1) \tag{LinCmeanT=1}$$

$$\mathbf{m_i}^\top \boldsymbol{\beta} = \beta_1 E(Y_{pi} \mid \mathbf{Y_i}, \varphi_Y, W_i, \mathbf{v_i}, T_i = 1) + \beta_2 \{E(Y_{pi} \mid \mathbf{Y_i}, \varphi_Y, W_i, \mathbf{v_i}, T_i = 1)\}^2 \tag{LinQuadCmeanT=1}$$

In (3.7) we also adjust for male and female risk factors denoted by $\mathbf{z_i}$ with corresponding parameter vector γ. Specifically, we consider four of the WHO male semen quality parameters, male smoking status and female smoking status. As male and female age are highly correlated, we incorporate the average and difference in male and female ages. Additionally, we account for the time (in days) from intercourse to ovulation, $d_{ijk} = k - O_{ij}$. We model $g(d_{ijk})$ as a smooth function estimated by $\hat{g}(\cdot) = \sum_{l=1}^{L} \alpha_l B_l(\cdot)$, where $\{B_1(\cdot), \ldots, B_L(\cdot)\}$ are the B-spline basis functions for a natural cubic spline. The intercept is represented by the parameter α_0. Previous specifications of $g(d_{ijk})$ found in the literature include a piecewise linear spline (Dominik et al. 2001), a quadratic function centered around ovulation (Dominik and Chen 2006), or a day-specific parameter with restriction of k to a small window of days around ovulation (see, e.g., Dunson and Stanford 2005).

We denote the parameters of the pregnancy model by $\boldsymbol{\varphi_A} = (\beta, \boldsymbol{\gamma}, \boldsymbol{\alpha})$. Assuming the components of $\boldsymbol{\varphi_A}$ are independent and are also independent of $\boldsymbol{\varphi_Y}$; we complete the model

structure by choosing noninformative uniform priors for each parameter in $\boldsymbol{\varphi_A}$.

Modeling the Probability of Pregnancy in the Enrollment Cycle

As we describe in Section 3.4.1, the ovulation day is estimated based on day level data collected using the Clearblue®Easy fertility monitor. Since couples were allowed to enroll on any day of the female menstrual cycle, the majority of the couples do not have sufficient monitor data throughout the enrollment cycle to estimate the ovulation day. Further, for the most fecund couples who become pregnant before the first observed bleeding event (approximately 10%), the enrollment cycle is their only cycle under observation. In order to include these couples in the analysis, we model the probability of becoming pregnant in the enrollment cycle as a function of the couple's baseline covariates, average number of intercourse acts (IntcFreq) and the female menstrual cycle length:

$$\text{logit}\{P(A_{i1}=1;\mathbf{m_i},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\lambda},\mathbf{z_i})\} = \mathbf{m_i}^\top\boldsymbol{\beta} + \mathbf{z_i}^\top\boldsymbol{\gamma} + \lambda_0 + \lambda_1\text{IntcFreq}_i + \lambda_2(\text{IntcFreq}_i)^2.$$

3.3. Estimation

We jointly model the longitudinal menstrual cycle lengths and the conditional probability of pregnancy per cycle, assuming these processes are independent conditional on the shared random effect W_i. Further, we assume cycle lengths, $(Y_{i1},\ldots,Y_{in_i})$, are independent given $(W_i,\mathbf{v_i})$. Let $\mathbf{Y_i^o}$ denote the observed menstrual cycle lengths and Y_i^u the unobserved length of the pregnancy cycle. Under these assumptions, the joint distribution of $(\mathbf{Y_i},\mathbf{A_i},Y_{pi},W_i)$ can be factored as

$$[\mathbf{Y_i},\mathbf{A_i},Y_{pi},W_i \mid \boldsymbol{\varphi_Y},\boldsymbol{\varphi_A}] = [\mathbf{A_i} \mid Y_{pi},\mathbf{z_i},\mathbf{x_i},\mathbf{d_i},\boldsymbol{\varphi_A}][Y_{pi} \mid \mathbf{Y_i},W_i,\mathbf{v_i},\boldsymbol{\varphi_Y}][\mathbf{Y_i} \mid W_i,\mathbf{v_i},\boldsymbol{\varphi_Y}][W_i \mid \boldsymbol{\varphi_Y}],$$

where

$$[\mathbf{Y_i} \mid W_i, \mathbf{v_i}, \boldsymbol{\varphi_Y}] = [\mathbf{Y_i^o}, Y_i^u \mid W_i, \mathbf{v_i}, \boldsymbol{\varphi_Y}]$$

$$= [Y_i^u \mid \mathbf{Y_i^o}, W_i, \mathbf{v_i}, \boldsymbol{\varphi_Y}]^{\delta_i} [\mathbf{Y_i^o} \mid W_i, \mathbf{v_i}, \boldsymbol{\varphi_Y}]$$

$$= [Y_i^u \mid \mathbf{Y_i^o}, W_i, \mathbf{v_i}, \boldsymbol{\varphi_Y}]^{\delta_i} [Y_{in_i}^o \mid W_i, \mathbf{v_i}, \boldsymbol{\varphi_Y}]^{(1-\delta_i)} \prod_{j=1}^{n_i-1} [Y_{ij}^o \mid W_i, \mathbf{v_i}, \boldsymbol{\varphi_Y}].$$

The unknown model parameters for couple i are: $Y_i^u, Y_{pi}, W_i, \boldsymbol{\varphi_Y}$, and $\boldsymbol{\varphi_A}$; and the observed data are $\mathbf{D} = (\mathbf{Y_i^o}, \mathbf{v_i}, \mathbf{A_i}, \mathbf{O_i}, \mathbf{z_i}, \mathbf{x_i}, \mathbf{d_i})$, where $\mathbf{x_i}$ is a vector composed of the stacked vectors of day-specific intercourse indicators. For a couple with $A_{in_i} = 0$ (i.e., no pregnancy), $\mathbf{Y_i^o} = \mathbf{Y_i}$ (i.e., the lengths of all cycles $j = 1, \ldots, n_i$ are observed) and the contribution to the joint posterior distribution is given by

$$[Y_{pi}, W_i, \boldsymbol{\varphi_Y}, \boldsymbol{\varphi_A} | \mathbf{D}; A_{in_i} = 0] \propto$$

$$\Big(\prod_{j=1}^{n_i} [Y_{ij}^o \mid W_i, \mathbf{v_i}, \boldsymbol{\varphi_Y}] [A_{ij} = 0 \mid A_{i1} = \ldots = A_{i(j-1)} = 0; Y_{pi}, \mathbf{x_i}, \mathbf{d_i}, \mathbf{z_i}, \boldsymbol{\varphi_A}] \Big)$$

$$\times [Y_{pi} \mid \mathbf{Y_i}, W_i, \mathbf{v_i}, \boldsymbol{\varphi_Y}] [W_i \mid \boldsymbol{\varphi_Y}] [\boldsymbol{\varphi_Y}] [\boldsymbol{\varphi_A}].$$

For a couple with $A_{in_i} = 1$ (i.e., pregnant), the contribution to the joint posterior distribution is given by

$$[Y_i^u, Y_{p_i}, W_i, \boldsymbol{\varphi_Y}, \boldsymbol{\varphi_A} | \mathbf{D}; A_{in_i} = 1] \propto$$

$$\Big(\prod_{j=1}^{n_i-1} [Y_{ij}^o \mid W_i, \mathbf{v_i}, \boldsymbol{\varphi_Y}] [A_{ij} = 0 \mid A_{i1} = \ldots = A_{i(j-1)} = 0; Y_{pi}, \mathbf{x_i}, \mathbf{d_i}, \mathbf{z_i}, \boldsymbol{\varphi_A}] \Big)^{\mathbb{1}\{n_i > 1\}}$$

$$\times [Y_i^u \mid \mathbf{Y_i^o}, W_i, \mathbf{v_i}, \boldsymbol{\varphi_Y}] [A_{in_i} = 1 \mid A_{i1} = \ldots = A_{i(n_i-1)} = 0; Y_{pi}, \mathbf{x_i}, \mathbf{d_i}, \mathbf{z_i}, \boldsymbol{\varphi_A}]$$

$$\times [Y_{pi} \mid \mathbf{Y_i}, W_i, \mathbf{v_i}, \boldsymbol{\varphi_Y}] [W_i \mid \boldsymbol{\varphi_Y}] [\boldsymbol{\varphi_Y}] [\boldsymbol{\varphi_A}].$$

We estimate the marginal posterior distributions of the parameters using Markov chain Monte Carlo (MCMC) integration. We describe the details of implementing our approach in Appendix A.3.

3.4. Application to the LIFE Study, a Prospective Pregnancy Study

The aim of our analysis is to investigate the relation between female menstrual cycle length and couple fecundity, while accounting for both female and male risk factors in keeping with the couple dependent nature of human reproduction. Specifically, we use the joint model detailed in Section 3.2 to model the female-specific menstrual cycle length distribution jointly with the couple's fecundity, as measured by the probability of pregnancy in a menstrual cycle. We assess the association between menstrual cycle length and the probability of pregnancy adjusting for male and female smoking status, average of male and female age, difference between male and female age, and four WHO semen quality parameters, specifically semen volume, sperm concentration, total sperm count, and morphology (strict criteria) as discussed in Cooper et al. (2010). While the focus of Cooper et al. (2010) is defining cutpoints for the semen quality parameters, we choose to use the continuous measurements. Furthermore, since total sperm count is the product of semen volume and sperm concentration, we fit two separate models adjusting for either total sperm count or both semen volume and sperm concentration.

3.4.1 The LIFE Study

We apply our joint modeling approach to the LIFE Study, a couple-based, prospective pregnancy study of 501 couples. Couples were accepted if they fulfilled the following inclusion criteria: married or in a committed relationship, females aged 18-40 and males over age 18 years, English or Spanish speaking, self-reported menstrual cycle lengths within 21-42 days, and no hormonal birth control injections in the past 12 months. Couples were followed until pregnancy, exit from the study or one year attempting pregnancy. Pregnancy was recognized using Clearblue®Easy pregnancy tests administered at home on the day menstruation is expected. For the length of the pregnancy attempt, both the male and female kept independent journals of daily intercourse frequency. If the intercourse question was not answered on a particular day in the female daily journal, we used the number of acts recorded in the

male daily journal. If the intercourse question was left blank by both members of the couple (approximately 2.9% of days), we assume the couple did not have intercourse on that day. Women recorded daily bleeding observed on a scale of 0 (none) to 4 (heavy). The beginning of the menstrual cycle is defined by menstruation, designated as the first day of bleeding followed within one day by at least two additional days bleeding. The length of the menstrual cycle is then defined as time (in days) from the first day of menstrual bleeding to the day preceding menstrual bleeding of the next cycle. Measurements of semen quality parameters were made at the National Institute for Occupational Safety and Health's andrology laboratory and a contract laboratory (Fertility Solutions) on samples collected at home via masturbation without the use of any lubricant following 2 days of abstinence. Morphology was assessed using the strict and WHO normal criteria (World Health Organization 1992; Rothmann et al. 2013). As both criteria measure morphology, we choose to use the strict criteria. Additional details on the collection of semen quality measurements can be found in Buck Louis et al. (2014).

Beginning on day six of the menstrual cycle, women used the Clearblue®Easy fertility monitor to measure daily levels of oestrone-$_3$-glucuronide, a metabolite of oestradiol, and luteinizing hormone (LH). The ovulation day was identified as the peak day detected using the Clearblue®Easy fertility monitor. In the case of two consecutive peak days, the latter of the two was designated as the ovulation day. If a peak was not detected and the female tested on at least 90% of the 20 testing days, we assumed the cycle was anovulatory (151 cycles) and therefore had zero risk for pregnancy; otherwise, we imputed the mean of the couple's observed ovulation days (250 cycles). If ovulation was not observed in any cycles, we imputed the ovulation day as the day with the highest LH peak (33 cycles). We excluded 20 cycles for which the ovulation day could not be detected or imputed due to non-compliance in testing.

For this analysis, we consider a subset of the LIFE Study restricted to 436 couples with complete data on the four WHO semen quality parameters so that we may investigate both

male and female risk factors for pregnancy. Couples who exited the study are censored on the last day of the cycle preceding their exit. We further excluded 10 couples who exited in the enrollment cycle and thus are missing data on intercourse days, menstrual cycle length, and pregnancy success, bringing the size of the subset to 426 couples (1934 menstrual cycles). Lastly, we excluded 23 cycles shorter than 9 or longer than 89 days. As the upper limit is more than twice the study's inclusion criteria of 42 days, lengths outside this range are likely an artifact of a missing cycle stop/start.

Details on the implementation of the proposed joint modeling approach are described in Appendix A.3 along with OpenBUGS (Thomas et al. 2006) model code in Appendix A.4.

3.4.2 Analysis of the LIFE Study

We first describe the relation between the female-specific mean menstrual cycle length and baseline age and smoking status. Table 3.1 shows the posterior medians and 95% equal tail credible intervals (CI) of the parameters of the regression model for mean menstrual cycle length (3.1). For the age range of the LIFE Study (18 to 40 years), we found that the female-specific mean cycle length decreased with increasing age (see Figure 3.1), which is in agreement with the findings by Harlow et al. (2000) in a study of menstrual cycle characteristics by age. We also included a quadratic age term, though this was not significant. For active smoking status, defined as cotinine level at least 10ng/mL (Pirkle et al. 1996), the posterior median of the coefficient was negative; however, the percentage of female active smokers was small and the posterior distribution included zero. In the model for the length of the enrollment cycle (3.2) we accounted for length-bias, which can be visualized by the right shift in the histogram of the female-specific mean of the enrollment cycle lengths compared to that of the post-enrollment cycles lengths (see Figure 3.2). Lastly, the parameters of the Gaussian-Gumbel mixture distribution indicate that the probability of a cycle from group 1 is approximately 79% (CI: (75%, 82%)), and that the mean and standard deviation

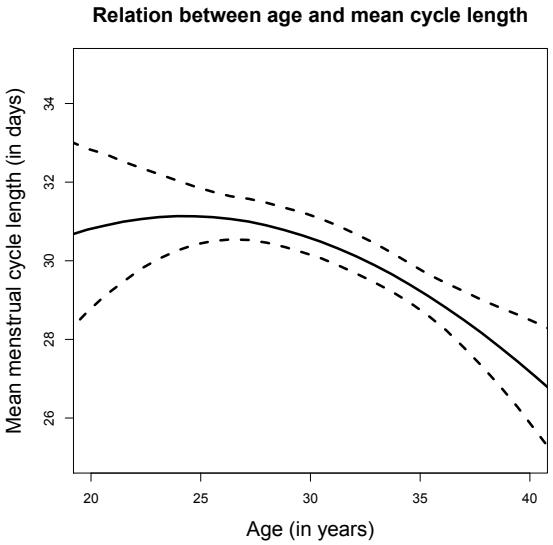


Figure 3.1: Median (solid) and 95% equal tail credible interval (dashed) for the posterior distribution of mean menstrual cycle length (in days) versus female age (in years).

of the Gaussian distribution are much smaller than that of the Gumbel distribution. In Section 3.A.1, we assess the fit of the menstrual cycle length submodel.

For the remainder of this section, we describe the results of the fecundity portion of the joint model. For each couple, we estimated the probability of pregnancy in a cycle conditional on no pregnancy in previous cycles of attempts, intercourse history, and couple covariates. Figure 3.3 shows boxplots of the conditional probability of pregnancy by cycle number. The conditional probability of pregnancy diminishes as the number of cycles without pregnancy increases.

Table 3.2 shows the posterior medians and 95% CIs of the parameters of the regression model for the day-specific probability of pregnancy on menstrual cycle length, without adjustment for other risk factors. We first fit a linear term for menstrual cycle length (model Lin) and then added a quadratic term (model LinQuad) which we found to be significant. For intercourse on a particular day of the menstrual cycle, we estimate the unconditional probability of pregnancy as a function of menstrual cycle length and find the optimal cycle length.

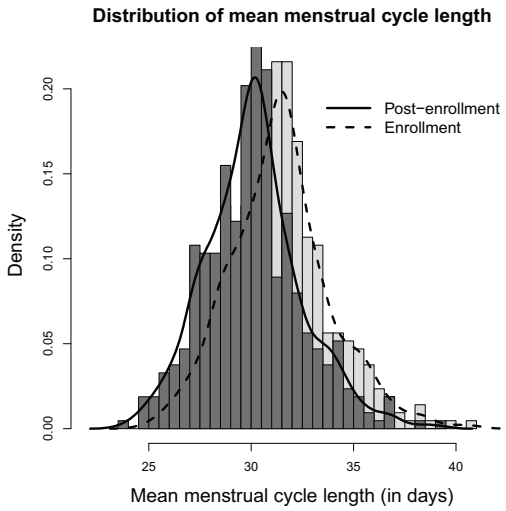


Figure 3.2: Histograms and density estimates of the estimated woman-specific mean menstrual cycle length (in days) for enrollment cycles (light gray bars, dashed line) compared to that of post-enrollment cycles (dark gray bars, solid line). Estimates shown are medians of the posterior distribution of woman-specific mean cycle length.

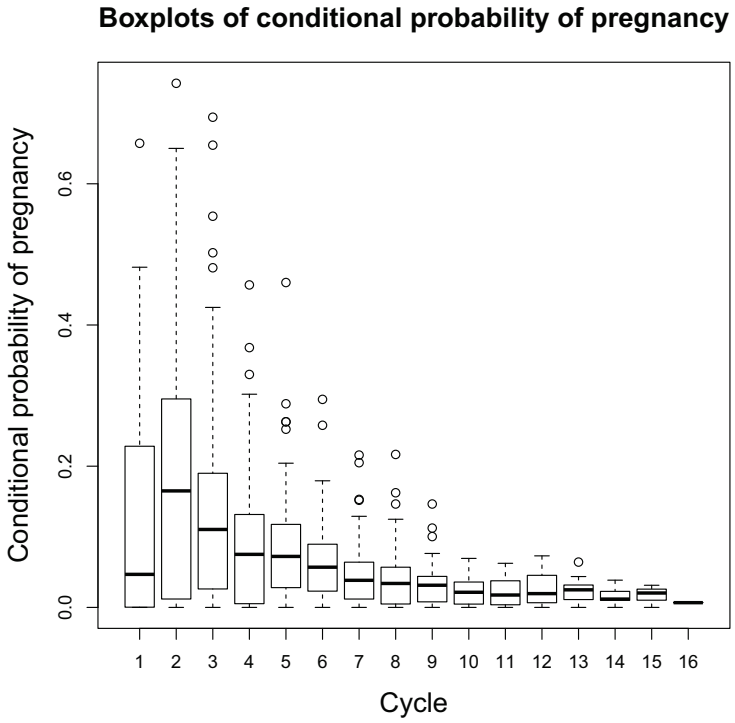


Figure 3.3: Box plots of conditional probability of pregnancy in a cycle given no pregnancy in previous cycles, intercourse history, menstrual cycle length, total sperm count, sperm morphology (strict criterion), mean of male and female age, difference between male and female age, male and female active smoking status (yes/no).

Table 3.1: Posterior median and 95% equal-tail credible intervals for menstrual cycle length parameters of joint model displayed as: $_{\text{Lo}}\text{Median}_{\text{Up}}$ (Louis and Zeger 2009). Age is standardized and active smoking status (yes/no) is defined as baseline cotinine level $\geq$ 10ng/mL. Restricted to 426 couples with data on semen quality who did not exit in the enrollment cycle.

	Parameter	$_{\text{Lo}}\text{Median}_{\text{Up}}$
η_1	Age	$_{-0.04}-0.03_{-0.01}$
η_2	Age (squared)	$_{-0.02}-0.01_{0.00}$
η_3	Active smoking (yes/no)	$_{-0.07}-0.02_{0.03}$
q	Probability of cycle from group 1	$_{0.75}0.79_{0.82}$
μ_N	Mean, group 1	$_{28.90}29.33_{29.85}$
σ_N	Standard deviation, group 1	$_{1.79}1.94_{2.11}$
μ_G	Mean, group 2	$_{33.84}35.18_{37.02}$
σ_G	Standard deviation, group 2	$_{11.06}12.34_{13.60}$
σ_W	Standard deviation, random effect	$_{0.08}0.09_{0.10}$

As an example, for a single intercourse act on the day before ovulation, we see that the optimal cycle length is approximately 31.8 days (see Figure 3.4a). In a separate model, we considered the relation between fecundity and menstrual cycle length conditional on cycles from group 1 (models LinT=1 and LinQuadT=1). For a single intercourse act on the day before ovulation, we found a similar concave relation with an optimal cycle length of 29.8 days (see Figure 3.4b). Lastly, we fit separate models to assess the relation between conditional mean menstrual cycle length and day-specific probability of pregnancy (models LinCmean-LinQuadCmeanT=1). For these models, the coefficient on the quadratic term was much smaller and was not significant (see bottom half of Table 3.2). It is also of considerable interest to assess the role of the conditional variance of menstrual cycle length in fecundity; however, a model containing both the mean and variance of cycle length did not converge due to high correlation between the mean and variance.

To further visualize the relation between menstrual cycle length and fecundity, we plot the day specific probability of pregnancy for the 25th, 50th, and 75th percentiles of menstrual

Table 3.2: Posterior median and 95% equal-tail credible intervals (CI) for association between (mean) menstrual cycle length and probability of pregnancy (see models Lin-LinQuadCmeanT=1) in Section 3.2.2. Posterior summaries displayed as: $_{\text{Lo}}\text{Median}_{\text{Up}}$ (Louis and Zeger 2009). Restricted to 426 couples with data on semen quality who did not exit in the enrollment cycle.

Model	Parameter	$_{\text{Lo}}\text{Median}_{\text{Up}}$
Lin	Menstrual cycle length	$_{0.17}0.34_{0.55}$
LinQuad	Menstrual cycle length	$_{0.14}0.46_{0.77}$
	Menstrual cycle length (squared)	$_{-0.62}-0.35_{-0.17}$
LinT=1	Menstrual cycle length, group 1	$_{0.09}0.25_{0.43}$
LinQuadT=1	Menstrual cycle length, group 1	$_{0.05}0.40_{0.72}$
	Menstrual cycle length, group 1 (squared)	$_{-0.70}-0.45_{-0.16}$
LinCmean	Conditional mean cycle length	$_{0.05}0.21_{0.38}$
LinQuadCmean	Conditional mean cycle length	$_{0.04}0.25_{0.45}$
	Conditional mean cycle length (squared)	$_{-0.28}-0.10_{0.05}$
LinCmeanT=1	Conditional mean cycle length, group 1	$_{0.06}0.21_{0.38}$
LinQuadCmeanT=1	Conditional mean cycle length, group 1	$_{0.02}0.23_{0.44}$
	Conditional mean cycle length, group 1 (squared)	$_{-0.27}-0.10_{0.03}$

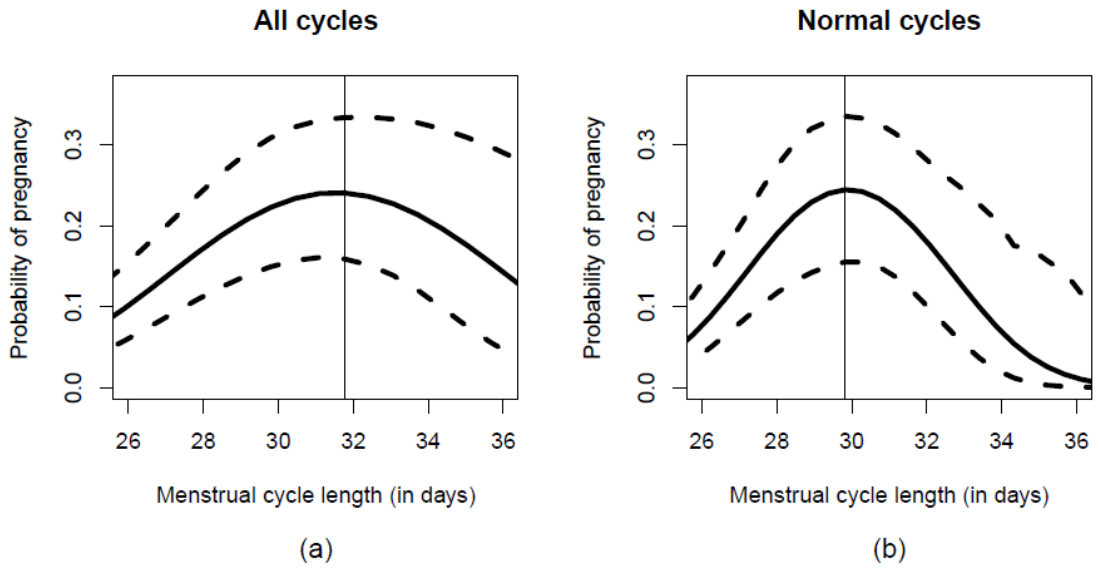


Figure 3.4: Median (solid) and 95% equal tail credible interval (dashed) for the unconditional probability of pregnancy due to intercourse on the day before ovulation versus menstrual cycle length (in days) for (a) all cycles (b) conditional on group 1.

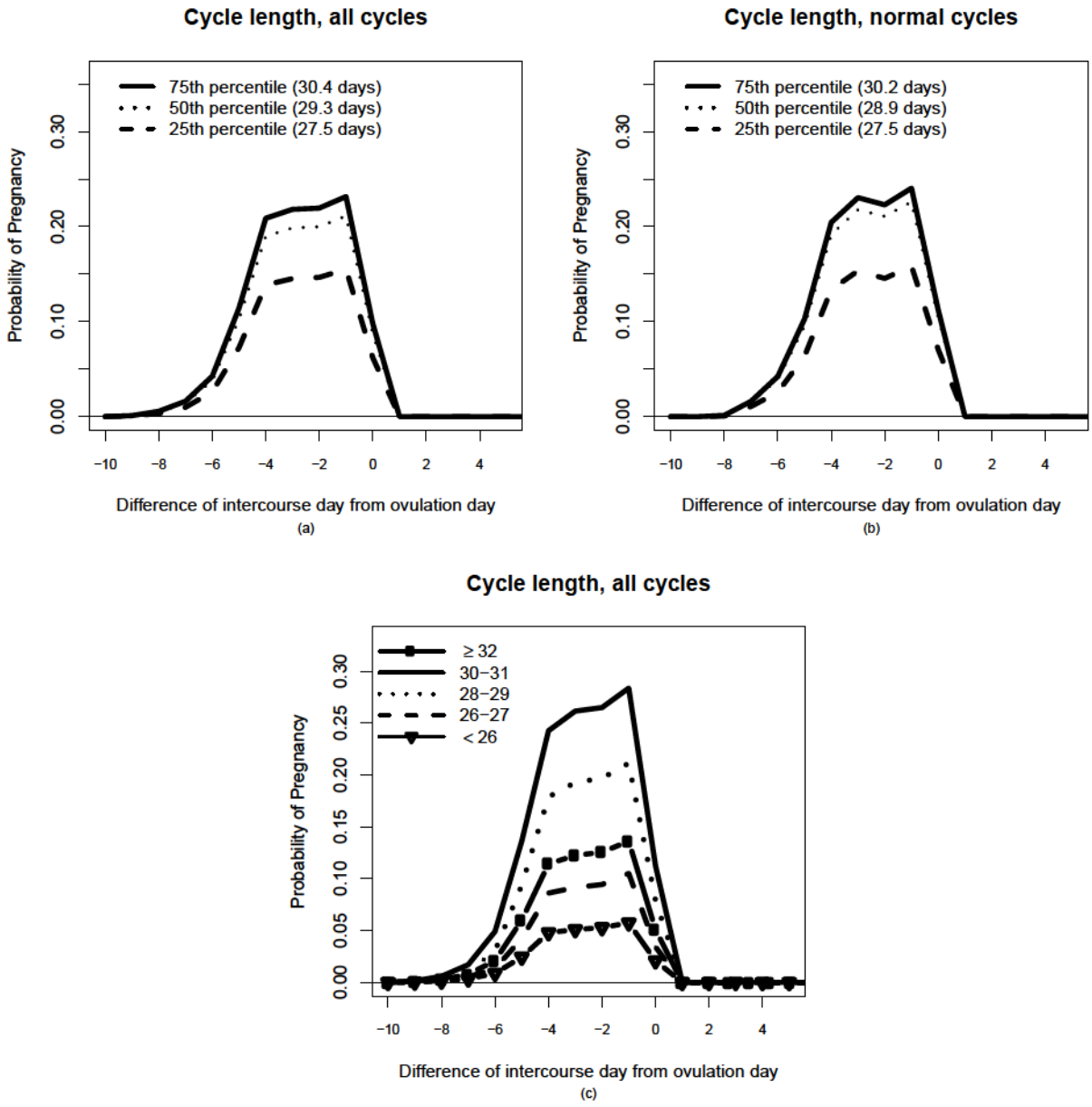


Figure 3.5: Percentiles of menstrual cycle length and unconditional probability of pregnancy due to a single act of intercourse by difference of intercourse day from ovulation day for (a) all cycles (b) conditional on group 1 (c) all cycles with cycle length (in days) categorized as in Small et al. (2006)

cycle length for all cycles and conditional on group 1 (see Figure 3.5a-b). The day specific probability of pregnancy is the unconditional probability of pregnancy due to intercourse on a single day, and the values shown in Figure 3.5 are the medians of the posterior distributions. The x-axis shows the difference in time between the intercourse day and the ovulation day. In both plots, the profile for probability of pregnancy is non-zero starting approximately seven days before ovulation, peaks the day prior to ovulation and then drops sharply to zero one day post ovulation. The probability of pregnancy is lower for the women in the 25th percentile of menstrual cycle length.

Our finding of peak probability of pregnancy corresponding to a menstrual cycle length of approximately 31.8 days agrees with the finding of Small et al. (2006) using data from the Mount Sinai Study of Women Office Workers. In this study of both pregnancy planners and nonplanners, menstrual cycle length was observed using daily diaries and then categorized into 5 categories. To compare our results, we fit a separate joint model in which menstrual cycle length is incorporated in the fecundity model as a categorical variable using cutpoints as in Small et al. (2006). As shown in Figure 3.5c, the group with menstrual cycle length of 30-31 days has the highest day specific probability of pregnancy. This plot also illustrates the difficulty in choosing a priori cutpoints of cycle length and then using these cutpoints to transform menstrual cycle length into a categorical covariate as is frequently done in studies of menstrual cycle length and pregnancy.

To assess the association between menstrual cycle length and the probability of pregnancy in the context of both male and female risk factors, we incorporated covariates in the fecundity model for sperm morphology (strict criteria), total sperm count, semen volume, sperm concentration, mean of male and female age, difference between male and female age, male active smoking status (yes/no), and female active smoking status (yes/no). We first fit separate models for each covariate to determine each one's unadjusted estimated association with the probability of pregnancy. As shown in the 'Unadjusted' column in Table 3.3, each of the semen quality parameters had a positive association with pregnancy, while the mean

Table 3.3: Posterior median and 95% equal-tail CI for pregnancy parameters of joint model displayed as: $_{\text{Lo}}\text{Median}_{\text{Up}}$ (Louis and Zeger 2009). Adjusted models A and B correspond to two combinations of uncorrelated semen quality parameters. Active smoking status (yes/no) is defined as baseline cotinine level $\geq$ 10ng/mL; and all other risk factors are standardized. Restricted to 426 couples with data on semen quality who did not exit in the enrollment cycle.

Parameters	Unadjusted	Model A	Model B
Menstrual cycle length	$_{0.14}0.46_{0.77}$	$_{-0.06}0.26_{0.59}$	$_{-0.05}0.27_{0.58}$
Menstrual cycle length (squared)	$_{-0.62}-0.35_{-0.17}$	$_{-0.52}-0.27_{-0.11}$	$_{-0.52}-0.27_{-0.12}$
Morphology (strict criteria)	$_{0.16}0.28_{0.40}$	$_{0.06}0.23_{0.41}$	$_{0.05}0.23_{0.40}$
Total sperm count (log)	$_{0.00}0.12_{0.25}$	$_{-0.09}0.08_{0.25}$	
Semen volume (log)	$_{-0.07}0.05_{0.17}$		$_{-0.13}0.03_{0.20}$
Sperm concentration (log)	$_{-0.02}0.11_{0.23}$		$_{-0.09}0.09_{0.27}$
Male active smoking (yes/no)	$_{-0.76}-0.38_{-0.02}$	$_{-0.86}-0.37_{0.13}$	$_{-0.87}-0.37_{0.13}$
Female active smoking (yes/no)	$_{-1.10}-0.37_{0.24}$	$_{-1.01}-0.13_{0.74}$	$_{-1.05}-0.15_{0.73}$
Male and female age (mean)	$_{-0.36}-0.24_{-0.11}$	$_{-0.34}-0.16_{0.02}$	$_{-0.35}-0.17_{0.01}$
Male and female age (difference)	$_{-0.12}0.01_{0.14}$	$_{-0.19}-0.02_{0.14}$	$_{-0.19}-0.03_{0.14}$
Intercourse frequency	$_{8.00}9.36_{9.94}$	$_{8.08}9.46_{9.96}$	$_{7.95}9.35_{9.96}$
Intercourse frequency (squared)	$_{-0.87}-0.79_{-0.66}$	$_{-0.87}-0.80_{-0.67}$	$_{-0.87}-0.79_{-0.66}$

couple age and smoking covariates had a negative association. Next, we fit the joint model adjusting for multiple covariates simultaneously in the fecundity model. Since total sperm count is the product of semen volume and sperm concentration, we fit two separate models (A and B), where model A adjusts for total sperm count and model B adjusts for semen volume and sperm concentration. In both adjusted models, the concave relation between female-specific menstrual cycle length and day-specific probability of pregnancy remained significant (see Table 3.3). For each of the semen quality parameters, we found a positive association with the probability of pregnancy though the coefficient was only significant for sperm morphology (strict criteria). Figure 3.6 illustrates the positive relation between sperm morphology and the day specific probability of pregnancy using the 25th, 50th, and 75th percentiles of morphology. We also found that the probability of pregnancy decreased with male and female smoking status and with mean age and difference in ages; however, none

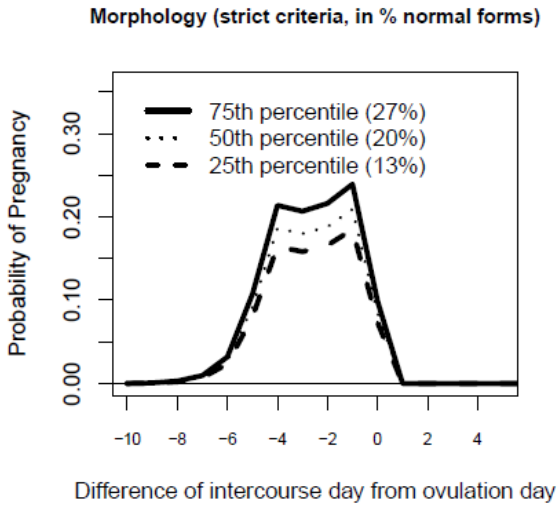


Figure 3.6: Percentiles of morphology (strict criteria, in % normal forms) and unconditional probability of pregnancy due to a single act of intercourse by difference of intercourse day from ovulation day. Estimates are adjusted for cycle length, total sperm count, mean of male and female age, difference between male and female age, female active smoking status (yes/no), and male active smoking status (yes/no).

of these were significant in the adjusted models. Lastly, for the enrollment cycles, we found a concave relation between probability of pregnancy and average intercourse frequency (see Figure 3.7) with a peak at 6 days of intercourse, adjusted for cycle length, total sperm count, mean of male and female age, difference between male and female age, female smoking status and male smoking status.

3.5. Discussion

We employed a Bayesian hierarchical modeling approach for assessing the relation between female menstrual cycle length and couple fecundity, while accounting for both male and female risk factors for pregnancy. In the menstrual cycle length submodel, we addressed several challenges including length-biased sampling of enrollment cycles, unobserved length of pregnancy cycles, and skewness of the underlying distribution. In the fecundity submodel, we proposed a model for the probability of pregnancy conditional on no pregnancy in previous cycles of attempts. We adjusted for the couples intercourse behavior in a cycle using a flexible

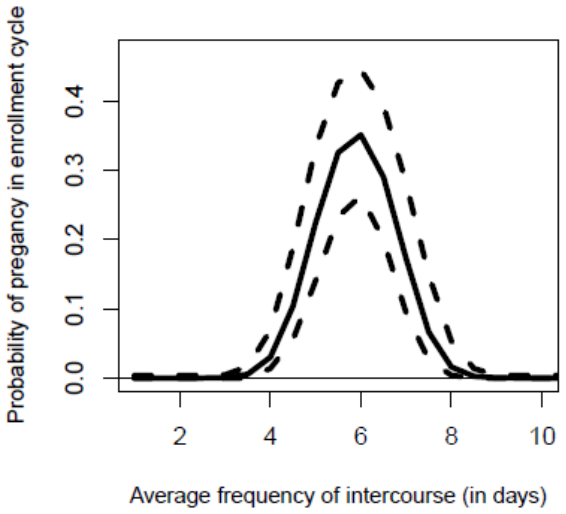


Figure 3.7: Median (solid) and 95% equal tail credible interval (dashed) for the probability of pregnancy in the enrollment cycle versus average number of days of intercourse. Estimates are adjusted for menstrual cycle length, total sperm count, sperm morphology (strict criteria), mean of male and female age, difference between male and female age, female active smoking status (yes/no), and male active smoking status (yes/no).

spline that is a function of the difference of day of intercourse from the ovulation day and gives weight to a broad window of days of intercourse in a cycle. We also incorporated a model for the probability of pregnancy in the enrollment cycle as a function of the average frequency of intercourse.

A key aspect of our approach is the inclusion of regression models in both the mean cycle length and day-specific probability of pregnancy models. For the LIFE Study population, we found that mean cycle length decreases with increasing age and that there is a concave relation between menstrual cycle length and the couple's probability of pregnancy, with an optimal cycle length of about 31.8 days. These findings are consistent with others in the literature; therefore, we have developed the framework for assessing the relation between exposures and both menstrual cycle length and fecundity. In Chapter 4, we build upon this model by incorporating baseline exposure variables to investigate potential associations between environmental chemicals, menstrual cycle length, and fecundity.

Finally an extension of this work that focuses on the fecundity model is to incorporate an

interaction between the spline function and menstrual cycle length. The trend in modeling fecundity has been to allow for flexibility in the length of the fertile window, i.e., the window of days around ovulation for which the probability of pregnancy is nonzero. In the proposed model, we estimated the day-specific probability of pregnancy for a broad window of days in the female menstrual cycle and found that the probability of pregnancy consistently declines to zero just one day post-ovulation; while pre-ovulation, we found more variation in the number of days with a non-zero probability of pregnancyacross female menstrual cycle length. Currently, we have incorporated menstrual cycle length in the fecundity model as a link-additive term, which allows the profile of day-specific probabilities to shift up or down. By incorporating an interaction with the spline function, we would allow for changes in the shape of the profile of probabilities.

Appendices

3.A.1. Fit of Menstrual Cycle Length Submodel

We examine the fit of the menstrual cycle length submodel (3.1) by comparing the predicted cycle length with that observed in the first cycle post-enrollment. We consider a subset of 284 couples with at least one post-enrollment cycle. In Figure 3.A.1, we plot the observed length and the median and 95% equal tail credible interval of the posterior predictive distribution for every fourth couple where the y-axis is the couple index and the couples are sorted by observed length. Overall, the observed cycle length is predicted well by the posterior median. In Figure 3.A.2, we focus on the smallest, middle, and largest 10% of observed cycle lengths. The observed lengths in the middle 10% are well predicted. The smallest observed lengths are slightly overestimated by the median but even the smallest still fall within the posterior predictive distribution. The largest 10% observed lengths are underestimated by the median and approximately 6 fall outside the posterior predictive distribution. We also fit the model assuming a log-normal distribution for the error variables and found a similar pattern of results (see Figures 3.A.3 and 3.A.4); however, the predictions were better and the deviance information criterion was smaller for the model assuming a Gaussian/Gumbel mixture.

We also assessed the fit of both menstrual cycle length models by estimating the posterior distribution of the standardized residuals, denoted r_i

$$r_i = \frac{y_i - E(Y_{pi} \mid \mathbf{Y_i})}{\sqrt{V(Y_{pi} \mid \mathbf{Y_i})}},$$

where y_i is the observed length of the first cycle post-enrollment. Figure 3.A.5 shows plots of the residuals versus the predicted lengths. For both models, the majority of the residuals are between -2 and 2 with some outlying large values. For the Gaussian-Gumbel mixture model, the residuals are smaller, indicating a better fit.

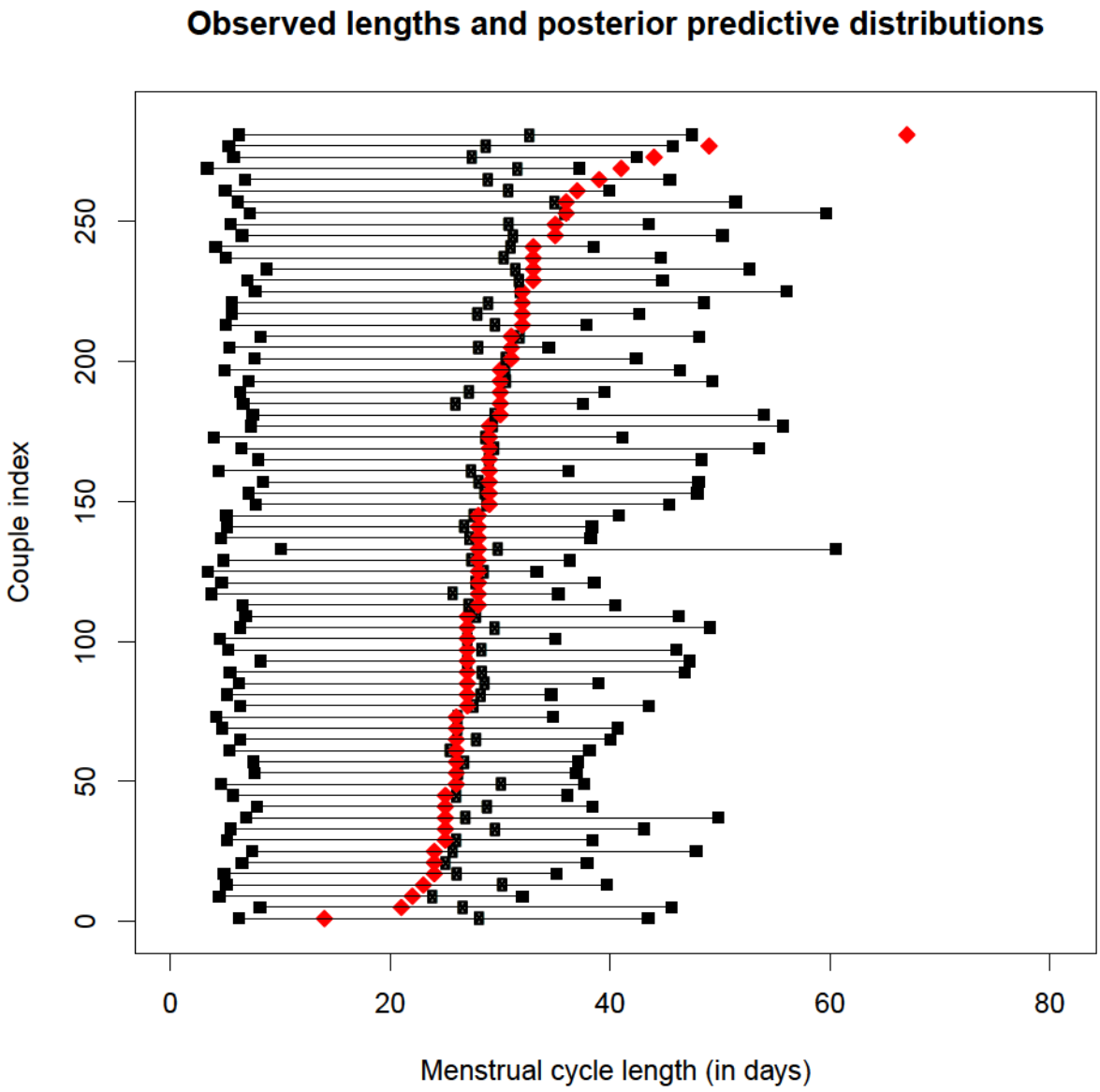


Figure 3.A.1: Posterior medians (black, circle) and 95% equal-tail credible intervals (black, squares) for predicted menstrual cycle length compared with observed length of first cycle post-enrollment (red, diamond). The predictive posterior distribution is estimated using the proposed menstrual cycle length model assuming a Gaussian/Gumbel mixture distribution for the error variables. Restricted to 284 couples with at least one observed cycle length post-enrollment. Couples are sorted by observed cycle length and results are shown for every fourth couple.

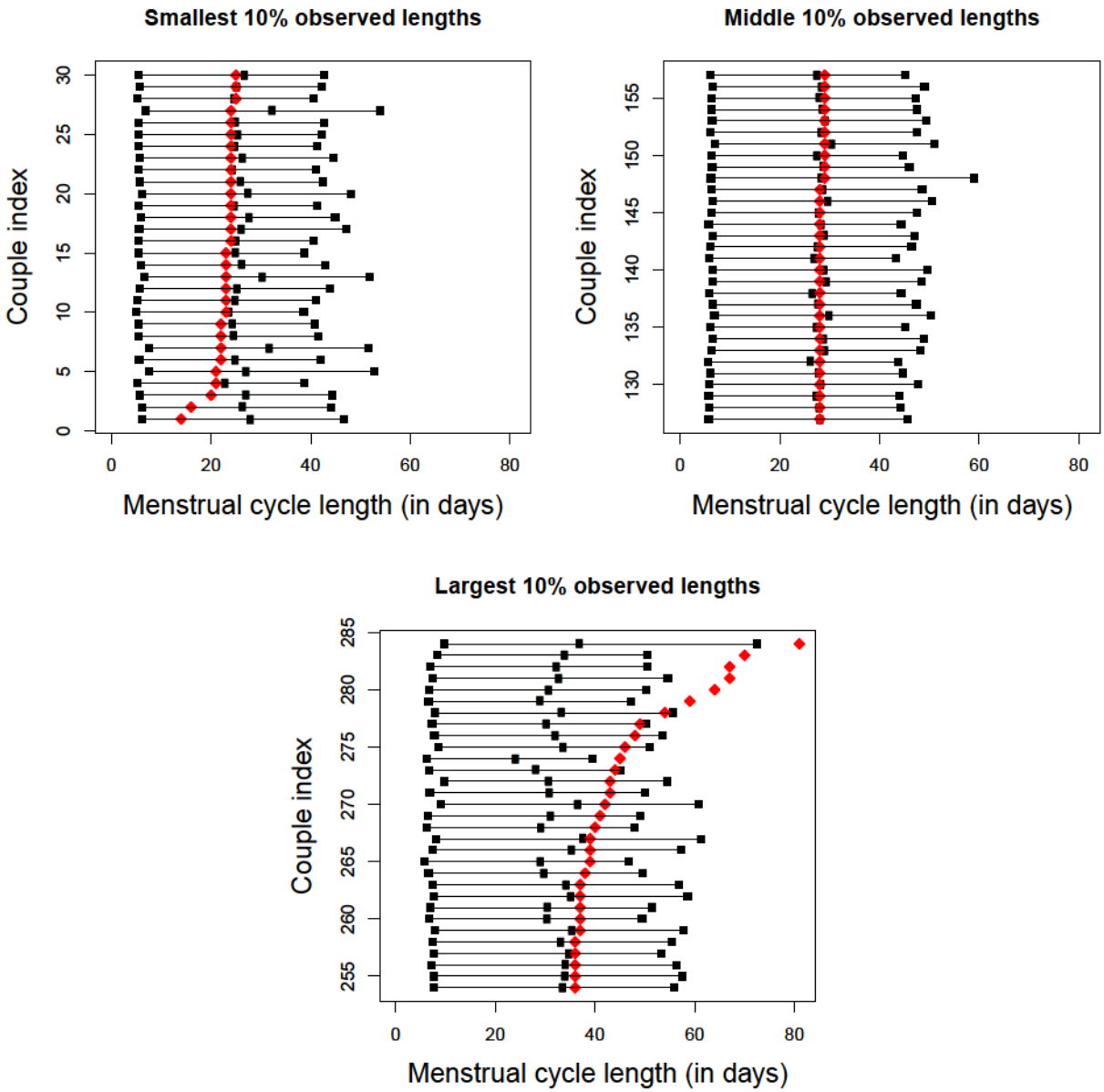


Figure 3.A.2: Posterior medians (black, circle) and 95% equal-tail credible intervals (black, squares) for predicted menstrual cycle length compared with observed length of first cycle post-enrollment (red, diamond). The predictive posterior distribution is estimated using the proposed menstrual cycle length model assuming a Gaussian/Gumbel mixture distribution for the error variables. Restricted to 284 couples with at least one observed cycle length post-enrollment. Couples are sorted by observed cycle length and results are shown for the smallest, middle, and largest 10%.

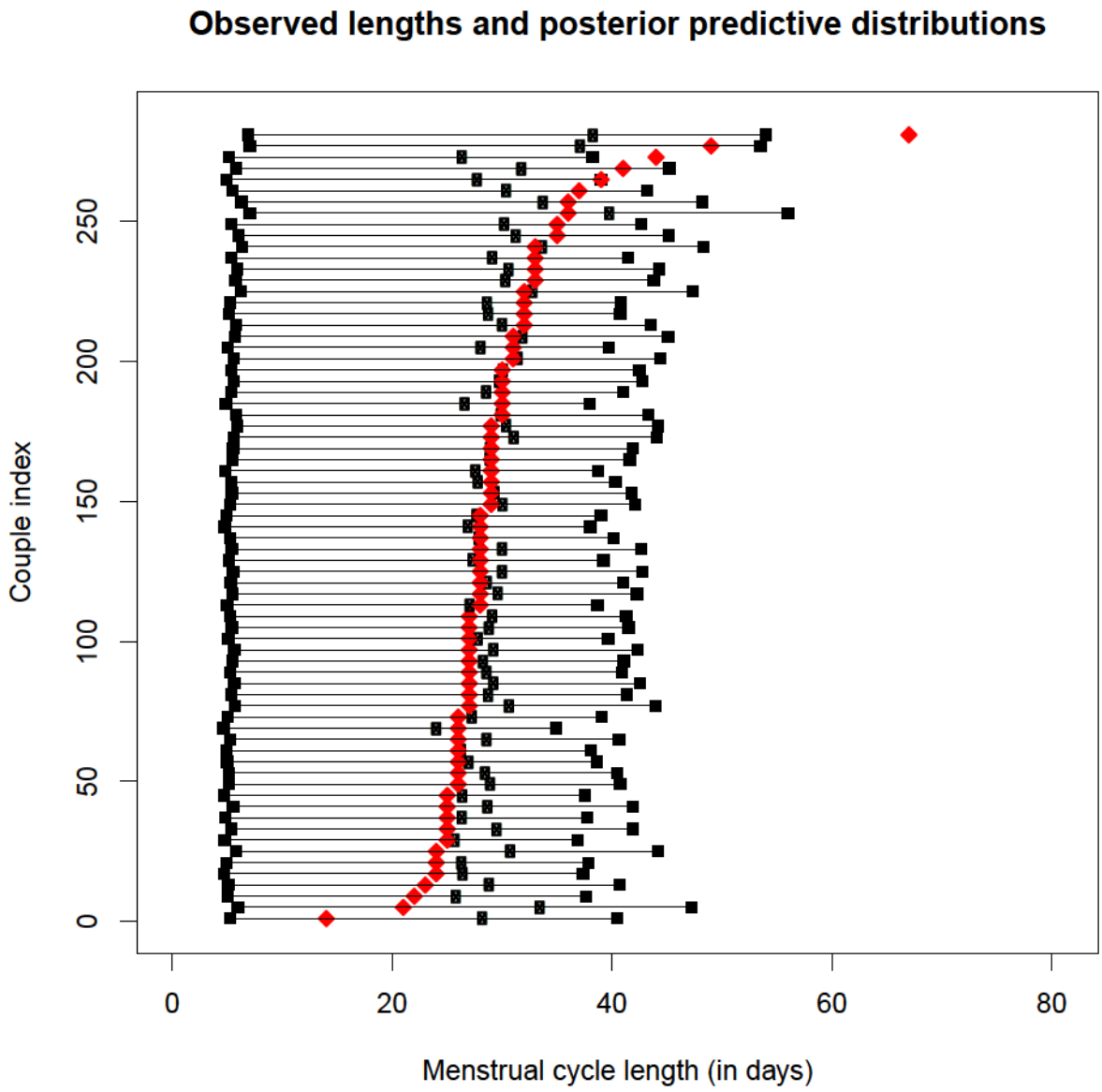


Figure 3.A.3: Posterior medians (black, circle) and 95% equal-tail credible intervals (black, squares) for predicted menstrual cycle length compared with observed length of first cycle post-enrollment (red, diamond). The predictive posterior distribution is estimated using the proposed menstrual cycle length model assuming a log-normal mixture distribution for the error variables. Restricted to 284 couples with at least one observed cycle length post-enrollment. Couples are sorted by observed cycle length and results are shown for every fourth couple.

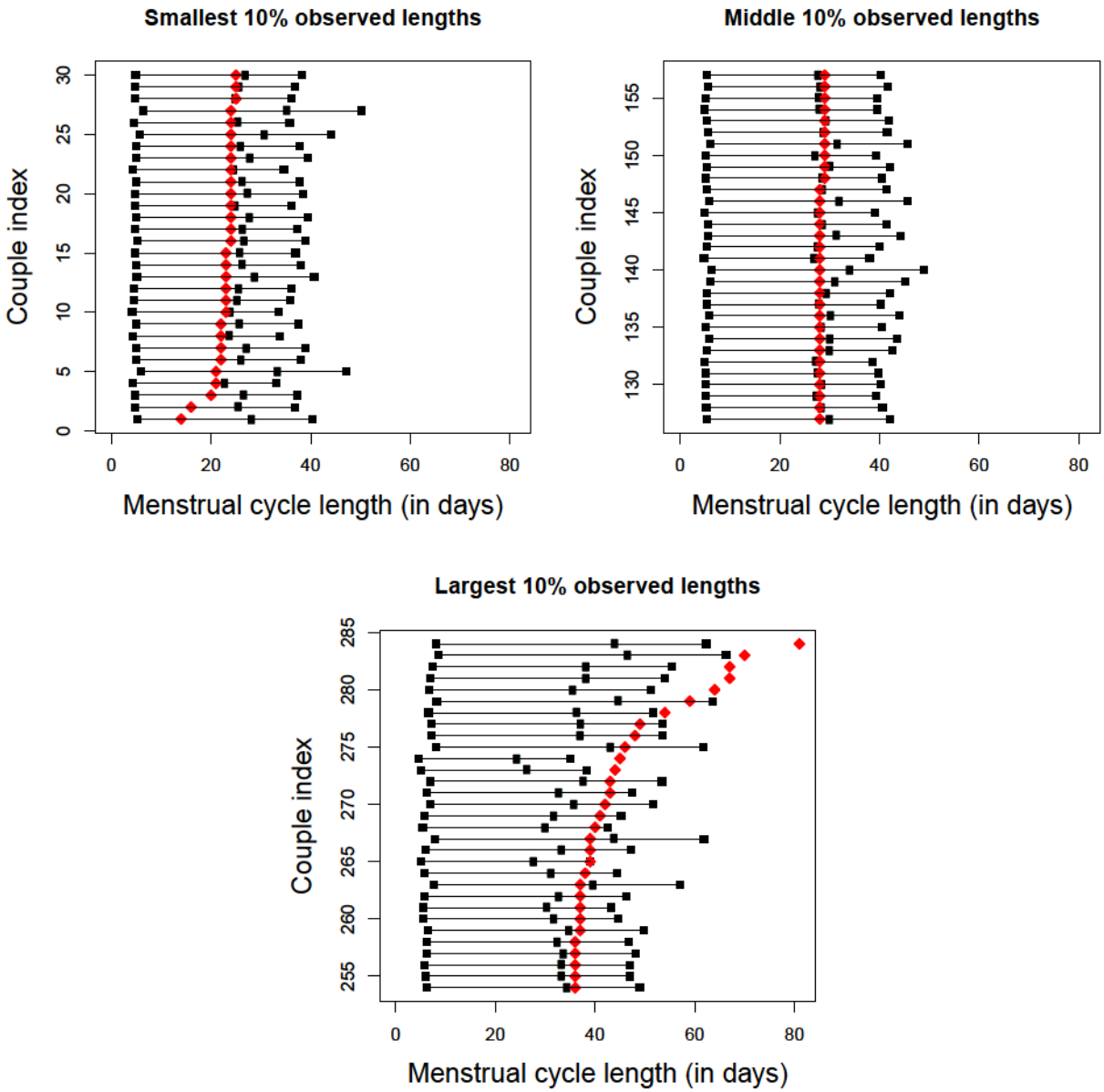


Figure 3.A.4: Posterior medians (black, circle) and 95% equal-tail credible intervals (black, squares) for predicted menstrual cycle length compared with observed length of first cycle post-enrollment (red, diamond). The predictive posterior distribution is estimated using the proposed menstrual cycle length model assuming a log-normal mixture distribution for the error variables. Restricted to 284 couples with at least one observed cycle length post-enrollment. Couples are sorted by observed cycle length and results are shown for the smallest, middle, and largest 10%.

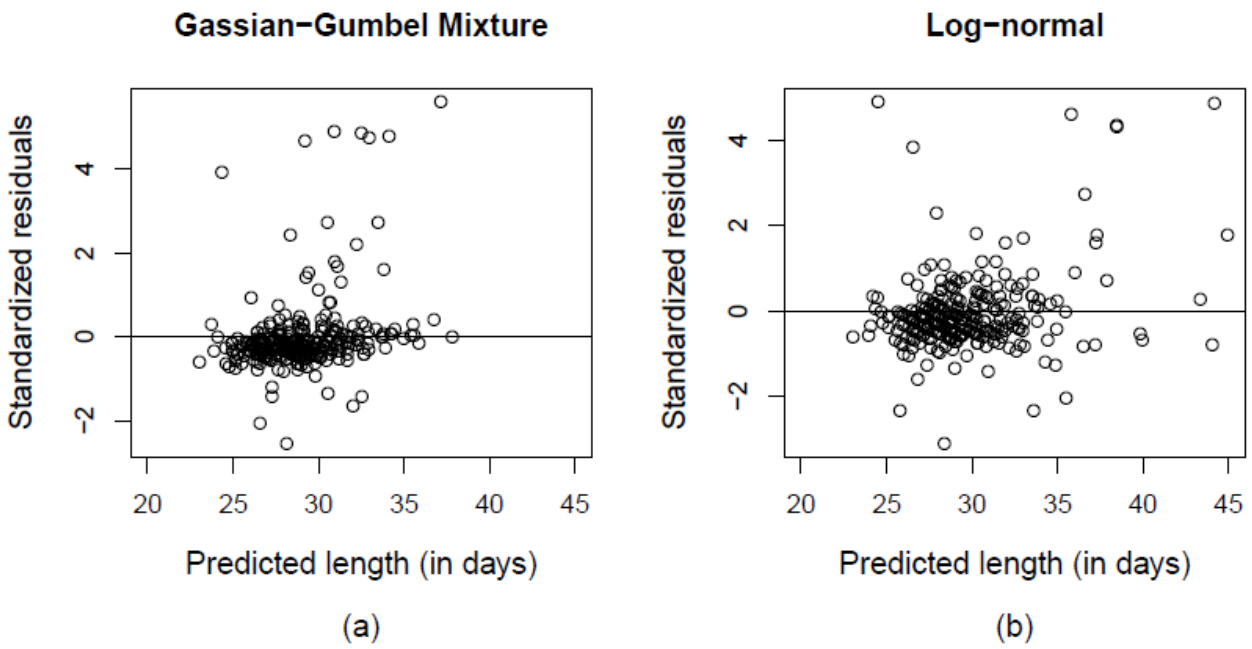


Figure 3.A.5: Standardized residuals versus predicted menstrual cycle lengths for (a) the model assuming Gaussian-Gumbel mixture error distribution and (b) the model assuming log-normal error distribution.

3.A.2. Time-to-Pregnancy Survival Distribution

An application of our fecundity modeling approach is the calculation of the time-to-pregnancy (TTP) survival distribution for various scenarios. This is potentially useful for counseling couples planning pregnancy. In Figure 3.A.6 we compare the survival probabilities (i.e. the probably that TTP is greater than j cycles) for various patterns of intercourse (fixed throughout cycles) adjusted for mean couple age, age difference, smoking status, sperm morphology and total sperm count. The intercourse pattern corresponding to the shortest estimated median TTP is intercourse every day for 7 days ending on ovulation day. The curve for intercourse on the two peak days is slightly lower than intercourse every other day. Figure 3.A.6 compares the TTP survival distributions varying the timing of intercourse for three consecutive acts. As expected the shortest median TTP corresponds to intercourse on peak days. An important counseling note is that the pattern of late timing corresponds to a much longer median TTP than that of early timing. Lastly, we can compare the TTP survival distribution for various risk factors fixing intercourse patterns over cycles. Figure 3.A.8 compares the four combinations of smoking fixing intercourse to one act on the peak day. Couples in which both partners smoke have the longest estimated median TTP while those in which only the female smokes have a slightly shorter estimated median TTP than couples in which only the male smokes.

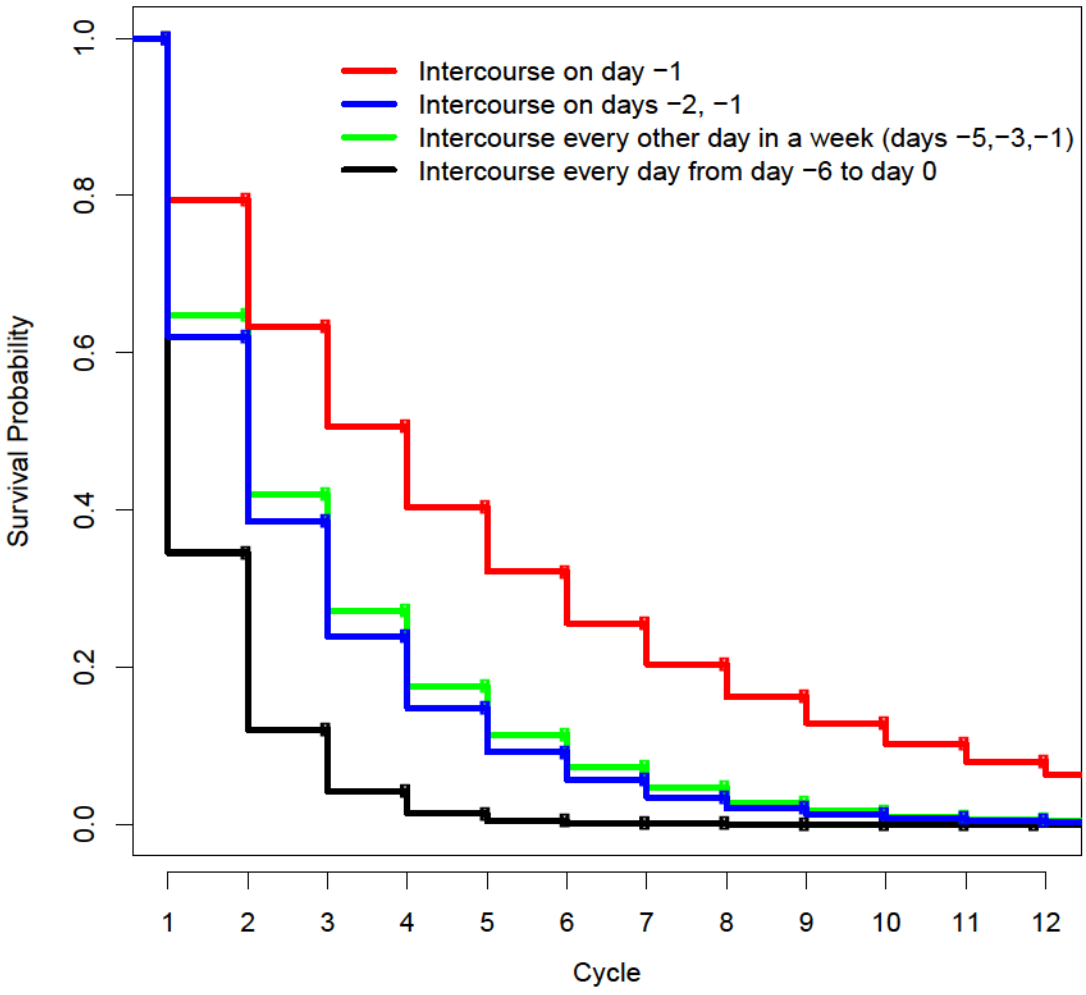


Figure 3.A.6: The probability that the time-to-pregnancy is greater than j cycles comparing intercourse patterns which are fixed throughout cycles. Menstrual cycle length, age, smoking status, and semen parameters are fixed at population mean values.

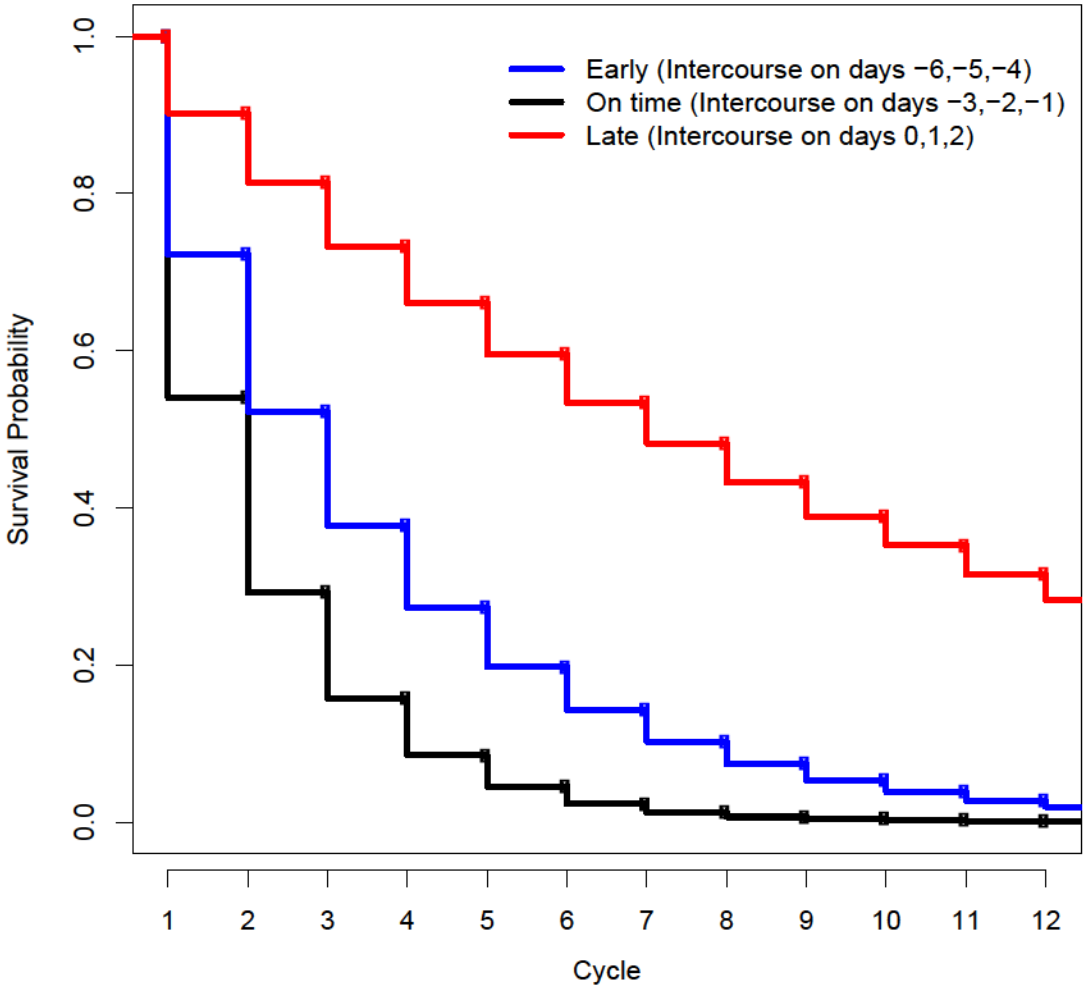


Figure 3.A.7: The probability that the time-to-pregnancy is greater than j cycles comparing intercourse timing patterns which are fixed throughout cycles. Menstrual cycle length, age, smoking status, and semen parameters are fixed at population mean values.

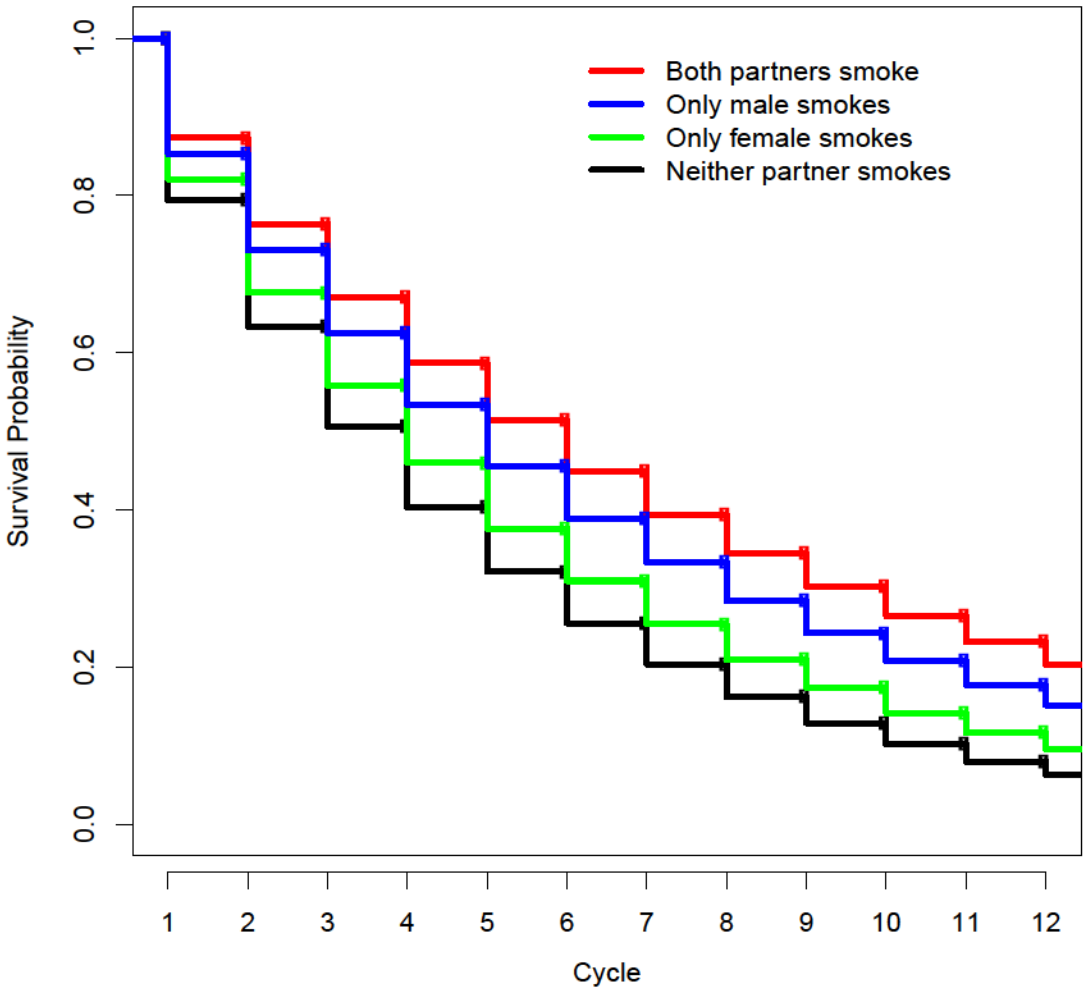


Figure 3.A.8: The probability that the time-to-pregnancy is greater than j cycles comparing smoking status. Intercourse pattern was fixed at one act on day -1 throughout cycles. Menstrual cycle length, age, and semen parameters are fixed at population mean values.

Chapter 4

Perfluoroalkyl Surfactants and Their Relations with Menstrual Cycle Length and Fecundity: the LIFE Study

4.1. Introduction

In Chapter 1 we provided some background on perfluoroalkyl surfactants (PFASs). Briefly, PFASs are ubiquitous environmental chemicals used in a variety of household products (e.g., surfactants and surface protectors in carpets, leather, paper, packaging, fabric and upholstery) (Giesy and Kannan 2002) and found at detectable levels in US and other populations around the world (Kannan et al. 2004; Kato et al. 2011; Olsen et al. 2012; Wang et al. 2011). There is considerable interest in studying the role of PFASs and other POPs in human fecundity, or the biologic capacity of men and women for reproduction (Buck Louis 2011). Epidemiologic studies of fecundity have reported negative associations, indicated by prolonged time-to-pregnancy, with perfluorooctane sulfonamide (PFOSA) (Buck Louis et al. 2012), and in a separate study with perfluorooctane sulfonate (PFOS), and perfluorooctanoate (PFOA) (Fei et al. 2009). For males, there is an evolving body of literature, though not consistent, for potential toxicity of PFASs with respect to semen quality parameters (Toft et al. 2012; Raymer et al. 2012; Buck Louis et al. 2014). However, the mechanism by which PFASs may be associated with couple fecundity is unclear. Although there is evidence to

support the important role of menstrual cycle length in female fecundity, see Chapter 3 and (Small et al. 2006, 2010; McLain et al. 2012), research on PFASs and menstrual cycle characteristics is limited, with only one study to our knowledge which examined PFOS and PFOA (Lyngsø et al. 2013). To address this data gap, we sought to investigate the association between female serum PFAS concentrations and (i) menstrual cycle length and (ii) fecundity in the context of menstrual cycle length.

4.2. Methods

4.2.1 Study Design And Cohort

The Longitudinal Investigation of Fertility and the Environment (LIFE) (Buck Louis et al. 2011), was a prospective pregnancy study in which 501 couples residing in Texas and Michigan between 2005-2009 and planning pregnancy within two months were enrolled. The couples were followed until pregnancy or up to one year of attempting pregnancy with the aim of studying the effects of environmental chemicals on couple fecundity. Couples were included in the study if they fulfilled the following criteria: married or in a committed relationship, females aged 18-40 and males age 18+ years, English or Spanish speaking, self-reported menstrual cycle lengths within 21-42 days, and no hormonal birth control injections in the past 12 months. Couples were excluded if they had been sterilized or told by a physician that they could not achieve pregnancy without medical aid. Full human subjects approval was obtained from all participating institutions, and all participants provided written informed consent prior to enrollment into the study.

4.2.2 Data and Biospecimen Collection

Baseline characteristics were collected by a research assistant in an at-home interview held on the day of enrollment following a pregnancy test to ensure that the female did not test positive for pregnancy. Also on enrollment day, 10cc of blood was collected by a nurse using equipment deemed free of environmental chemicals under study and transported on

ice to a laboratory for processing in which 2 ml of serum were used for analysis of PFAS concentrations.

Data on menstrual bleeding intensity was collected in a daily journal and used in conjunction with an at home fertility monitor to define cycles distinct from episodic bleeding. The length of the menstrual cycle was defined as the time (days) from the onset of bleeding that increased in intensity and lasted two or more days to the day preceding the next similar bleeding event. As follow up began on any day of the females menstrual cycle, the length of the cycle in which the couple was enrolled was calculated as the sum of the time since last menstrual period, recalled at enrollment, and the prospectively observed time from enrollment to the next bleeding event. If the couple became pregnant, increased levels of progesterone precluded menstrual bleeding in preparation for implantation. To avoid excluding couples who became pregnant in the first cycle, the length of the cycle in which the couple became pregnant was right-censored. Couples who did not become pregnant were censored on the last day of the menstrual cycle preceding their withdrawal from the study or at one year of follow-up, thereby excluding 16 couples who withdrew before the first observed bleeding event. We excluded cycles which were shorter than 9 days or longer than 89 days (1%), leaving 2,198 cycles remaining.

The female partner was instructed by the research nurse in the use of Clearblue®Easy pregnancy tests which reliably detect human chorionic gonadotropin (hCG) levels over 25mIU/mL, and the Clearblue®Easy fertility monitor which tracks daily levels of oestrone-3-glucuronide and luteinizing hormone (LH). The ovulation day in a cycle was estimated by the peak fertile day identified by the monitor, or the second peak day in the case of two consecutive peak days. In cycles in which the peak was not detected, if the woman tested for at least 90% of 20 test days, the cycle was assumed to be anovulatory (8.5%) and thus not at risk for pregnancy; otherwise, in 14% of cycles, the ovulation day was imputed using the mean of the ovulation day of the females other cycles. If there were no other cycles in which the ovulation day could be estimated, the day of peak LH was imputed as the ovulation day.

Less than 1% of cycles were excluded for which no ovulation dates could be estimated due to noncompliance in monitor use resulting in a final count of 483 couples and 2,174 cycles.

Daily data on the couples intercourse pattern frequency was reported independently in male and female journals. The female journal was used to calculate the timing of intercourse relative to the estimated ovulation day. If the intercourse question was left blank in the female journal, then the male journal was consulted. If the intercourse question was blank in both journals, then it was assumed that the couple did not have intercourse on this day ($< 3\%$ of days).

4.2.3 Toxicological Analysis

Quantification of PFASs in serum, reported in ng/mL, was done at the Division of Laboratory Sciences in the National Center for Environmental Health, Centers for Disease Control and Prevention, using isotope dilution high performance liquid-chromatography-tandem mass spectrometry and established operating procedures (Kato et al. 2011; Kuklenyik et al. 2005). Seven PFASs were analyzed: 2-(N-ethyl-perfluorooctane sulfonamido) acetate (EtFOSAA), 2-(N-methyl-perfluorooctane sulfonamido) acetate (MeFOSAA), perfluorodecanoate (PFDA), perfluorononanoate (PFNA), perfluorooctane sulfonamide (PFOSA), perfluorooctane sulfonate (PFOS), and perfluorooctanoate (PFOA). The limit of detection (LOD) was 0.1 for PFNA, PFOSA, and PFOA and 0.1 for the remaining 4 PFASs. Measurement of cotinine in serum (in ng/ml) was conducted using liquid chromatography-isotope dilution tandem mass spectrometry (Bernert et al. 1997). We used a cut point of 10 ng/mL to define active smoking (yes/no) (Pirkle et al. 1996).

4.2.4 Statistical Analysis

In the descriptive phase, we describe female characteristics and serum PFAS concentrations by mean menstrual cycle length. For descriptive analyses only, we categorized mean menstrual cycle length using cutoffs based on those found in previous literature (Small et al.

2007; Kolstad et al. 1999; Buck Louis et al. 2011).

In the analytic phase, we first fit a two stage model where in stage (i), we model the relation between individual serum concentrations of 7 PFASs and female menstrual cycle length, and in stage (ii) we model the relation between PFASs and couple fecundity in the context of the fitted menstrual cycle length from stage (i). We compare this two stage approach with the Bayesian joint model of menstrual cycle length and fecundity developed in Chapter 3, which accounts for the uncertainty in estimating the distribution of menstrual cycle length.

In both the two stage and joint models, we model the role of each PFAS in menstrual cycle length (continuous) using the Bayesian hierarchical accelerated failure time model in (3.1) with the exposure incorporated in the covariate vector, $\mathbf{v}$. In this model, the exponential of the regression coefficient, $\exp(\eta)$, represents the multiplicative acceleration factor (AF) by which the menstrual cycle length is increased (AF>1) or decreased (AF<1) for each unit change in the PFAS covariate conditional on the female-specific random effect. We also adjusted for a priori potential confounders (i.e., age (years), body mass index (BMI) (categorical), and active smoking at enrollment (yes/no)). To allow for the possibility of extremely short or long cycle lengths we assumed a mixture of Gaussian and Gumbel error distributions. We address length-bias in the distribution of length for the cycle in which the couple enrolls (Lum et al. 2014) and right censoring for the length of the cycle in which the couple becomes pregnant.

Next, we use (3.6) and(3.7) to model the association between female-specific menstrual cycle length and fecundity as measured by the probability of pregnancy in a cycle conditional on previous cycles ending without pregnancy. We incorporate menstrual cycle length in stage (ii) of the two-stage model using the median of its posterior predictive distribution; while in the joint model, we average over the full posterior predictive distribution. We include female PFAS concentration in the covariate vector $\mathbf{z}$ in (3.7) to investigate the relation between female PFAS exposure and day-specific probability of pregnancy with adjustment

for the couples intercourse pattern in the current and previous cycles and a priori potential confounders (i.e., female age (years), BMI (categorical), active smoking at enrollment (yes/no)).

For the aforementioned models, we fit both a single model for each PFAS and a model consisting of all 7 PFAS concentrations simultaneously. In addition, to determine if a weighted combination of multiple PFASs might be related to the endpoints and to reduce issues related to multicollinearity, we conducted unadjusted and adjusted principal component regression in both the menstrual cycle length model and the fecundity model of the joint model. This was implemented by first calculating the principal components of the instrument observed concentrations of the 7 PFASs and then including the first three principal components in the regression models.

A challenge that must be addressed is how to handle concentrations below the limit of detection (LOD). According to (Armbruster and Pry 2008):

> The limit of detection is the lowest analyte concentration likely to be reliably distinguished from the [limit of blank] (LoB) and at which detection is feasible. The LoB is the highest *apparent* analyte concentration expected to be found when replicates of a blank sample containing no analyte are tested.

One approach is substitution by $\frac{1}{2}$LOD, or $\frac{1}{\sqrt{2}}$LOD or 0; however this has been shown to introduce bias (Richardson and Ciampi 2003; Schisterman et al. 2006). Therefore, we used the instrument observed concentrations without substitution for values below LOD. As the units of measurement are small we scaled the instrument observed concentrations by the interquartile range (IQR), with the exception of PFOSA and EtFOSAA for which the IQR could not be determined. We also considered log transformations of the instrument observed concentrations. As several of the PFASs have some values below the limit of detection (LOD), we fit separate models replacing the continuous PFAS concentrations with categorical variables, specifically tertiles of concentration with the lowest as the reference group. We used dichotomous categorizations (e.g., above and below LOD, above and below 75th percentile)

Table 4.1: Characteristics by mean menstrual cycle length (days), LIFE Study, 2005-2009. Data are mean ± standard deviation for continuous variables and no. (%) for categorical variables. Active smoking is defined as baseline cotinine ≥ 10ng/ml.

Characteristic	Short Mean Cycle Length (≤ 24 days) (n=21)	Average Mean Cycle Length (25-31 days) (n=305)	Long Mean Cycle Length (≥ 32 days) (n=157)
Age (years)	32.1 ± 4.8	30.5 ± 4.2	29.0 ± 3.6
BMI(kg/m2)	28.3 ± 9.3	27.3 ± 6.5	27.9 ± 8.2
Active smoking (yes/no)	3 (14)	29 (10)	22 (14)
Gravid	12 (57)	186 (61)	79 (50)
Parous	12 (57)	153 (51)	61 (39)
Nonwhite	1 (5)	42 (14)	24 (15)
≤ High school graduate/GED	0 (0)	16 (5)	9 (6)
No health insurance	2 (10)	17 (6)	17 (11)

for PFOSA and EtFOSAA as more than 90% of values of these PFASs were below LOD. We conducted all analyses under the missing-at-random assumption and used Markov Chain Monte Carlo methods to impute missing PFAS concentrations and cotinine levels (≤4%) data due to insufficient blood for quantification (Schafer 2010). We did not adjust for multiple comparisons in keeping with the exploratory purposes of this investigation.

4.3. Results

Baseline female characteristics by mean menstrual cycle length are shown in Table 4.1. Mean cycle length tended to decrease with age. The geometric means (GMs) and 95% confidence intervals for PFAS serum concentrations by menstrual cycle length category are shown in Table 4.2. For calculation of the GM, zeros were replaced with half the value of the minimum positive concentration. None of the GMs differed significantly by mean menstrual cycle length category; however, the number of women in the short mean cycle length category was very small.

Table 4.2: Geometric means (GM) (95% CI) of preconception serum concentrations of PFASs (ng/mL) by mean menstrual cycle length (days), LIFE study, 2005-2009.

PFAS (ng/mL)	LOD (ng/mL)	<LOD (%)	Short Mean Cycle Length ($\leq$ 24 days) GM (95% CI)	Average Mean Cycle Length (25-31 days) GM (95% CI)	Long Mean Cycle Length ($\geq$ 32 days) GM (95% CI)
EtFOSAA	0.2	97	$_{0.05}0.06_{0.07}$	$_{0.06}0.06_{0.07}$	$_{0.06}0.06_{0.06}$
MeFOSAA	0.2	25	$_{0.26}0.40_{0.63}$	$_{0.25}0.27_{0.30}$	$_{0.25}0.29_{0.33}$
PFDA	0.2	9	$_{0.19}0.29_{0.42}$	$_{0.34}0.37_{0.40}$	$_{0.29}0.33_{0.36}$
PFNA	0.1	1	$_{0.58}0.93_{1.51}$	$_{1.08}1.16_{1.25}$	$_{0.95}1.06_{1.17}$
PFOSA	0.1	89	$_{0.05}0.06_{0.07}$	$_{0.05}0.05_{0.06}$	$_{0.05}0.05_{0.06}$
PFOS	0.2	<1	$_{9.70}12.38_{15.81}$	$_{11.08}11.92_{12.81}$	$_{9.69}10.73_{11.88}$
PFOA	0.1	<1	$_{2.43}2.99_{3.68}$	$_{2.98}3.20_{3.43}$	$_{2.59}2.89_{3.21}$

4.3.1 Role of PFASs in Menstrual Cycle Length

The AF and 95% credible interval (CI) corresponding to each PFAS in the menstrual cycle length stage of the two stage model are presented in Table 4.3. For each PFAS, we first display the results using the categorical form of the PFAS followed by the results using the continuous (scaled) form of the PFAS. The AF is the exponentiated form of the regression coefficient. For the categorized exposures, the AF is the factor by which the menstrual cycle length is increased (AF>1) or decreased (AF<1) for a woman with serum PFAS concentration in the 2nd or 3rd tertile compared to the 1st tertile of exposure, conditional on the random effect. Both the 2nd and 3rd tertiles of concentrations of PFOA and the 2nd tertile of MeFOSAA were associated with shorter menstrual cycle lengths compared with the respective 1st tertile of each PFAS. When adjusting for age (years), BMI (categorical), and active smoking at enrollment (yes/no), the 2nd and 3rd tertiles of PFOA were again associated with shorter lengths, while the 2nd tertile of PFDA was associated with longer lengths. As an illustration, we present boxplots of predicted menstrual cycle length by PFOA tertile in Figure 4.1. Each of the aforementioned associations were also observed in the model which included all 7 PFASs simultaneously. Some moderate to high correlations in concentration were observed between PFASs (Tables 4.5 and 4.6). PFNA was correlated with PFDA, PFOS, and PFOA;

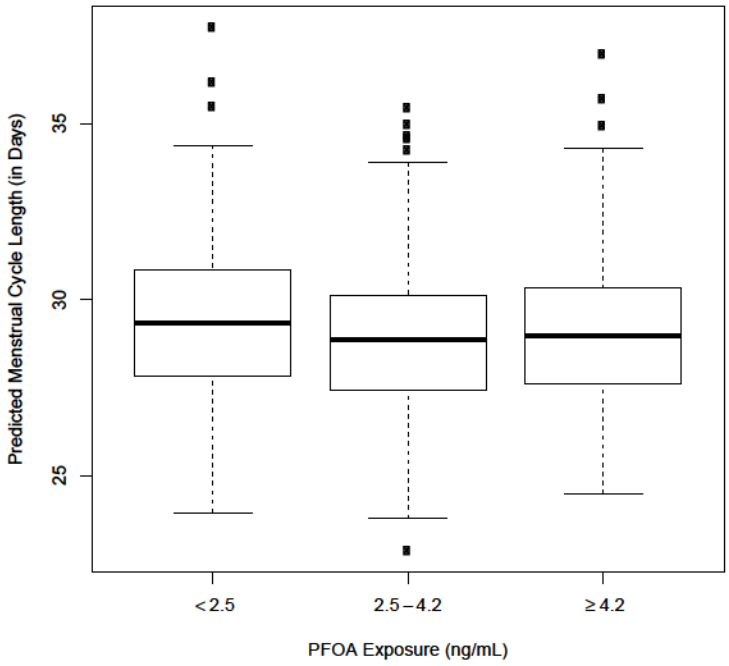


Figure 4.1: Boxplots of predicted menstrual cycle length (days) by tertile of PFOA exposure (ng/mL).

and PFDA was moderately correlated with PFOS and PFOA. In continuous (scaled) or log-transformed form, none of the PFASs were significantly associated with menstrual cycle length. In the menstrual cycle length part of the joint model (see Table 4.4), we again observed associations with cycle length for PFOA, MeFOSAA, and PFDA when modeled simultaneously. When modeled individually, the 2nd tertile of PFOA, MeFOSAA, and PFDA were each associated with changes in length compared with the corresponding 1st tertile in unadjusted models. With adjustment for age (years), BMI (categorical), and active smoking at enrollment (yes/no), the 2nd tertile of PFDA and the 3rd tertile of PFOA were associated with changes in menstrual cycle length.

Table 4.3: Two stage model: unadjusted and adjusted AF and 95%CI (displayed as: $_{\text{Lo}}\text{AF}_{\text{Up}}$ (Louis and Zeger 2009)) of menstrual cycle length (days) associated with preconception serum PFAS concentrations (ng/mL), LIFE Study, 2005-2009. Adjusted for age (years), body mass index (categorized), active smoking at enrollment (yes/no). For models with categorized PFAS concentrations, the interpretation is multiplicative change in length for respective tertile compared with 1st tertile of exposure. For models with continuous PFAS concentrations, the interpretation is multiplicative change in length per IQR change in concentration.

	Menstrual Cycle Length Submodel			
	Single PFAS Model		Multiple PFAS Model	
PFAS (ng/mL)	Unadjusted $_{\text{Lo}}\text{AF}_{\text{Up}}$	Adjusted $_{\text{Lo}}\text{AF}_{\text{Up}}$	Unadjusted $_{\text{Lo}}\text{AF}_{\text{Up}}$	Adjusted $_{\text{Lo}}\text{AF}_{\text{Up}}$
EtFOSAA				
$\geq$ 0.055	$_{0.97}1.00_{1.02}$	$_{0.99}1.01_{1.04}$	$_{0.96}1.00_{1.02}$	$_{0.98}1.01_{1.04}$
Continuous	$_{0.96}0.99_{1.04}$	$_{0.97}1.01_{1.05}$	$_{0.95}0.97_{1.00}$	$_{0.97}1.01_{1.06}$
MeFOSAA				
0.02-0.03	$_{0.94}0.97_{0.99}$	$_{0.95}0.97_{1.01}$	$_{0.93}0.97_{0.99}$	$_{0.95}0.98_{1.01}$
$\geq$ 0.04	$_{0.96}0.99_{1.02}$	$_{0.97}1.00_{1.03}$	$_{0.96}1.00_{1.03}$	$_{0.97}1.00_{1.04}$
Continuous (scaled by IQR)	$_{0.99}1.00_{1.00}$	$_{0.99}1.00_{1.00}$	$_{0.97}0.99_{1.00}$	$_{0.99}1.00_{1.01}$
PFDA				
0.03-0.04	$_{1.00}1.03_{1.06}$	$_{1.01}1.03_{1.06}$	$_{0.99}1.03_{1.06}$	$_{1.00}1.04_{1.07}$
$\geq$ 0.05	$_{0.97}1.00_{1.03}$	$_{0.99}1.01_{1.04}$	$_{0.95}0.99_{1.04}$	$_{0.96}1.01_{1.06}$
Continuous (scaled by IQR)	$_{0.99}1.00_{1.01}$	$_{0.99}1.00_{1.01}$	$_{0.98}1.00_{1.07}$	$_{0.99}1.01_{1.02}$
PFNA				
0.09-1.4	$_{0.99}1.02_{1.05}$	$_{1.00}1.02_{1.05}$	$_{0.99}1.02_{1.06}$	$_{0.98}1.02_{1.06}$
$\geq$ 1.5	$_{0.98}1.00_{1.03}$	$_{0.99}1.01_{1.04}$	$_{0.98}1.02_{1.07}$	$_{0.98}1.02_{1.07}$
Continuous (scaled by IQR)	$_{0.99}1.00_{1.01}$	$_{0.99}1.00_{1.01}$	$_{0.99}1.01_{1.18}$	$_{0.99}1.00_{1.02}$
PFOSA				
$\geq$ 0.1	$_{0.96}0.99_{1.03}$	$_{0.96}0.99_{1.03}$	$_{0.95}0.98_{1.02}$	$_{0.95}0.98_{1.02}$
Continuous	$_{0.71}0.91_{1.15}$	$_{0.75}0.93_{1.15}$	$_{0.78}1.05_{2.55}$	$_{0.74}0.99_{1.26}$
PFOS				
9.5-15.1	$_{0.97}1.00_{1.03}$	$_{0.98}1.01_{1.04}$	$_{0.98}0.99_{1.04}$	$_{0.98}1.02_{1.05}$
$\geq$ 15.2	$_{0.97}1.00_{1.03}$	$_{0.98}1.01_{1.04}$	$_{0.99}1.02_{1.06}$	$_{0.99}1.03_{1.07}$
Continuous (scaled by IQR)	$_{0.98}1.00_{1.01}$	$_{0.99}1.00_{1.01}$	$_{0.92}1.00_{1.01}$	$_{0.99}1.00_{1.01}$
PFOA				
2.5-4.1	$_{0.94}0.97_{1.00}$	$_{0.95}0.98_{1.00}$	$_{0.92}0.95_{0.98}$	$_{0.92}0.95_{0.98}$
$\geq$ 4.2	$_{0.95}0.97_{1.00}$	$_{0.95}0.97_{1.00}$	$_{0.92}0.95_{0.99}$	$_{0.91}0.95_{0.98}$
Continuous (scaled by IQR)	$_{0.97}0.99_{1.01}$	$_{0.98}0.99_{1.01}$	$_{0.90}0.99_{1.01}$	$_{0.97}0.99_{1.01}$

Table 4.4: Joint model: unadjusted and adjusted AF and 95%CI (displayed as: $_{\text{Lo}}\text{AF}_{\text{Up}}$ (Louis and Zeger 2009)) of menstrual cycle length (days) associated with preconception serum PFAS concentrations (ng/mL), LIFE Study, 2005-2009. Adjusted for age (years), body mass index (categorized), active smoking at enrollment (yes/no). For models with categorized PFAS concentrations, the interpretation is multiplicative change in length for respective tertile compared with 1st tertile of exposure. For models with continuous PFAS concentrations, the interpretation is multiplicative change in length per IQR change in concentration.

	Menstrual Cycle Length Submodel			
	Single PFAS Model		Multiple PFAS Model	
PFAS (ng/mL)	Unadjusted $_{\text{Lo}}\text{AF}_{\text{Up}}$	Adjusted $_{\text{Lo}}\text{AF}_{\text{Up}}$	Unadjusted $_{\text{Lo}}\text{AF}_{\text{Up}}$	Adjusted $_{\text{Lo}}\text{AF}_{\text{Up}}$
EtFOSAA				
$\geq$ 0.055	$_{0.97}1.00_{1.02}$	$_{0.99}1.01_{1.04}$	$_{0.97}1.00_{1.02}$	$_{0.98}1.01_{1.04}$
Continuous	$_{0.96}1.00_{1.04}$	$_{0.97}1.01_{1.05}$	$_{0.96}1.00_{1.04}$	$_{0.97}1.02_{1.06}$
MeFOSAA				
0.02-0.03	$_{0.94}0.97_{0.99}$	$_{0.95}0.98_{1.01}$	$_{0.94}0.97_{0.99}$	$_{0.95}0.98_{1.00}$
$\geq$ 0.04	$_{0.96}0.99_{1.01}$	$_{0.97}1.00_{1.03}$	$_{0.96}0.99_{1.02}$	$_{0.96}1.00_{1.02}$
Continuous (scaled by IQR)	$_{0.99}0.99_{1.00}$	$_{0.99}1.02_{1.00}$	$_{0.99}1.00_{1.00}$	$_{0.98}1.00_{1.00}$
PFDA				
0.03-0.04	$_{1.00}1.03_{1.05}$	$_{1.00}1.03_{1.05}$	$_{0.99}1.02_{1.06}$	$_{1.00}1.03_{1.07}$
$\geq$ 0.05	$_{0.98}1.00_{1.03}$	$_{0.99}1.01_{1.04}$	$_{0.95}0.99_{1.03}$	$_{0.98}1.02_{1.06}$
Continuous (scaled by IQR)	$_{0.99}1.00_{1.01}$	$_{0.99}1.00_{1.01}$	$_{0.99}1.00_{1.02}$	$_{1.00}1.01_{1.02}$
PFNA				
0.09-1.4	$_{0.99}1.02_{1.04}$	$_{0.99}1.02_{1.04}$	$_{1.00}1.03_{1.07}$	$_{0.98}1.02_{1.05}$
$\geq$ 1.5	$_{0.98}1.00_{1.03}$	$_{0.99}1.01_{1.04}$	$_{0.98}1.03_{1.08}$	$_{0.98}1.02_{1.06}$
Continuous (scaled by IQR)	$_{0.99}1.00_{1.01}$	$_{0.99}1.00_{1.01}$	$_{0.98}1.00_{1.02}$	$_{0.98}1.00_{1.01}$
PFOSA				
$\geq$ 0.1	$_{0.96}0.99_{1.02}$	$_{0.96}0.99_{1.02}$	$_{0.95}1.01_{1.02}$	$_{0.95}0.98_{1.02}$
Continuous	$_{0.72}0.90_{1.12}$	$_{0.75}0.93_{1.15}$	$_{0.72}0.91_{1.18}$	$_{0.76}0.99_{1.24}$
PFOS				
9.5-15.1	$_{0.98}1.01_{1.03}$	$_{0.98}1.01_{1.03}$	$_{0.98}1.01_{1.04}$	$_{0.99}1.02_{1.04}$
$\geq$ 15.2	$_{0.98}1.00_{1.03}$	$_{0.98}1.01_{1.03}$	$_{0.98}1.02_{1.05}$	$_{0.99}1.02_{1.06}$
Continuous (scaled by IQR)	$_{0.99}1.00_{1.00}$	$_{0.99}1.00_{1.01}$	$_{0.99}1.00_{1.01}$	$_{0.99}1.00_{1.02}$
PFOA				
2.5-4.1	$_{0.95}0.97_{1.00}$	$_{0.95}0.98_{1.01}$	$_{0.93}0.95_{0.99}$	$_{0.93}0.95_{0.99}$
$\geq$ 4.2	$_{0.95}0.98_{1.00}$	$_{0.96}0.98_{1.00}$	$_{0.93}0.96_{1.00}$	$_{0.92}0.95_{0.98}$
Continuous (scaled by IQR)	$_{0.98}0.99_{1.01}$	$_{0.98}0.99_{1.01}$	$_{0.97}0.99_{1.01}$	$_{0.97}0.99_{1.01}$

Table 4.5: Correlations among serum concentrations of PFASs, LIFE Study, 2005-2009.

	MeFOSAA	PFDA	PFNA	PFOSA	PFOS	PFOA
EtFOSAA	0.03	0.04	0.05	0.03	0.06	0.05
MeFOSAA		0.03	0.10	0.46	0.24	0.09
PFDA			0.73	-0.03	0.56	0.55
PFNA				-0.06	0.60	0.60
PFOSA					0.13	-0.06
PFOS						0.45

Table 4.6: Correlations among log transformed serum concentrations of PFASs, LIFE Study, 2005-2009.

	MeFOSAA	PFDA	PFNA	PFOSA	PFOS	PFOA
EtFOSAA	0.11	0.06	0.07	0.08	0.11	0.07
MeFOSAA		0.04	0.08	0.45	0.26	0.07
PFDA			0.81	-0.07	0.56	0.56
PFNA				-0.11	0.60	0.64
PFOSA					0.12	-0.07
PFOS						0.57

4.3.2 Role of PFASs in Probability of Pregnancy

The odds ratio (OR) and 95% credible interval (CI) corresponding to each PFAS in the pregnancy stage of the two stage model are presented in Table 4.7. The OR is the exponentiated coefficient corresponding to the PFAS covariate in $\mathbf{z}$ of (3.7). An OR significantly less than 1 indicates the corresponding PFAS is negatively associated with the probability of pregnancy. We found increased concentrations of PFOSA were negatively associated with the probability of pregnancy in both the individual and multiple PFAS models, while accounting for menstrual cycle length estimated in the first stage. This negative association was observed when PFOSA was included in the model either as continuous or dichotomized as above versus below LOD. In Figure 4.2, we illustrate this negative role of PFOSA with box plots of the conditional probability of pregnancy by cycle for females with PFOSA concentration above ('high') and below ('low') LOD. In the adjusted multiple PFAS model, the 2nd tertile of PFNA was also negatively associated with the probability of pregnancy in comparison to the 1st tertile of PFNA. This association was also observed in the joint model (Table 4.8) which accounts for the uncertainty in estimating the females menstrual cycle length. Also of note is the quadratic relation between menstrual cycle length and the probability of pregnancy found in both the two stage and joint models, which indicates both short and long cycle lengths are negatively associated with the probability of pregnancy.

For the principal component (PC) analysis, the proportional contribution of each PFAS is shown in Table 4.9. We included the first three PCs in both the menstrual cycle length submodel and the pregnancy submodel of the joint model; however, the credible interval for the exponentiated coefficient associated with each PC included 1 (Tables 4.10-4.11). We also fit the joint model with the PCs only in the menstrual cycle length submodel, but the results did not change.

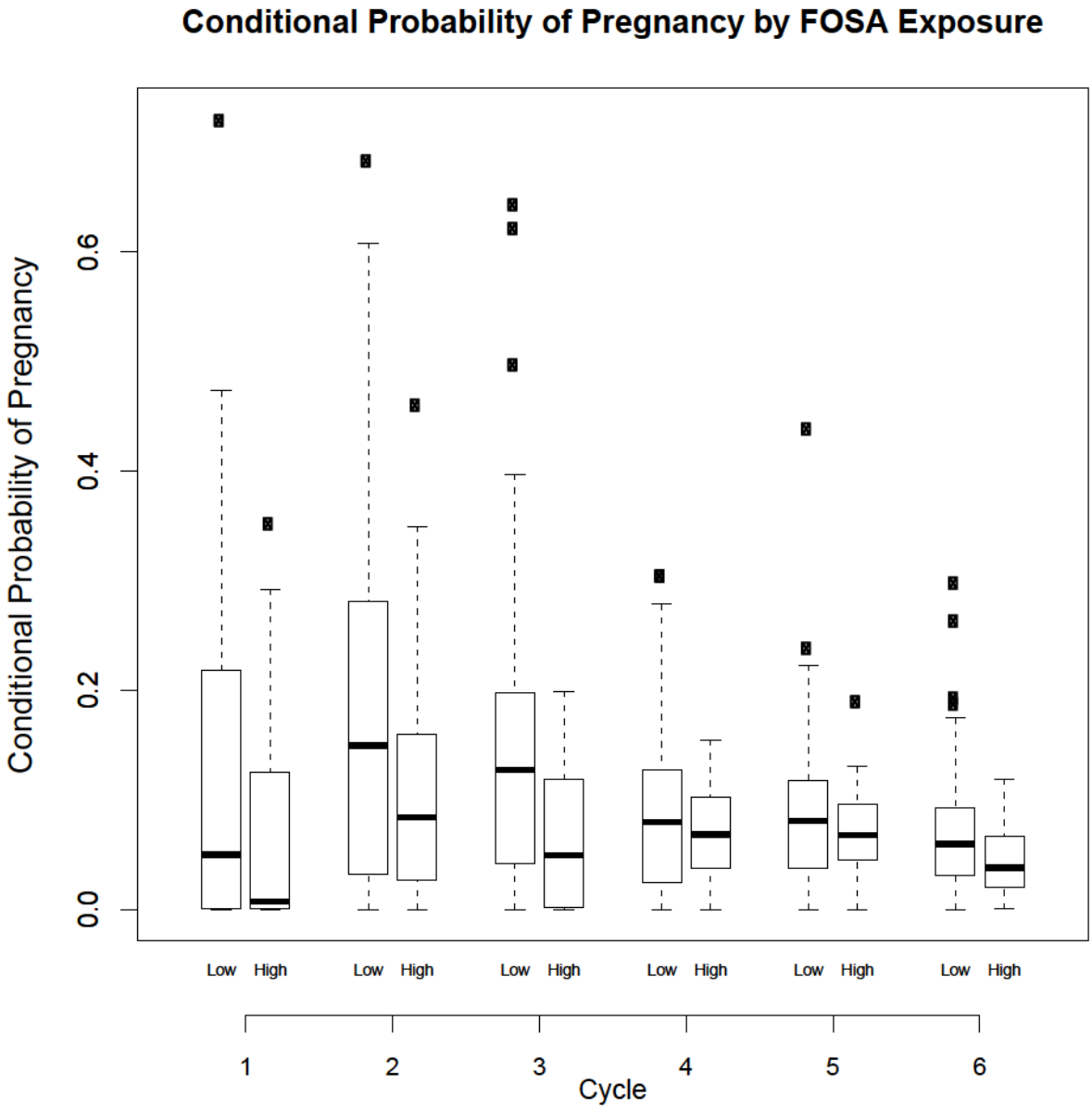


Figure 4.2: Box plots of conditional probability of pregnancy in a cycle by PFOSA exposure level (Low='<0.1ng/mL', High='$\geq$ 0.1ng/mL') conditional on no pregnancy in previous cycles, intercourse history, menstrual cycle length (days), menstrual cycle length (squared, days2), female age (years), female body mass index (categorical), female active smoking at enrollment (yes/no). Only the first six cycles are shown.

Table 4.7: Two stage model: unadjusted and adjusted OR and 95%CI (displayed as: $_{\text{Lo}}\text{OR}_{\text{Up}}$ (Louis and Zeger 2009)) for direct association of change in PFAS concentrations with day-specific probability of pregnancy, LIFE Study, 2005-2009. For models with categorized PFAS concentrations, the reference category is that with the lowest concentrations. For models with continuous PFAS concentrations, the interpretation is per IQR change in concentration. Adjusted for menstrual cycle length (days), menstrual cycle length (squared, days^2), female age (years), body mass index (categorical), and active smoking at enrollment (yes/no).

	Pregnancy Submodel			
	Single PFAS Model		Multiple PFAS Model	
	Unadjusted	Adjusted	Unadjusted	Adjusted
PFAS (ng/mL)	$_{\text{Lo}}\text{OR}_{\text{Up}}$	$_{\text{Lo}}\text{OR}_{\text{Up}}$	$_{\text{Lo}}\text{OR}_{\text{Up}}$	$_{\text{Lo}}\text{OR}_{\text{Up}}$
EtFOSAA				
≥ 0.055	$_{0.64}0.85_{1.13}$	$_{0.57}0.76_{0.99}$	$_{0.56}0.76_{1.04}$	$_{0.59}0.81_{1.08}$
Continuous	$_{0.33}0.80_{1.02}$	$_{0.33}0.82_{1.30}$	$_{0.31}0.78_{1.28}$	$_{0.33}0.84_{1.33}$
MeFOSAA				
0.02-0.03	$_{0.68}0.92_{1.26}$	$_{0.68}0.98_{1.30}$	$_{0.77}1.06_{1.48}$	$_{0.75}1.08_{1.47}$
≥ 0.04	$_{0.59}0.79_{1.08}$	$_{0.65}0.89_{1.19}$	$_{0.72}1.02_{1.02}$	$_{0.71}1.06_{1.47}$
Continuous (scaled by IQR)	$_{0.87}0.96_{1.04}$	$_{0.91}1.00_{1.09}$	$_{0.93}1.03_{1.12}$	$_{0.94}1.04_{1.13}$
PFDA				
0.03-0.04	$_{0.63}0.86_{1.18}$	$_{0.51}0.73_{1.00}$	$_{0.56}0.83_{1.24}$	$_{0.52}0.84_{1.22}$
≥ 0.05	$_{0.80}0.93_{1.47}$	$_{0.72}1.01_{1.36}$	$_{0.81}1.34_{2.24}$	$_{0.74}1.35_{2.13}$
Continuous (scaled by IQR)	$_{0.84}0.96_{1.08}$	$_{0.84}0.96_{1.09}$	$_{0.84}1.01_{1.20}$	$_{0.85}1.03_{1.21}$
PFNA				
0.09-1.4	$_{0.55}0.74_{1.01}$	$_{0.54}0.75_{1.01}$	$_{0.44}0.66_{1.00}$	$_{0.41}0.64_{0.95}$
≥ 1.5	$_{0.75}0.98_{1.29}$	$_{0.69}0.96_{1.26}$	$_{0.43}0.72_{1.18}$	$_{0.38}0.72_{1.13}$
Continuous (scaled by IQR	$_{0.83}0.94_{1.06}$	$_{0.83}0.94_{1.06}$	$_{0.76}0.95_{1.14}$	$_{0.77}0.94_{1.11}$
PFOSA				
≥ 0.1	$_{0.32}0.51_{0.76}$	$_{0.34}0.56_{0.81}$	$_{0.29}0.47_{0.75}$	$_{0.28}0.48_{0.73}$
Continuous	$_{0.14}0.19_{0.59}$	$_{0.14}0.20_{0.54}$	$_{0.14}0.18_{0.52}$	$_{0.14}0.19_{0.47}$
PFOS				
9.5-15.1	$_{0.65}0.89_{1.20}$	$_{0.62}0.88_{1.15}$	$_{0.82}1.02_{1.69}$	$_{0.80}1.20_{1.69}$
≥ 15.2	$_{0.69}0.92_{1.23}$	$_{0.68}0.93_{1.24}$	$_{0.80}1.21_{1.82}$	$_{0.77}1.20_{1.76}$
Continuous (scaled by IQR)	$_{0.85}0.97_{1.09}$	$_{0.85}0.97_{1.08}$	$_{0.84}1.02_{1.17}$	$_{0.85}1.02_{1.19}$
PFOA				
2.5-4.1	$_{0.74}1.00_{1.35}$	$_{0.75}1.06_{1.39}$	$_{0.71}1.01_{1.45}$	$_{0.70}1.05_{1.47}$
≥ 4.2	$_{0.66}0.89_{1.19}$	$_{0.62}0.86_{1.13}$	$_{0.52}0.76_{1.13}$	$_{0.47}0.75_{1.10}$
Continuous (scaled by IQR)	$_{0.79}0.93_{1.08}$	$_{0.78}0.93_{1.08}$	$_{0.77}0.95_{1.18}$	$_{0.75}0.94_{1.15}$
Menstrual cycle length (days)	$_{1.02}1.16_{1.33}$	$_{0.99}1.13_{1.31}$	$_{1.00}1.15_{1.32}$	$_{1.01}1.17_{1.36}$
Menstrual cycle length (days^2)	$_{0.73}0.81_{0.89}$	$_{0.75}0.83_{0.91}$	$_{0.74}0.82_{0.90}$	$_{0.75}0.83_{0.92}$

Table 4.8: Joint model: unadjusted and adjusted OR and 95%CI (displayed as: $_{\text{Lo}}\text{OR}_{\text{Up}}$ (Louis and Zeger 2009)) for direct association of change in PFAS concentrations with day-specific probability of pregnancy, LIFE Study, 2005-2009. For models with categorized PFAS concentrations, the reference category is that with the lowest concentrations. For models with continuous PFAS concentrations, the interpretation is per IQR change in concentration. Adjusted for menstrual cycle length (days), menstrual cycle length (squared, days^2), female age (years), body mass index (categorical), and active smoking at enrollment (yes/no).

	Pregnancy Submodel			
	Single PFAS Model		Multiple PFAS Model	
	Unadjusted	Adjusted	Unadjusted	Adjusted
PFAS (ng/mL)	$_{\text{Lo}}\text{OR}_{\text{Up}}$	$_{\text{Lo}}\text{OR}_{\text{Up}}$	$_{\text{Lo}}\text{OR}_{\text{Up}}$	$_{\text{Lo}}\text{OR}_{\text{Up}}$
EtFOSAA				
$\geq$ 0.055	$_{0.55}0.83_{1.17}$	$_{0.60}0.88_{1.29}$	$_{0.59}0.87_{1.30}$	$_{0.57}0.91_{1.32}$
Continuous	$_{0.40}0.84_{1.55}$	$_{0.42}0.87_{1.62}$	$_{0.42}0.88_{1.55}$	$_{0.40}0.89_{1.45}$
MeFOSAA				
0.02-0.03	$_{0.69}1.06_{1.57}$	$_{0.74}1.09_{1.68}$	$_{0.72}1.09_{1.66}$	$_{0.70}1.12_{1.63}$
$\geq$ 0.04	$_{0.61}0.93_{1.35}$	$_{0.65}0.98_{1.44}$	$_{0.65}1.01_{1.56}$	$_{0.62}1.03_{1.53}$
Continuous (scaled by IQR)	$_{0.89}1.00_{1.13}$	$_{0.91}1.02_{1.14}$	$_{0.90}1.02_{1.14}$	$_{0.92}1.04_{1.16}$
PFDA				
0.03-0.04	$_{0.46}0.73_{1.06}$	$_{0.47}0.71_{1.06}$	$_{0.55}0.92_{1.55}$	$_{0.49}0.89_{1.40}$
$\geq$ 0.05	$_{0.55}0.86_{1.23}$	$_{0.60}0.90_{1.34}$	$_{0.72}1.37_{2.59}$	$_{0.68}1.40_{2.47}$
Continuous (scaled by IQR)	$_{0.75}0.89_{1.05}$	$_{0.74}0.89_{1.06}$	$_{0.76}0.97_{1.31}$	$_{0.74}0.96_{1.18}$
PFNA				
0.09-1.4	$_{0.45}0.69_{1.02}$	$_{0.46}0.69_{1.03}$	$_{0.38}0.64_{1.09}$	$_{0.35}0.64_{1.00}$
$\geq$ 1.5	$_{0.53}0.78_{1.08}$	$_{0.56}0.81_{1.15}$	$_{0.37}0.68_{1.30}$	$_{0.32}0.67_{1.14}$
Continuous (scaled by IQR)	$_{0.75}0.87_{1.02}$	$_{0.75}0.88_{1.02}$	$_{0.72}0.91_{1.12}$	$_{0.71}0.90_{1.08}$
PFOSA				
$\geq$ 0.1	$_{0.36}0.64_{1.15}$	$_{0.35}0.63_{1.12}$	$_{0.33}0.61_{1.12}$	$_{0.27}0.59_{1.03}$
Continuous	$_{0.14}0.32_{2.64}$	$_{0.14}0.31_{2.63}$	$_{0.14}0.28_{3.23}$	$_{0.14}0.28_{1.93}$
PFOS				
9.5-15.1	$_{0.63}1.00_{1.42}$	$_{0.69}1.01_{1.52}$	$_{0.80}1.25_{1.98}$	$_{0.76}1.24_{1.84}$
$\geq$ 15.2	$_{0.57}0.86_{1.20}$	$_{0.61}0.90_{1.30}$	$_{0.69}1.14_{1.91}$	$_{0.65}1.14_{1.82}$
Continuous (scaled by IQR)	$_{0.80}0.94_{1.10}$	$_{0.81}0.95_{1.11}$	$_{0.86}1.05_{1.26}$	$_{0.86}1.06_{1.24}$
PFOA				
2.5-4.1	$_{0.60}0.94_{1.36}$	$_{0.68}1.00_{1.51}$	$_{0.64}0.72_{1.64}$	$_{0.65}1.11_{1.69}$
$\geq$ 4.2	$_{0.47}0.69_{0.98}$	$_{0.50}0.73_{1.06}$	$_{0.44}1.04_{1.18}$	$_{0.42}0.75_{1.19}$
Continuous (scaled by IQR)	$_{0.70}0.85_{1.03}$	$_{0.71}0.86_{1.06}$	$_{0.72}0.92_{1.18}$	$_{0.75}0.95_{1.17}$
Menstrual cycle length (days)	$_{1.10}1.53_{2.12}$	$_{0.99}1.40_{1.92}$	$_{1.08}1.52_{2.09}$	$_{1.10}1.54_{2.17}$
Menstrual cycle length (days^2)	$_{0.53}0.68_{0.84}$	$_{0.56}0.71_{0.87}$	$_{0.55}0.73_{0.88}$	$_{0.57}0.76_{0.94}$

Table 4.9: Proportional contribution of PFASs to each principal component, LIFE Study, 2005-2009.

	Principal Component						
PFAS (ng/mL)	1st	2nd	3rd	4th	5th	6th	7th
EtFOSAA	0.03	0.03	0.84	0.00	0.01	0.01	0.00
MeFOSAA	0.06	0.36	0.05	0.25	0.20	0.08	0.06
PFDA	0.23	0.07	0.04	0.13	0.02	0.26	0.32
PFNA	0.24	0.05	0.03	0.01	0.05	0.14	0.43
PFOSA	0.01	0.37	0.03	0.20	0.26	0.04	0.06
PFOS	0.22	0.07	0.02	0.15	0.16	0.34	0.06
PFOA	0.21	0.05	0.00	0.25	0.30	0.12	0.06
Cumulative Proportion of Variance	0.39	0.61	0.75	0.83	0.90	0.96	1.00

Table 4.10: Joint model principal components regression: unadjusted and adjusted AF and 95%CI (displayed as: $_{\text{Lo}}\text{AF}_{\text{Up}}$ (Louis and Zeger 2009)) for association of principal component with menstrual cycle length, LIFE Study, 2005-2009. Adjusted for female age (years), body mass index (categorical), and active smoking at enrollment (yes/no).

	Menstrual Cycle Length Submodel	
	Unadjusted	Adjusted
Principal Component	$_{\text{Lo}}\text{AF}_{\text{Up}}$	$_{\text{Lo}}\text{AF}_{\text{Up}}$
1st	$_{0.99}1.00_{1.00}$	$_{0.99}1.00_{1.01}$
2nd	$_{0.99}1.00_{1.00}$	$_{0.99}1.00_{1.00}$
3rd	$_{0.99}1.00_{1.01}$	$_{0.99}1.00_{1.01}$

Table 4.11: Joint model principal components regression: unadjusted and adjusted OR and 95%CI (displayed as: $_{\text{Lo}}\text{OR}_{\text{Up}}$ (Louis and Zeger 2009)) for association of principal component with day-specific probability of pregnancy, LIFE Study, 2005-2009. Adjusted for menstrual cycle length (days), menstrual cycle length (squared, days^2), female age (years), body mass index (categorical), and active smoking at enrollment (yes/no).

	Pregnancy Submodel	
	Unadjusted	Adjusted
Principal Component	$_{\text{Lo}}\text{OR}_{\text{Up}}$	$_{\text{Lo}}\text{OR}_{\text{Up}}$
1st	$_{0.84}0.92_{1.01}$	$_{0.84}0.93_{1.02}$
2nd	$_{0.82}0.95_{1.10}$	$_{0.83}0.97_{1.13}$
3rd	$_{0.80}0.98_{1.11}$	$_{0.80}0.98_{1.14}$

4.4. Discussion

Based on the LIFE Study, a prospective cohort of couples enrolled prior to conception and followed for up to one year while attempting to become pregnant, we found that select PFASs may be adversely associated with menstrual cycle length. In particular, increased concentrations of PFOA and MeFOSAA were associated with decreases in cycle length while increased concentration of PFDA was associated with increases in cycle length. Moreover, we observed an adverse association between short and long menstrual cycle and the probability of pregnancy. This finding underscores the importance of menstrual cycle length in evaluating the role of PFAS in fecundity. Additionally, we found that the PFASs associated with the probability of pregnancy (PFOSA and PFNA) were different from the ones associated with menstrual cycle length suggesting there are multiple roles in which PFASs may serve in relation to fecundity. It is important to note that only 11% of PFOSA concentrations were above the LOD, potentially reflective of ceased production of PFOSA in the U.S. in 2002.

In a recent study of PFOS, PFOA and menstrual cycle characteristics, Lyngsø et al. (2013) found that PFOA was associated with long mean cycle length ($\geq$32 days) compared to normal mean cycle length (25-31 days). Our findings also suggest a negative role for PFOA though in terms of shortening cycle length. However, comparison of these two findings is impeded by important differences between the INUENDO cohort and the LIFE Study, including antenatal enrollment and quantification of PFOS and PFOA and retrospective ascertainment of menstrual cycle length in the INUENDO study. In the LIFE Study, blood for PFAS quantification was obtained at preconception enrollment and data on menstrual cycle length was collected prospectively and modeled as a continuous outcome. Furthermore, the level of concentrations of PFOA found in the LIFE Study is slightly higher than that found in the INUENDO study.

Recent analyses of LIFE Study data and fecundity with respect to time-to-pregnancy and semen quality parameters are also suggestive of a role for PFOSA and PFNA in fecundity. In an analysis of male and female exposure to POPs and couple fecundity as measured

by time-to-pregnancy, increased concentration of PFOSA was associated with reduced fecundity (Buck Louis et al. 2012). Moreover, in an assessment of the relation between male PFASs and semen quality, increased concentration of PFOSA was associated with smaller sperm head area and perimeter, and a higher percentage of bicephalic and immature sperm while increased concentration of PFNA was associated with a lower percentage of sperm with coiled tails (Buck Louis et al. 2014).

In conclusion, for this prospective cohort, increased concentration of select PFASs measured preconception were associated with changes in menstrual cycle length, both in shorter and longer lengths. Other PFASs were negatively associated with the probability of pregnancy. Our findings await corroboration as investigation in the role of PFASs and the female component of fecundity has only recently begun.

Chapter 5

Concluding Remarks

We develop and apply a joint modeling approach to assess the role of female menstrual cycle length in a couple's probability of pregnancy. We address several challenges including modeling skewed distribution of menstrual cycle length accounting for length-bias and right censoring, and modeling the probability of pregnancy incorporating biological factors from both partners while also accounting for the couple's daily history of intercourse acts relative to ovulation. In a prospective study with preconception enrollment of couples trying to become pregnant, we find evidence of a quadratic relation between cycle length and the probability of pregnancy such that both short and long cycle lengths are associated with decreased pregnancy probabilities. To investigate the role of environmental chemicals, we develop a framework that allows for the assessment of time-constant exposures on the probability of pregnancy as well as those mediated through menstrual cycle length. In particular we investigate perfluoroalkyl surfactants (PFASs) and find evidence that increased concentration of select PFASs, measured in females at enrollment, are associated with changes in menstrual cycle length (both shortening and lengthening). We also found evidence that two of the PFASs are negatively associated with the couple's probability of pregnancy and these are different from the PFASs associated with changes in cycle length.

To develop a model for menstrual cycle length from prospective pregnancy data, we first need to account for the sampling process, in particular length-bias and selection effects. We develop and apply a two-stage semi-parametric approach for enrollment cycle data to

estimate the distribution of menstrual cycle length along with the shape of the enrollment probability. We evaluate our approach in simulation and find that we obtain approximately unbiased estimates of the parameters of the cycle length distribution with very little increase in variance for estimating the shape of the enrollment probability.

5.1. Implications and Future Work

The methods developed in this research are utilized for investigating the role of environmental and biological risk factors in a couple's probability of pregnancy and jointly in female menstrual cycle length. Our methods are also applicable to the study of other environmental pollutants including pesticides, phthalates, and phytoestrogens to name a few; therefore, this work may serve as a platform for collaboration between biostatisticians, epidemiologists, and environmental health scientists to develop informed and appropriate analyses of the role of environmental pollutants and fecundity. Moreover, this work will generate ideas for analyzing existing data and for improving sampling plans, study design, and data collection of future prospective pregnancy studies.

Although we apply our methods to the study of fecundity, the methods we develop are relevant to other applications as well. For example, several of the challenges addressed here are also faced in modeling human immunodeficiency virus (HIV) infection and progression to acquired immunodeficiency syndrome (AIDS). As is the case with pregnancy, a person must be at-risk in order for infection to occur; thus models of HIV infection due to intercourse with an infected partner must account for the observed history of intercourse acts and their unknown outcomes. As is also the case in sampling couples attempting pregnancy, a prevalent sampling plan is typically utilized to study the rate of progression to AIDS. Features of this sampling plan such as length-bias and selection effects on the time since HIV infection are relevant in estimating the distribution of time-to-AIDS diagnosis for the population.

In Chapter 3, we found evidence of an association between female menstrual cycle length and the couple's probability of pregnancy. As the length is actually the sum of three phases,

it would be of interest to develop a joint model of the bleeding, proliferative, and secretory phases and fecundity. The length of the proliferative phase of the menstrual cycle is the time from the first day without bleeding to the ovulation day. The range of days on which the female uses the fertility monitor to test for ovulation is limited to at most 20 (e.g., day 6 through 26 of the cycle); therefore, there is the additional challenge of interval censoring. Using the predictive distribution, one could predict the unknown ovulation day as an alternative to imputation of the average ovulation day. The length of the secretory phase is the time from ovulation day to the next bleeding event. One could develop a model to study the variability of both the proliferative and secretory phases and the relation between these variabilities and fecundity.

Much of this research is based on a Bayesian hierarchical modeling framework. An advantage of this is the ability to make predictions of pregnancy success while accounting for complex sources of uncertainty and variability. For example, given a couples biological risk factors, history of failed attempts, pattern of intercourse, and level of environmental pollutants, the joint model developed in Chapters 3- 4 could be used to obtain the posterior predictive distribution of the probability of the couple becoming pregnant, marginalized over the posterior of the unknown parameters. A web tool such as a calculator could be programmed to calculate the posterior predictive distribution and provide output and interpretation for clinicians or users. For example, based on such a prediction and his/her medical expertise, a clinician may advise the couple to continue attempting pregnancy on their own, perhaps with changes in their behavior, or s/he may suggest fertility counseling or treatment.

This dissertation contributes to the study of human fecundity, in particular couple probability of pregnancy and female menstrual cycle length. We build a framework for investigating the complex relation between these two processes along with assessment for environmental exposures on both processes. We do so by integrating methodology from the disciplines of biostatistics, epidemiology, and environmental health.

Bibliography

Aalen, O. O. and Husebye, E. (1991). Statistical analysis of repeated events forming renewal processes. *Statistics in Medicine* **10,** 1227–1240.

Armbruster, D. A. and Pry, T. (2008). Limit of blank, limit of detection and limit of quantitation. *Clin Biochem Rev* **29,** S49–S52.

Asgharian, M., M'Lan, C. E., and Wolfson, D. B. (2002). Length-biased sampling with right censoring : an unconditional approach. *Journal of the American Statistical Association* **97,** 201–209.

Axmon, A., Rylander, L., Albin, M., and Hagmar, L. (2006). Factors affecting time to pregnancy. *Human Reproduction* **21,** 1279–1284.

Bernert, J. T., Turner, W. E., Pirkle, J. L., Sosnoff, C. S., Akins, J. R., Waldrep, M. K., Ann, Q., Covey, T. R., Whitfield, W. E., Gunter, E. W., et al. (1997). Development and validation of sensitive method for determination of serum cotinine in smokers and nonsmokers by liquid chromatography/atmospheric pressure ionization tandem mass spectrometry. *Clinical Chemistry* **43,** 2281–2291.

Bortot, P., Masarotto, G., and Scarpa, B. (2010). Sequential predictions of menstrual cycle lengths. *Biostatistics* **11,** 741–755.

Brookmeyer, R. and Gail, M. H. (1987). Biases in prevalent cohorts. *Biometrics* **43,** 739–749.

Brown, E. R., Ibrahim, J. G., and DeGruttola, V. (2005). A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics* **61,** 64–73.

Buck Louis, G., Chen, Z., Schisterman, E., Kim, S., Sweeney, A., Sundaram, R., Lynch, C., Gore-Langton, R., and Barr, D. (2014). Perfluorochemicals and human semen quality: The life study. *Environmental health perspectives* .

Buck Louis, G., Sundaram, R., Schisterman, E. F., Sweeney, A. M., Lynch, C. D., Gore-Langton, R. E., Maisog, J., Kim, S., Chen, Z., and Barr, D. B. (2012). Persistent environmental pollutants and couple fecundity: the life study. *Environmental health perspectives* **121,** 231–236.

Buck Louis, G. M. (2011). Fecundity and fertility. In Buck Louis, G. M. and Platt, R. W., editors, *Reproductive and Perinatal Epidemiology*. Oxford University Press., New York.

Buck Louis, G. M., Rios, L. I., McLain, A., Cooney, M. A., Kostyniak, P. J., and Sundaram, R. (2011). Persistent organochlorine pollutants and menstrual cycle characteristics. *Chemosphere* **85,** 1742–1748.

Buck Louis, G. M., Schisterman, E. F., Sweeney, A. M., Wilcosky, T. C., Gore-Langton, R. E., Lynch, C. D., Boyd Barr, D., Schrader, S. M., Kim, S., Chen, Z., Sundaram, R., and on behalf of the LIFE Study (2011). Designing Prospective Cohort Studies for Assessing Reproductive and Developmental Toxicity During Sensitive Windows of Human Reproduction and Development - the LIFE Study. *Paediatric and Perinatal Epidemiology* **25,** 413–424.

Buck Louis, G. M., Sundaram, R., Schisterman, E. F., Sweeney, A., Lynch, C. D., Kim, S., Maisog, J. M., Gore-Langton, R., Eisenberg, M. L., and Chen, Z. (2014). Semen Quality and Time to Pregnancy: the longitudinal investigation of fertility and the environment study. *Fertility and Sterility* **101,** 453–462.

Carlin, B. P. and Louis, T. A. (2009). *Bayesian Methods for Data Analysis, 3rd edition*. Chapman and Hall/CRC Press, Boca Raton, FL, 3rd edition.

Cole, L. A. (2011). The utility of six over-the-counter (home) pregnancy tests. *Clinical Chemistry and Laboratory Medicine* **49,** 1317–1322.

Conway, J. B. (1990). *A Course in Functional Analysis*, volume 96. Springer.

Cooper, G. and Sandler, D. (1997). Long-term effects of reproductive-age menstrual cycle patterns on peri-and postmenopausal fracture risk. *American journal of epidemiology* **145,** 804–809.

Cooper, T. G., Noonan, E., von Eckardstein, S., Auger, J., Gordon Baker, H. W., Behre, H. M., Haugen, T. B., Kruger, T., Wang, C., Mbizvo, M. T., and Vogelsong, K. M. (2010). World health organization Reference Values for Human Semen Characteristics. *Human Reproduction Update* **16,** 231–245.

Cox, D. R. (1962). *Renewal Theory*. Methuen, The University of Michigan.

Cox, D. R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling*. John Wiley and Sons, Inc., New York.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford University Press, New York.

Dominik, R. and Chen, P.-L. (2006). Day-specific pregnancy probability estimation in barrier contraceptive effectiveness trials. *Paediatric and Perinatal Epidemiology* **20,** 38–42.

Dominik, R., Zhou, H., and Cai, J. (2001). A statistical model for the evaluation of barrier contraceptive efficacy. *Statistics in Medicine* **20,** 3279–3294.

Dunson, D. B. and Stanford, J. B. (2005). Bayesian inferences on predictors of conception probabilities. *Biometrics* **61,** 126–133.

Ecochard, R. (2006). Heterogeneity in fecundability studies: Issues and modelling. *Statistical Methods in Medical Research* **15,** 141–160.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association* **82,** 171–185.

Eunice Kennedy Shriver National Institute of Child Health and Human Development (2011). NICHD's scientific vision: The next decade. `http://www.nichd.nih.gov/vision/vision_themes/index.cfm`.

Fehring, R. J., Schneider, M., and Raviele, K. (2006). Variability in the phases of the menstrual cycle. *Journal of Obstetric, Gynecologic, & Neonatal Nursing* **35,** 376–384.

Fei, C., McLaughlin, J. K., Lipworth, L., and Olsen, J. (2009). Maternal levels of perfluorinated chemicals and subfecundity. *Human Reproduction* .

Feller, W. (1971). *An Introduction to Probability Theory and its Application*, volume 2. Wiley, The University of California.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis* **1,** 515–534.

Giesy, J. P. and Kannan, K. (2002). Peer reviewed: perfluorochemical surfactants in the environment. *Environmental science & technology* **36,** 146A–152A.

Gill, R. D. and Keiding, N. (2010). Product-limit estimators of the gap time distribution of a renewal process under different sampling patterns. *Lifetime Data Analysis* **16,** 571–579.

Guo, Y., Manatunga, A. K., Chen, S., and Marcus, M. (2006). Modeling menstrual cycle length using a mixture distribution. *Biostatistics* **7,** 100–114.

Harlow, S. D., Lin, X., and Ho, M. J. (2000). Analysis of Menstrual Diary Data Across the Reproductive Life Span Applicability of the Bipartite Model Approach and the Importance of Within-woman Variance. *Journal of Clinical Epidemiology* **53,** 722–733.

Harlow, S. D. and Zeger, S. L. (1991). An application of longitudinal methods to the analysis of menstrual diary data. *Journal of Clinical Epidemiology* **44,** 1015 – 1025.

Huang, C.-Y. and Qin, J. (2011). Nonparametric estimation for length-biased and right-censored data. *Biometrika* **98,** 177–186.

Jensen, T., Scheike, T., Keiding, N., Schaumburg, I., and Grandjean, P. (1999a). Fecundability in relation to body mass and menstrual cycle patterns. *Epidemiology* pages 422–428.

Jensen, T. K., Scheike, T., Keiding, N., Schaumburg, I., and Grandjean, P. (1999b). Fecundability in relation to body mass and menstrual cycle patterns. *Epidemiology* pages 422–428.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New Jersey, 2nd edition.

Kannan, K., Corsolini, S., Falandysz, J., Fillmann, G., Kumar, K. S., Loganathan, B. G., Mohd, M. A., Olivero, J., Wouwe, N. V., Yang, J. H., et al. (2004). Perfluorooctanesulfonate and related fluorochemicals in human blood from several countries. *Environmental science & technology* **38,** 4489–4495.

Kato, K., Wong, L.-Y., Jia, L. T., Kuklenyik, Z., and Calafat, A. M. (2011). Trends in exposure to polyfluoroalkyl chemicals in the us population: 1999- 2008. *Environmental science & technology* **45,** 8037–8045.

Keiding, N. and Gill, R. D. (1990). Random truncation models and markov processes. *The Annals of Statistics* **18,** 582–602.

Keiding, N., Hansen, O. K. H., Sorensen, D. N., and Slama, R. (2012). The current duration approach to estimating time to pregnancy. *Scandinavian Journal of Statistics* **39,** 185–204.

Keiding, N., Kvist, K., Hartvig, H., Tvede, M., and Juul, S. (2002). Estimating time to pregnancy from current durations in a cross-sectional sample. *Biostatistics* **3,** 565–578.

Keulers, M., Hamilton, C., Franx, A., Evers, J., and Bots, R. (2007). The length of the fertile

window is associated with the chance of spontaneously conceiving an ongoing pregnancy in subfertile couples. *Human Reproduction* **22,** 1652–1656.

Kolstad, H. A., Bonde, J. P., Hjøllund, N. H., Jensen, T. K., Henriksen, T. B., Ernst, E., Giwercman, A., Skakkebæk, N. E., and Olsen, J. (1999). Menstrual cycle pattern and fertility: a prospective follow-up study of pregnancy and early embryonal loss in 295 couples who were planning their first pregnancy. *Fertility and sterility* **71,** 490–496.

Kuklenyik, Z., Needham, L. L., and Calafat, A. M. (2005). Measurement of 18 perfluorinated organic acids and amides in human serum using on-line solid-phase extraction. *Analytical chemistry* **77,** 6085–6091.

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

Lin, X., Raz, J., and Harlow, S. D. (1997). Linear mixed models with heterogeneous within-cluster variances. *Biometrics* **53,** 910–923.

Louis, T. A. and Zeger, S. L. (2009). Effective Communication of Standard Errors and Confidence Intervals. *Biostatistics* **10,** 1–2.

Lum, K. J., Sundaram, R., and Louis, T. A. (2014). Accounting for length-bias and selection effects in estimating the distribution of menstrual cycle length. *Biostatistics* .

Lynch, C. D., Jackson, L. W., and Buck Louis, G. M. (2006). Estimation of the Day-specific Probabilities of Conception: Current State of the Knowledge and the Relevance for Epidemiological Research. *Paediatric and Perinatal Epidemiology* **20,** 3–12.

Lyngsø, J., Ramlau-Hansen, C., Høyer, B. B., Støvring, H., Bonde, J., Jönsson, B., Lindh, C., Pedersen, H., Ludwicki, J., Zviezdai, V., et al. (2013). Menstrual cycle characteristics in fertile women from greenland, poland and ukraine exposed to perfluorinated chemicals: a cross-sectional study. *Human Reproduction* page det390.

McLain, A. C., Lum, K. J., and Sundaram, R. (2012). A Joint Mixed Effects Dispersion Model for Menstrual Cycle Length and Time-to-pregnancy. *Biometrics* **68,** 648–656.

Murphy, S. A., Bentley, G. R., and O'Hanesian, M. A. (1995). An analysis for menstrual data with time-varying covariates. *Statistics in Medicine* **14,** 1843–1857.

Norwitz, E. R., Schust, D. J., and Fisher, S. J. (2001). Implantation and the survival of early pregnancy. *New England Journal of Medicine* **345,** 1400–1408.

Olsen, G. W., Lange, C. C., Ellefson, M. E., Mair, D. C., Church, T. R., Goldberg, C. L., Herron, R. M., Medhdizadehkashi, Z., Nobiletti, J. B., Rios, J. A., et al. (2012). Temporal trends of perfluoroalkyl concentrations in american red cross adult blood donors, 2000–2010. *Environmental science & technology* **46,** 6330–6338.

Pirkle, J. L., Flegal, K. M., Bernert, J. T., Brody, D. J., Etzel, R. A., and Maurer, K. R. (1996). Exposure of the us population to environmental tobacco smoke: the third national health and nutrition examination survey, 1988 to 1991. *The Journal of the American Medical Association* **275,** 1233–1240.

Ramlau-Hansen, C., Thulstrup, A. M., Nohr, E., Bonde, J. P., Sørensen, T., and Olsen, J. (2007). Subfecundity in overweight and obese couples. *Human reproduction* **22,** 1634–1637.

Raymer, J. H., Michael, L. C., Studabaker, W. B., Olsen, G. W., Sloan, C. S., Wilcosky, T., and Walmer, D. K. (2012). Concentrations of perfluorooctane sulfonate (pfos) and perfluorooctanoate (pfoa) and their associations with human semen quality measurements. *Reproductive Toxicology* **33,** 419–427.

Richardson, D. B. and Ciampi, A. (2003). Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *American Journal of Epidemiology* **157,** 355–363.

Rothmann, S. A., Bort, A.-M., Quigley, J., and Pillow, R. (2013). Sperm Morphology Classification: a Rational Method for Schemes Adopted by the world health organization. In Carrell, D. T. and Aston, K. L., editors, *Spermatogenesis: Methods and Protocols*, pages 27–38. Humana Press, New York.

Royston, P. and Ferreira, A. (1999). A new approach to modeling daily probabilities of conception. *Biometrics* **55,** 1005–1013.

Schafer, J. L. (2010). *Analysis of incomplete multivariate data*. CRC press.

Schisterman, E. F., Vexler, A., Whitcomb, B. W., and Liu, A. (2006). The limitations due to exposure detection limits for regression models. *American Journal of Epidemiology* **163,** 374–383.

Small, C., Manatunga, A., Klein, M., Dominguez, C., Feigelson, H., McChesney, R., and Marcus, M. (2010). Menstrual cycle variability and the likelihood of achieving pregnancy. *Reviews on environmental health* **25,** 369–378.

Small, C., Manatunga, A., and Marcus, M. (2007). Validity of self-reported menstrual cycle length. *Annals of epidemiology* **17,** 163–170.

Small, C. M., Manatunga, A. K., Klein, M., Feigelson, H. S., Dominguez, C. E., McChesney, R., and Marcus, M. (2006). Menstrual Cycle Characteristics: Associations With Fertility and Spontaneous Abortion. *Epidemiology* **17,** 52–60.

Thomas, A., OHara, R., Ligges, U., and Sturtz, S. (2006). Making bugs open. *R News* **6,** 12–17.

Tingen, C., Stanford, J. B., and Dunson, D. B. (2004). Methodologic and statistical approaches to studying human fertility and environmental exposure. *Environmental health perspectives* **112,** 87.

Toft, G., Jönsson, B., Lindh, C., Giwercman, A., Spano, M., Heederik, D., Lenters, V., Vermeulen, R., Rylander, L., Pedersen, H., et al. (2012). Exposure to perfluorinated compounds and human semen quality in arctic and european populations. *Human Reproduction* pages 2532–40.

Treloar, A., Boynton, R., Behn, B., Brown, B., et al. (1967). Variation of the human menstrual cycle through reproductive life. *Inter J Fertility* **12,** 77.

Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika* **76,** 751–761.

Verbeke, G. and Molenberghs, G. (2009). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Vollman, R. (1977). *The menstrual cycle, Vol. 7*. WB Saunders Company, Philadelphia, PA.

Wand, M. P. and Jones, M. C. (1994). *Kernel Smoothing*. Chapman & Hall/CRC.

Wang, M., Park, J.-S., and Petreas, M. (2011). Temporal changes in the levels of perfluorinated compounds in california women's serum over the past 50 years. *Environmental science & technology* **45,** 7510–7516.

Wang, M.-C. (1989). A semiparametric model for randomly truncated data. *Journal of the American Statistical Association* **84,** 742–748.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association* **86,** 130–143.

Wang, M.-C., Jewell, N. P., and Tsai, W. Y. (1986). Asymptotic properties of the product limit estimate under random truncation. *The Annals of Statistics* **14,** 1597–1605.

Weinberg, C. R., Baird, D. D., and Wilcox, A. J. (1994). Sources of bias in studies of time to pregnancy. *Statistics in medicine* **13,** 671–681.

Weinberg, C. R., Gladen, B. C., and Wilcox, A. J. (1994). Models relating the timing of intercourse to the probability of conception and the sex of the baby. *Biometrics* **50,** 358–367.

Whelan, E. A., Sandler, D. P., Root, J. L., Smith, K. R., and Weinberg, C. R. (1994). Menstrual cycle patterns and risk of breast cancer. *American journal of epidemiology* **140,** 1081–1090.

Winter, B. B. (1989). Joint simulation of backward and forward recurrence times in a renewal process. *Journal of Applied Probability* **26,** 404–407.

Winter, B. B. and Foldes, A. (1988). A product-limit estimator for use with length-biased data. *Canadian Journal of Statistics* **16,** 337–355.

World Health Organization (1992). *WHO Laboratory Manual for the Examination of Human Semen and Sperm-cervical Mucus Interaction.* Cambridge University Press, Cambridge, United Kingdom, 3rd edition.

Zaadstra, B. M., Seidell, J. C., Van Noord, P. A., te Velde, E. R., Habbema, J. D., Vrieswijk, B., and Karbaat, J. (1993). Fat and female fecundity: prospective study of effect of body fat distribution on conception rates. *BMJ* **306,** 484–487.

Appendix A

Guidance and Code for Simulations and Analyses

In this chapter, we provide code for the simulation and analyses in this dissertation and some guidance on computational challenges.

A.1. Steps for Simulating Data

In Section 2.6 of the manuscript, we describe a simulation study of the performance of the proposed semi-parametric estimation approach for estimating the parameters of the target distribution. Here we describe steps for generating the simulated data.

In each sample, the data for the i-th individual are generated according to the following plan:

1. Generate $\tilde{Y}_{0i}$ from the normal-Gumbel mixture distribution with density $\tilde{Y}_0 \sim qf_{normal}(\mu_N, \sigma_N^2) + (1-q)f_{Gumbel}(\mu_G, \sigma_G^2)$, with $\theta = (q = 0.65, \mu_N = 29.5, \sigma_N = 0.75, \mu_G = 36.5, \sigma_G = 14.3)$, where q is the proportion of normally distributed lengths and (μ_N, σ_N) and (μ_G, σ_G) are the mean and standard deviation of the normal and Gumbel distributions, respectively.

2. Generate covariate observation $Z_i \sim Bern(p_c)$ with $p_c = 0.5$.

3. Generate random effect $W_i \sim Gamma(\alpha, 1/\alpha)$, so $E(W) = 1$ and $SD(W) = \sqrt{1/\alpha} = 0.10$.

4. Given $Z_i = z_i$ and $W_i = w_i$, generate $Y_i = Y_{0i} w_i e^{\beta z_i}$ with $\beta = -0.05$.

5. Generate censoring time $C_i \sim exponential(\varphi)$ with φ chosen to result in an overall censoring percentage of 13%.

6. Set $X_i = Y_i \wedge C_i$.

7. Given $X_i = x_i$, draw γ_i from $[\gamma_i \mid x_i] \sim \mathcal{B}ern(p_{lb})$ with $p_{lb} = \frac{x_i}{k}$ where k is a constant chosen so that $0 \leq p_{lb} \leq 1$. If γ_i=0, discard X_i and start a new observation at step 1. If $\gamma_i = 1$, keep X_i and go on to the next step.

8. Given $X_i = x_i$, generate backward recurrence time B_i using the method by Winter (1989).

9. Enroll (X_i, δ_i, B_i) with probability: $\pi(B_i) = \exp(\zeta B_i/100)$, $\zeta = -5.5$.

Additionally, to study the case of a constant $\pi(\cdot)$, this plan is repeated changing ζ in step 9 to $\zeta = 0$.

A.2. R code for Estimation of Menstrual Cycle Length Distribution Accounting for Length-bias and Selection Bias from Simulated Data

The following R (R Development Core Team) computer code can be used to estimate the distribution of menstrual cycle length and the probability of enrollment for the simulated data (see Section 2.6).

```
#Purpose:  Estimate probability of enrollment as a function of
            backward recurrence time. Use this in the full likelihood
            to estimate the menstrual cycle length distribution.
#Note:      This code is for data from simulations.


#Define functions
ParamNonsmooth<-function(data,indices=1:(dim(data)[1])){


#Function to calculate log likelihood (conditional likelihood)
condLikgaussquadprob = function(arg,nnodes){
    re.nodenwt=gauss.quad.prob(nnodes,dist="gamma",alpha=arg[6],beta=1/arg[6])
fy0marg=sapply(1:nsub,function(i){dfmargGaussQuadProbwCov(x=ys[i],funct='pdf',
  nodesnwt = re.nodenwt, pmix.in=arg[1],
  norm.mean.in=arg[2],norm.sd.in=arg[3],
  gumbel.mean.in=arg[4],gumbel.sd.in=arg[5],
  beta.slope=arg[7], z.cov=covs[i])})
Sy0marg.ys=sapply(1:nsub,function(i){dfmargGaussQuadProbwCov(x=ys[i],funct='surv',
nodesnwt = re.nodenwt, pmix.in=arg[1],
norm.mean.in=arg[2],norm.sd.in=arg[3],
gumbel.mean.in=arg[4],gumbel.sd.in=arg[5],
beta.slope=arg[7], z.cov=covs[i])})
```

```
Sy0marg.bs=sapply(1:nsub,function(i){dfmargGaussQuadProbwCov(x=bs[i],funct='surv',
nodesnwt = re.nodenwt, pmix.in=arg[1],
norm.mean.in=arg[2],norm.sd.in=arg[3],
gumbel.mean.in=arg[4],gumbel.sd.in=arg[5],
beta.slope=arg[7], z.cov=covs[i])})
loglik = sum(delta*log(fy0marg) + (1-delta)*log(Sy0marg.ys) - log(Sy0marg.bs))

if(is.finite(loglik)==F) loglik = -10^10
#output log likelihood
loglik
}
#Full Likelihood
fullLikMCint = function(arg,nnodes,n.mc){
re.nodenwt=gauss.quad.prob(nnodes,dist="gamma",alpha=arg[6],beta=1/arg[6])
fy0marg=sapply(1:nsub,function(i){dfmargGaussQuadProbwCov(x=ys[i],funct='pdf',
 nodesnwt = re.nodenwt, pmix.in=arg[1],norm.mean.in=arg[2],
 norm.sd.in=arg[3],gumbel.mean.in=arg[4],gumbel.sd.in=arg[5],
 beta.slope=arg[7], z.cov=covs[i])})
Sy0marg.ys=sapply(1:nsub,function(i){dfmargGaussQuadProbwCov(x=ys[i],funct='surv',
nodesnwt = re.nodenwt, pmix.in=arg[1],norm.mean.in=arg[2],
norm.sd.in=arg[3],gumbel.mean.in=arg[4],gumbel.sd.in=arg[5],
beta.slope=arg[7], z.cov=covs[i])})
W.draws = rgamma(n.mc,shape=arg[6],scale=(1/arg[6]))
y0.draws = fy0rand(n.mc,pmix=arg[1],norm.mean=arg[2],norm.sd=arg[3],
gumbel.mean=arg[4],gumbel.sd=arg[5])
doubleintcov0 = 1/n.mc*sum(outerf.cov0(y0.draws*W.draws))
doubleintcov1 = 1/n.mc*sum(outerf.cov1(y0.draws*W.draws*exp(arg[7])))
```

```
loglik=sum(delta*log(fy0marg) + (1-delta)*log(Sy0marg.ys) -
(1-covs)*log(doubleintcov0) - (covs)*log(doubleintcov1))

if(is.finite(loglik)==F) loglik = -10^10
#output log likelihood
return(loglik)
}

ys = data[indices,1]
bs = data[indices,2]
delta = data[indices,3]
covs = data[indices,4]
nsub=length(ys)
results = rep(NA,35)

optim.res = optim(par=initial,fn=condLikgaussquadprob, nnodes=nnodesForGQP,
method='L-BFGS-B', lower=lowerbounds, upper=upperbounds,
control=list(fnscale=-1),hessian=FALSE)
if(optim.res$convergence!=52 & optim.res$par[1]!=lowerbounds[1] &
optim.res$par[1]!=upperbounds[1] & optim.res$par[6]!=lowerbounds[6]
& optim.res$par[3]!=lowerbounds[3] & optim.res$par[5]!=lowerbounds[5]){
#Store results using conditional
results[1:7] = optim.res$par
p.est.CL = results[1]
norm.mean.est.CL = results[2]
norm.sd.est.CL = results[3]
gumbel.mean.est.CL = results[4]
```

```
gumbel.sd.est.CL = results[5]
alpha.est.CL = results[6]
beta.slope.est.CL = results[7]

#E[y0] and SD[y0]
results[8:9] = y0meanSD(pmix=p.est.CL,norm.mean=norm.mean.est.CL,
norm.sd=norm.sd.est.CL,gumbel.mean=gumbel.mean.est.CL,
gumbel.sd=gumbel.sd.est.CL)
meanY0.est.CL=results[8]
sdY0.est.CL=results[9]
#E[Y] = y0mean
results[10] = meanY0.est.CL
results[11] = meanY0.est.CL*exp(beta.slope.est.CL)
#SD[Y] = sqrt(y0sd^2*(1/alpha + 1) + y0mean^2/alpha
results[12]= sqrt(sdY0.est.CL^2*(1/alpha.est.CL + 1) +
meanY0.est.CL^2/alpha.est.CL)
results[13]= sqrt(sdY0.est.CL^2*exp(2*beta.slope.est.CL)*(1/alpha.est.CL + 1) +
meanY0.est.CL^2*exp(2*beta.slope.est.CL)/alpha.est.CL)
results[14]=sqrt(1/alpha.est.CL)

# #Get fixed estimate of f_B|R(b|R=1) (in paper, this is g_B(b)
bs.cov0 = bs[covs==0]
numuniqueb.cov0 = length(unique(bs.cov0))
  fbgivR.disc.cov0=data.frame(matrix(rep(0,2*numuniqueb.cov0),numuniqueb.cov0,2))
  names(fbgivR.disc.cov0)=c('grid','estimates')
  fbgivR.disc.cov0[,1]=unique(sort(bs.cov0))
  fbgivR.disc.cov0[,2]=sapply(1:numuniqueb.cov0,function(i){
```

```
length(bs.cov0[bs.cov0==fbgivR.disc.cov0$grid[i]])/length(bs.cov0)})

#Initialize pi
pi.est.disc.cov0=data.frame(matrix(rep(0,2*numuniqueb.cov0),numuniqueb.cov0,2))
names(pi.est.disc.cov0)=c('grid','estimates')
   pi.est.disc.cov0$grid=unique(sort(bs.cov0))
   pi.est.disc.cov0$estimates=rep(1,numuniqueb.cov0)

#Estimate S(b)
nodesandweights=gauss.quad.prob(nnodesForGQP,dist="gamma",alpha=alpha.est.CL,
beta=1/alpha.est.CL)
Survy0marg.cov0 = sapply(1:numuniqueb.cov0,
function(i){dfmargGaussQuadProbwCov(x=fbgivR.disc.cov0$grid[i],
funct='surv', nodesnwt = nodesandweights,
pmix.in=results[1],norm.mean.in=results[2],
norm.sd.in=results[3],gumbel.mean.in=results[4],
gumbel.sd.in=results[5],
beta.slope=results[7], z.cov=0)})

#Estimate pi
pi.est.disc.cov0$estimates=fbgivR.disc.cov0$estimates/Survy0marg.cov0
pi.est.disc.cov0$estimates[pi.est.disc.cov0$estimates==Inf]=
fbgivR.disc.cov0$estimates.cov0[pi.est.disc.cov0$estimates==Inf]/.001

outerf.cov0<-function(yarg) {
sapply(yarg, function(yarg) {
sum(pi.est.disc.cov0$estimates[pi.est.disc.cov0$grid<=yarg])
```

```
})
}

bs.cov1 = bs[covs==1]
numuniqueb.cov1 = length(unique(bs.cov1))

  fbgivR.disc.cov1=data.frame(matrix(rep(0,2*numuniqueb.cov1),numuniqueb.cov1,2))
  names(fbgivR.disc.cov1)=c('grid','estimates')
  fbgivR.disc.cov1[,1]=unique(sort(bs.cov1))
  fbgivR.disc.cov1[,2]=sapply(1:numuniqueb.cov1,function(i){
length(bs.cov1[bs.cov1==fbgivR.disc.cov1$grid[i]])/length(bs.cov1)})

#Initialize pi
pi.est.disc.cov1=data.frame(matrix(rep(0,2*numuniqueb.cov1),numuniqueb.cov1,2))
names(pi.est.disc.cov1)=c('grid','estimates')
   pi.est.disc.cov1$grid=unique(sort(bs.cov1))
   pi.est.disc.cov1$estimates=rep(1,numuniqueb.cov1)

#Estimate S(b)
SurvyOmarg.cov1 = sapply(1:numuniqueb.cov1,function(i){
dfmargGaussQuadProbwCov(x=fbgivR.disc.cov1$grid[i],funct='surv',
nodesnwt = nodesandweights, pmix.in=results[1],
                            norm.mean.in=results[2],
norm.sd.in=results[3],gumbel.mean.in=results[4],
                            gumbel.sd.in=results[5],
beta.slope=results[7], z.cov=1)})
```

```
#Estimate pi
pi.est.disc.cov1$estimates=fbgivR.disc.cov1$estimates/Survy0marg.cov1
pi.est.disc.cov1$estimates[pi.est.disc.cov1$estimates==Inf]
=fbgivR.disc.cov1$estimates.cov1
[pi.est.disc.cov1$estimates==Inf]/.001

outerf.cov1<-function(yarg) {
sapply(yarg, function(yarg) {
sum(pi.est.disc.cov1$estimates[pi.est.disc.cov1$grid<=yarg])
})
}

optim.resfull = optim(par=results[1:7],fn=fullLikMCint, nnodes=nnodesForGQP,
n.mc=n.montecarlo, method='L-BFGS-B',
lower=lowerbounds,upper=upperbounds,
control=list(fnscale=-1),hessian=F)
if(optim.resfull$convergence!=52 & optim.resfull$par[1]!=lowerbounds[1]
& optim.resfull$par[1]!=upperbounds[1]
& optim.resfull$par[6]!=lowerbounds[6]
&optim.resfull$par[3]!=lowerbounds[3]
& optim.resfull$par[5]!=lowerbounds[5]){

results[15:21] = optim.resfull$par
p.est.FullL = results[15]
norm.mean.est.FullL = results[16]
norm.sd.est.FullL = results[17]
gumbel.mean.est.FullL = results[18]
```

```
gumbel.sd.est.FullL = results[19]
alpha.est.FullL = results[20]
beta.slope.est.FullL = results[21]

#E[y0] and SD[y0]
results[22:23] = y0meanSD(pmix=p.est.FullL,norm.mean=norm.mean.est.FullL,
norm.sd=norm.sd.est.FullL,gumbel.mean=gumbel.mean.est.FullL,
gumbel.sd=gumbel.sd.est.FullL)
meanY0.est.FullL=results[22]
sdY0.est.FullL=results[23]

#E[Y] = y0mean
results[24] = meanY0.est.FullL
results[25] = meanY0.est.FullL*exp(beta.slope.est.FullL)

#SD[Y] = sqrt(y0sd^2*(1/alpha + 1) + y0mean^2/alpha
results[26]= sqrt(sdY0.est.FullL^2*(1/alpha.est.FullL + 1) +
meanY0.est.FullL^2/alpha.est.FullL)
results[27]= sqrt(sdY0.est.FullL^2*exp(2*beta.slope.est.FullL)*
(1/alpha.est.FullL + 1) +
meanY0.est.FullL^2*exp(2*beta.slope.est.FullL)/alpha.est.FullL)
results[28]=sqrt(1/alpha.est.FullL)
results[29] = fullLikMCint(results[1:7],nnodes=nnodesForGQP,n.mc=n.montecarlo)
        results[30] = optim.resfull$value
results[31] = optim.res$counts
results[32] = optim.resfull$counts
results[33] = 2*7 - 2*optim.resfull$value
```

```
results[34] = optim.res$convergence
results[35] = optim.resfull$convergence
}
}
return(results)
}

#For post enrollment cycles
ParamNonsmoothPostEnroll<-function(data,indices=1:(dim(data)[1])){
loglikelihoodfunc = function(arg,nnodes){
re.nodenwt<-gauss.quad.prob(nnodes,dist="gamma",alpha=arg[6],beta=1/arg[6])
fy0marg=sapply(1:nsub,function(i){dfmargGaussQuadProbwCov(x=ys[i],funct='pdf',
nodesnwt = re.nodenwt,
pmix.in=arg[1],
norm.mean.in=arg[2],norm.sd.in=arg[3],gumbel.mean.in=arg[4],gumbel.sd.in=arg[5],
beta.slope=arg[7], z.cov=covs[i])}
loglik = sum(log(fy0marg))

if(is.finite(loglik)==F) loglik = -10^10
#output log likelihood
loglik
}

ys = data[indices,1]
covs = data[indices,2]
nsub = length(ys)
results = rep(NA,14)
```

```
maxitvalue = 100
optim.res = optim(par=initial,fn=loglikelihoodfunc, nnodes=nnodesForGQP,
method='L-BFGS-B',
lower=lowerbounds, upper=upperbounds,control=list(fnscale=-1),hessian=FALSE)

if(optim.res$convergence!=52 & optim.res$par[1]!=lowerbounds[1] &
optim.res$par[1]!=upperbounds[1]  &
optim.res$par[6]!=lowerbounds[6] &
optim.res$par[3]!=lowerbounds[3] &
& optim.res$par[5]!=lowerbounds[5]){
results[1:7] = optim.res$par

#Store results using conditional
p.est.CL = results[1]
norm.mean.est.CL = results[2]
norm.sd.est.CL = results[3]
gumbel.mean.est.CL = results[4]
gumbel.sd.est.CL = results[5]
alpha.est.CL = results[6]
beta.slope.est.CL = results[7]

#E[y0] and SD[y0]
results[8:9] = y0meanSD(pmix=p.est.CL,norm.mean=norm.mean.est.CL,
norm.sd=norm.sd.est.CL,gumbel.mean=gumbel.mean.est.CL,
gumbel.sd=gumbel.sd.est.CL)
meanY0.est.CL=results[8]
sdY0.est.CL=results[9]
```

```
#E[Y] = y0mean
results[10] = meanY0.est.CL
results[11] = meanY0.est.CL*exp(beta.slope.est.CL)

#SD[Y] = sqrt(y0sd^2*(1/alpha + 1) + y0mean^2/alpha
results[12]= sqrt(sdY0.est.CL^2*(1/alpha.est.CL + 1)+meanY0.est.CL^2/alpha.est.CL)
results[13]= sqrt(sdY0.est.CL^2*exp(2*beta.slope.est.CL)*(1/alpha.est.CL + 1) +
meanY0.est.CL^2*exp(2*beta.slope.est.CL)/alpha.est.CL)
results[14]=sqrt(1/alpha.est.CL)
}
return(results)
}
```

A.3. Implementation Details for Joint Model of Cycle Length and Fecundity

The posterior distributions of the population and couple-specific parameters of the joint model described in Section 3.3 are estimated using MCMC sampling methods implemented using the OpenBUGS (Thomas et al. 2006) software using the code in Section A.4. Convergence of parameters is checked by visual inspection of trace plots. Based on these trace plots, we use a burn-in of 10000 iterations and a sample of 30000 iterations. While the regression coefficients of both submodels exhibit rapid mixing, the spline parameters are slower to converge. Convergence can be improved by using fewer knots.

To assess the sensitivity to the choice of uniform priors for $(\sigma_1, \sigma_2, \sigma_W)$, we fit the model assuming uniform priors on the scale of the log standard deviations; and we found that the corresponding posterior distributions were approximately the same. As a separate sensitivity analysis, we also fit the model assuming a log-normal distribution for $W_i, i = 1, \ldots, n$. Again, there was very little difference in the posterior distributions.

For the shape of the profile of pregnancy probabilities in the fecundity model, we assume $g(\cdot)$ is a smooth function and estimate g by $\hat{g}(\cdot) = \sum_{l=1}^{L} \alpha_l B_l(\cdot)$, where $\{B_1(\cdot), \ldots, B_L(\cdot)\}$ are the B-spline basis functions for a natural cubic spline. Brown et al. (2005) suggest choosing the number of knots based on the model with the smallest deviance information criterion (DIC). We separately fit models with 4-10 knots at locations based on percentiles of the intercourse day data. We found that the DIC was smaller for the models with 8, 9, or 10 knots; therefore, we chose to use 8 knots so that the profile of pregnancy probabilities is informed by the intercourse data without the addition of too many parameters. We exclude intercourse acts occurring more than 16 days prior to ovulation or 18 days post ovulation, or on the days following a positive pregnancy test as these occur after conception; and we also exclude intercourse acts on days occurring more than 45 days after the start of a cycle.

A.4. Model Code for Joint Model of Menstrual Cycle Length and Fecundity

The following OpenBUGS computer code can be used to estimate the parameters of the joint model of menstrual cycle length and fecundity developed and implemented in Chapter 3. This code can also be used to do the analyses in Chapter 4 by putting in the data on PFAS exposure as specified below.

```
#Purpose: Estimate posterior distributions of parameters of joint model of
  menstrual cycle length and fecundity


#Start model
model{
const <-10000 #for ones trick


#Regression for menstrual cycle length model
for(i in 1:n){
w[i] ~ dgamma(inv.sigma.squared.w,inv.sigma.squared.w)


    #Put PFAS exposure here
vTeta[i]<- eta.age1*agef[i]+eta.smokef*smokef[i]+ eta.PFAS*PFAS[i]


risk[i] <- exp(vTeta[i])
mult[i]<- w[i]*risk[i]
EYgivWz[i] <- mult[i]*EY0
SDYgivWz[i] <- mult[i]*SDY0


y0.pred[i]<-y.pred[i]*pow(mult[i],-1)
fy.pred[i] <- pow(mult[i],-1)*(p.normal*sqrt(inv.sigma.squared.normal/2/3.14159)*
```

```
  exp(-1/2*inv.sigma.squared.normal*pow(y0.pred[i]-mean.normal,2))+
  p.Gumbel/scale*exp(-(y0.pred[i]-loc)/scale-exp(-(y0.pred[i]-loc)/scale)))
One.pred[i] <- 1
One.pred[i] ~ dbern(p.Onestrick.pred[i])
p.Onestrick.pred[i]  <- fy.pred[i]/const
y.pred[i] ~ dflat()

y.pred.cen[i]<- y.pred[i] - 29.47438 #updated
y.pred.std[i]<- y.pred.cen[i]/3.261141  #updated

    #Regression for pregnancy model
#cycle level data
for (j in offsets[i]:(offsets[i+1]-1)){
#day level data
for(k in dayoffsets[j]:(dayoffsets[j+1]-1)){
#splines
for(m in 1:etamalpha){
gsmoothmat[k,m]<-   alpha[m]*basis[(intcdayrel[k]+addconst+1),m]
}
gsmooth[k]<-sum(gsmoothmat[k,])
innersum[k] <--1*log(1 + exp(beta.MCL[1]*y.pred.std[i] +
   beta.MCL[2]*pow(y.pred.std[i],2)  +
   gamma.STCRIT*STCRIT[i]+ gamma.logSPCOUNT*logSPCOUNT[i]+
   gamma.logSEVOLUME*logSEVOLUME[i] +
   gamma.averageofagefandagem*averageofagefandagem[i] +
   gamma.agemminusagef*agemminusagef[i] +
   gamma.smokef*smokef[i] + gamma.smokem*smokem[i] +
```

```
   (1-nopeakInd[k])*(mu0 + gsmooth[k]) +
   nopeakInd[k]*(lambda[1] + lambda[2]*AvgIntc[i] +
   lambda[3]*pow(AvgIntc[i],2))))
}
log.comp.pcycle[j] <- sum(innersum[dayoffsets[j]:(dayoffsets[j+1]-1)])
log.comp.pcycle.withannov[j] <- (1-annovInd[j])*log.comp.pcycle[j]
pcycle.withannov[j]<- 1-exp(log.comp.pcycle.withannov[j])
}
pttp[i,1]<-pcycle.withannov[offsets[i]]
for(cyc in 2:etamcyc[i]){
pttp[i,cyc]<-pcycle.withannov[offsets[i]+cyc-1]*
exp(sum(log.comp.pcycle.withannov[offsets[i]:(offsets[i]+cyc-2)]))
}
#This will pad the ragged arrays with 1's (change to NA in R) max(etamcyc)=16
for(futcyc in (etamcyc[i]+1):17){
pttp[i,futcyc] <- 1
}
#For last cycle, the probability of Aini = 1  (have to put annov on outside)
log.p.cycle[i] <- (1-annovInd[(offsets[i+1]-1)])*
log(1-exp(log.comp.pcycle[(offsets[i+1]-1)]))
loglik.pregcycle[i]  <-  delta[i]*log.p.cycle[i]
loglik.bycouple[i] <- loglik.pregcycle[i]+indttpgr1[i]*
sum(log.comp.pcycle.withannov[offsets[i]:(offsets[i]+
(etamcyc[i]-1)-delta[i]*indttpgr1[i])])

#Ones trick
z[i] <-1
```

```
z[i]  ~  dbern(pz[i])
pz[i]<- L[i]/const
L[i] <- exp(loglik.bycouple[i])
}

for(j in 1:Ncyc.AllcycCensPregAtOvul){
y0[j] <- y.AllcycCensPregAtOvul[j]*pow(mult[woman.AllcycCensPregAtOvul[j]],-1)
fy[j] <- pow(mult[woman.AllcycCensPregAtOvul[j]],-1)*
(p.normal*sqrt(inv.sigma.squared.normal/2/3.14159)*
exp(-1/2*inv.sigma.squared.normal*pow(y0[j]-mean.normal,2))
p.Gumbel/scale*exp(-(y0[j]-loc)/scale-exp(-(y0[j]-loc)/scale)))
Sy[j] <- p.normal*phi((mean.normal-y0[j])/sigma.normal) +
 p.Gumbel*(1-exp(-exp(-(y0[j]-loc)/scale)))
One[j] <- 1
One[j] ~ dbern(p.Onestrick[j])
p.Onestrick[j]  <- pow(EYgivWz[woman.AllcycCensPregAtOvul[j]],-1*enrollcycInd[j])*
   pow(fy[j],1-pregind[j])*pow(Sy[j],pregind[j])/const
}

EY0<-p.normal*mean.normal+p.Gumbel*mean.Gumbel

#Specification of priors
p.Gumbel <-1-p.normal
p.normal ~ dunif(0,1)  #Alternative: ddirch
mean.normal  ~  dunif(0,50)
sigma.normal ~ dunif(0,10)
inv.sigma.squared.normal <-1/pow(sigma.normal,2)
```

```
lam ~ dunif(0,20)
mean.Gumbel <- mean.normal + lam
sigma.Gumbel ~ dunif(0,40)
scale  <-sigma.Gumbel/3.14159*sqrt(6)
loc <-mean.Gumbel - scale*0.5772156649
sigma.w ~ dunif(0,0.5) #Play with this
inv.sigma.squared.w <-1/pow(sigma.w,2)
eta.age1  ~   dunif(-1,1)
eta.smokef ~   dunif(-1,1)
eta.PFAS ~   dunif(-1,1)

beta.MCL[1] ~ dunif(-2,2)
beta.MCL[2] ~ dunif(-2,2)

gamma.agemminussagef ~ dunif(-2,2)
gamma.averageofagefandagem ~ dunif(-2,2)
gamma.smokef ~ dunif(-2,2)
gamma.smokem ~ dunif(-2,2)

gamma.STCRIT ~ dunif(-2,2)
gamma.logSPCOUNT~ dunif(-2,2)
gamma.logSEVOLUME~ dunif(-2,2)
for(m in 1:etamalpha){
alpha[m] ~ dunif(-1000,1000)
}
mu0 ~ dunif(-1000,100)
lambda[1]~dunif(-30,30)
```

```
lambda[2]~dunif(-10,10)

lambda[3]~dunif(-2,2)

} #End model
```

Kirsten J. Lum

Curriculum Vitæ

(Updated October 26, 2014)

Johns Hopkins University
Department of Biostatistics
615 N. Wolfe St.
Baltimore, MD 21205

Email: klum@jhu.edu

Education

Ph.D. candidate, Biostatistics, Johns Hopkins University (JHU), 2009–present

Thesis topic: *Joint modeling of hierarchical data with application to prospective pregnancy studies*

Advisor: Dr. Thomas A. Louis

M.S., Statistics, American University, 2008

Thesis topic: *Joint modeling of longitudinal binary data and time to an event: an application to human fecundity models*

Advisor: Dr. Elizabeth Malloy

M.S.E, Biomedical Engineering, Johns Hopkins University, 2004

Thesis topic: *Automatic methods for classifying sulcal regions on the human brain cortex*

Advisor: Dr. Jerry Prince

B.S. Electrical Engineering, *cum laude*, University of Maryland in College Park, 2001

Employment

Pre-doctoral Fellow and Summer Intern, Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), National Institutes of Health (NIH), Summers 2007, 2010, and years 2012-present

Teaching Assistant, Department of Biostatistics, JHU, 2010–present

Magnetic Resonance Engineer, Center for the Study of Brain, Mind, and Behavior, Princeton University, 2003–2006

Research Fellow, Lab of Personality and Cognition, National Institute on Aging, NIH, 2003–2006

Honors and Awards

The Helen Abbey Award for excellence in teaching, JHU, 2013

Pre-doctoral Intramural Research Training Award, NICHD, NIH, 2012

Louis I. and Thomas D. Dublin Award for the advancement of Epidemiology and Biostatistics, JHU, 2012

Pre-doctoral Training Award in Environmental Biostatistics, National Institute of Environmental Health Sciences, NIH, 2010-2012

7th Annual Summer Institute in Reproductive and Perinatal Epidemiology, NICHD and CIHRs Institute of Human Development, Child and Youth Health, 2011

Wray Jackson Smith Scholarship, American Statistical Association, 2008

Outstanding Graduate Student Award, Washington Statistical Society, 2007

Professional Memberships

American Statistical Association

Eastern North American Region/International Biometric Society

Caucus for Women in Statistics

Washington Statistical Society

Publications

Published

Lum KJ, Sundaram R., Louis TA (2014). Accounting for length-bias and selection effects in estimating the distribution of menstrual cycle length. Biostatistics. Advanced access published July 14, 2014, doi: 10.1093/biostatistics/kxu035.

McLain A, **Lum KJ**, Sundaram R (2012). A joint mixed effects dispersion model for menstrual cycle length and time-to-pregnancy. Biometrics, 68: 648-656. PMID: 22321128.

Lynch, Courtney D., Sundaram R, Buck Louis GM, **Lum KJ**, Pyper C (2012). Are increased levels of self-reported psychosocial stress, anxiety, and depression associated with fecundity?. Fertility and Sterility, 98: 453-458. PMID: 22698634.

Lum KJ, Sundaram R, Buck Louis GM (2011). Women's lifestyle behaviors while trying to become pregnant: evidence supporting preconception guidance. American Journal of Obstetrics and Gynecology, 205: 203.e1-7. PMID: 21658667.

Buck Louis GM, **Lum KJ**, Sundaram R, Chen Z, Kim SD, Lynch CJ, Schisterman EF, Pyper C (2011). Stress reduces conception probabilities across the fertile window: evidence in support of relaxation. Fertility and Sterility, 95: 2184-2189. PMID: 20688324.

Neta GI, von Ehrenstein OS, Goldman LR, **Lum KJ**, Sundaram R, Andrews W, Zhang J (2010). Umbilical cord serum cytokine levels and risks of small-for-gestational-age and preterm birth. American Journal of Epidemiology, 171: 859-867. PMID: 20348155.

Pollack AZ, Buck Louis GM, Sundaram R, **Lum KJ** (2010). Caffeine consumption and miscarriage: a prospective cohort study. Fertility and Sterility, 93: 304-306. PMID: 19732873.

Behnke KJ, Rettmann ME, Pham DL, Shen D, Resnick SM, Davatzikos C, Prince JL (2003). Automatic classification of sulcal regions of the human brain cortex using pattern recognition. Medical Imaging, Proc of SPIE, 5032: 1499-1510.

In Progress

Lum KJ, Sundaram R, Buck Louis GM, Louis TA. A Bayesian approach to joint modeling of menstrual cycle length and fecundity, (Aug 2014). Johns Hopkins University, Dept. of Biostatistics Working Papers. `http://biostats.bepress.com/jhubiostat/`

Lum KJ, Sundaram R, Louis TA, Buck Louis GM. Perfluoroalkyl substances and their relations with menstrual cycle length and fecundity: the LIFE Study.

Teaching

Teaching Assistant

Non-Inferiority and Equivalence Clinical Trials, JHU, 2014

Biostatistics in Medical Product Regulation, JHU, 2013-2014

Lead TA, Statistical Methods in Public Health I-II, JHU, 2011, 2012

Statistical Methods in Public Health I-IV, JHU, 2010

Basic Algebra, American University, 2006-2007

Guest Lectures

Statistical Methods in Public Health II, JHU, 2012

Presentations and Posters

A Bayesian Approach to Joint Modeling of Longitudinal Menstrual Cycle Length and the Probability of Pregnancy. Joint Statistical Meetings (JSM), Aug 2014, Boston, MA [speed presentation, poster combination]

Assessing the Association Between Time-to-Pregnancy and Woman-Specific Menstrual Cycle Length Mean and Variability Using a Bayesian Hierarchical Model. Joint Statistical Meetings (JSM), August 2013, Montreal, Canada. [presentation]

Accounting for Length-bias and Selection Bias in Estimating Menstrual Cycle Length. Eastern North American Region (ENAR) Spring Meeting, March 2013, Orlando, FL. [presentation]

Incorporating sampling plan and competing risks in analysis of prospective pregnancy studies. ENAR, April 2012, Washington, DC. [poster]

Joint modeling of menstrual cycle length and fecundity. ENAR, March 2011, Miami, FL. [poster]

Womens lifestyle behaviors during the periconception window: evidence supporting need for preconception guidance. 1st European Congress 2010: Preconception Care and Preconception Health, October 2010, Brussels, Belgium. [presentation]

Joint modeling of binary longitudinal data with discrete time-to-event: an application to human fecundity models. JSM, August 2009, Washington, DC. [presentation]

Do women change behavior while attempting pregnancy? Society for Epidemiologic Research 42nd Annual Meeting, June 2009, Anaheim, CA. [poster]

Joint model of longitudinal binary data with discrete survival time. JSM, August 2008, Denver, CO. [presentation]

A model for human intercourse behavior and its importance in modeling time to pregnancy. American University CAS Robyn Rafferty Mathias Student Research Conference, March 2008, Washington, DC. [presentation]

Incorporating libido into human fecundability models. ENAR, March 2008, Arlington, VA. [presentation]

Computing Skills

R, S-PLUS, Matlab, OpenBUGS, SAS (certified SAS 9), PROC IML, STATA, C, C++, IDEA (certified, Siemens Corporation)

Leadership and Service Activities

Organizer and chair of invited session at ENAR, 2014

Organizer of JHU Biostatistics Student Welcome Picnic, 2011

Leader of JHU Biostatistics Computing Club, 2010-2011

Organizer of JHU Biostatistics Student Camping Trip, 2010