2018

Effective exposure: lag-parameterized exponential models for exposure risk

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

EFFECTIVE EXPOSURE: LAG-PARAMETERIZED EXPONENTIAL MODELS FOR EXPOSURE RISK

by

HANNA GERLOVIN

BA in Mathematical Sciences, Colby College, 2008
MA in Biostatistics, Boston University, 2012

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2018

Approved by

First Reader

David R. Gagnon, MD MPH PhD
Research Professor, Biostatistics

Second Reader

Michael LaValley, PhD
Professor, Biostatistics

Third Reader

Ludovic Trinquart, PhD
Assistant Professor, Biostatistics

Fourth Reader

Elena Losina, PhD
Adjunct Professor, Biostatistics

DEDICATION

This dissertation is dedicated to my late parents, Dina and Emmanuel Gerlovin. May this work be a testament to their perseverance and love - for if you can leave your entire life behind at the age of 40, to come to a new country and world, anything else is possible. Thank you for your unconditional support. I miss you.

ACKNOWLEDGMENTS

First, and foremost, THANK YOU to my advisor, Dr. David Gagnon, for his help with the concepts, ideas, and never-ending patience. It has been a pleasure to learn from you, and I am especially grateful for your consistent and unwavering support of both my methodologic advancements and professional development.

Additionally, I would like to thank my committee members, Drs. Timothy Heeren, Michael LaValley, Elena Losina, and Ludovic Trinquart, for the enthusiasm and "big picture" perspectives that helped me shape this dissertation beyond the programmatic functions. I look forward to working with you all for years to come, specifically, as I prepare manuscripts for submission. There are too many Biostatistics department faculty to thank individually, but your faith in me has been paramount.

I would also like to acknowledge the many supporters from my research tenure at the Slone Epidemiology Center, in particular, on the Black Women's Health Study. Thank you for trusting me with your data and for giving me a place to call "home" over the last six years. Thank you to Drs. Julie Palmer, Lynn Rosenberg, Kim Bertrand, and Edward Ruiz-Narvaez for the mentorship and career support throughout the doctoral program.

Thank you to my friends and family for the emotional support, especially my brothers, Lev and Mark, and parents-in-law, Jan and Craig. You provided me with guidance and confidence when I needed it most.

Finally, thank you to my rock, my heart, my husband, Josh. You have gotten me through the hardest of times, and always challenged me to persevere. Your encouragement, love, and partnership, have given me the strength to reach higher, and I will forever be grateful.

EFFECTIVE EXPOSURE: LAG-PARAMETERIZED EXPONENTIAL MODELS FOR EXPOSURE RISK

HANNA GERLOVIN

Boston University, Graduate School of Arts and Sciences, 2018

Major Professor: David R. Gagnon, PhD

ABSTRACT

Many observational studies assessing the effects of treatments or exposures are limited to comparisons between treatment users and nonusers or exposed and unexposed participants at study entry. However, the underlying and etiologically relevant exposure may gradually increase over time before reaching some plateau. This amount of time required for this latent cumulative exposure to reach a maximum hazard will be referred to as the "lag", coming from the concept that the association between exposure and outcome is lagged or delayed. Accounting for the lag is essential when analyzing exposure-response associations adequately. My challenge was to simultaneously estimate the lag-time and the exposure's lagged-association with the outcome at plateau.

In this dissertation, I draw an analogy with the pharmacokinetic one-compartment model (OCM). OCM describes the accumulation of a medication in the body based on an exponential cumulative density function whose rate of increase is defined by a half-life parameter. Upon discontinuation, the OCM assumes that a medication will eliminate at the same half-life rate. The decline, for my purposes, can be interpreted as the time to return to a null effect of exposure, which occurs at roughly 4-5 half-lives.

My methods model the association of a latent exposure and dichotomous outcome using a half-life of effect, similar to the OCM, in longitudinal analyses of single and repeated exposures. I derive profile likelihood-based algorithms to estimate of the upper limit of association simultaneously with the rate of latent exposure growth towards or away from plateau. Lastly, I extend this approach to allow different half-life parameters for incline and decline.

Using simulations, I analyze the performance of my approach by comparing bias and coverage of the estimates for the half-life and effect parameters. With data from the Black Women's Health Study Cohort (a prospective cohort of 59,000 women followed 1995-2015), I show that prolonged cigarette smoking is associated with a maximum hazard of cardiovascular disease (CVD) at 2.5 times the hazard of never smokers. Additionally, I estimate that it takes about 7 years of smoking cessation for an individual's hazard of CVD to decrease by 50%.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AIC ...	Akaike's Information Criterion
b	Starting Time
BWHS .	Black Women's Health Study
CI	Confidence Interval
CP ...	Coverage Probability
CPH ..	Cox Proportional Hazards Regression
CVD ..	Cardiovascular Disease
D	Current Dose
EE, E_{it}, E_j	Effective Exposure
f	Stopping Time
HR ...	Hazards Ratio
λ, h ...	Lag/Half-Life Parameter in OPEE
λ_1, h_1 ..	Incline Lag/Half-Life Parameter in TPEE
λ_2, h_2 ..	Decline Lag/Half-Life Parameter in TPEE
OPEE ..	One-parameter effective exposure model
PLL ...	Profile Log–Likelihood
PLR ...	Pooled Logistic Regression
t	Time
TPEE ..	Two-parameter effective exposure model
z_1, z_2 ..	Time since start, stop in Single-Dose OPEE and TPEE
z_3, z_4 ..	Time since start, stop for Second Dosing in multiple exposures context

CHAPTER 1

Introduction

Public health research has long been aware of the detrimental effects of smoking on health, particularly the increased risk of cardiovascular diseases (CVD).(U.S. Department of Health and Human Services, 1990; WHO, 2004) Cigarettes and tobacco cause plaque build-up in the arteries, leading to an increased risk of CVD for individuals who smoke. Few would consider a heart attack occurring after one week of smoking to be caused by use *alone*. Conversely amongst those who smoke for an extended period of time, one would not expect the CVD hazard to simply disappear following cessation. Thus, the question is how best to account for transitioning individuals when analyzing risk in a population-level model?

Clinicians and health professionals agree on the benefits of smoking cessation in terms of reducing risks, though literature has been mixed regarding the amount of time required to return to "normal".(Kawachi et al., 1994; Rachet et al., 2003; Rosenberg et al., 1990) Recommendations for smoking cessation are made, with the intention of lowering the smoking-associated health risks. However, estimation of the hazard is complicated by the fact that it takes time for the impact of a history of smoking to go away completely. For the "on-again off-again" life-course of many smokers, it is additionally challenging to minimize misclassification of exposure, which can bias estimation of risk. That is, how does one appropriately classify the exposure for someone who is not consistent in their cessation or habits of smoking?

The time-to-effect of an exposure on an outcome can be thought of in terms of "lag" - i.e. a period of time that must elapse, following exposure, prior to seeing a measurable change in risk. Lagged, or delayed, effects have been studied in a breadth of examples, including the multiple conditions for which smoking cessa-

tion is considered to reduce risk over time.(U.S. Department of Health and Human Services, 1990)

In particular, Rachet et al. (2003) explored the distribution of the lag in the association between smoking cessation and heart attack, using data from the Framingham Heart Study. The method here was limited to individuals with successful smoking cessation, yet, given that the average smoker may try to quit 30 times before success(Chaiton et al., 2016), there is a clear need for models that can account for and handle more complex scenarios.

Another approach to deal with lagged exposure-response associations has been to look at the cumulative dose, such as the total years smoked or pack-years, at the time of event. The downside to this technique comes from its inability to account for discontinued use, i.e. the latent exposure may subside, while a cumulative dose is assumed to stay constant.

For this dissertation, the delay, or "lag", will be defined as the amount of time between exposure initiation or discontinuation and the time to saturation or elimination of the underlying hazard. This differs from the epidemiologic concept of a "latency period", in that an event may occur during the lag-time of an effect, but is not expected to occur within the former period of time specification. The overall goal of this thesis will be to introduce novel statistical methods for estimating a lagged effect, and the lag-time associated with that effect size.

The analogous structure of the novel models I introduce comes from the pharmacokinetic one-compartment model (OCM) in that the effective amount of a single or set of protracted exposures, over time, follows an exponential accumulation or decay curve. The first-order elimination rate parameter in OCM has an intuitive "half-life" interpretation, which can be used to describe the amount of time

required for the risk to rise or fall half-way. In my implementation, the volume and clearance parameters are factored out of the equations, leaving only the relative concentration, rate of growth/decay, and time parameters. Together, these parameters model the shape of a latent risk curve, which looks similar to the OCM, but reflects the increase or decrease in the effective level of exposure relative to a maximum hazard.

Unlike the pharmacological effect in the OCM, the biologic effective exposure value [that parallels the "concentration" in the OCM equation] may not be an easily measurable quantity or readily available. To illustrate this, consider the effect of prednisone or corticosteroids (CS) on the risk of fracture. The biologic mechanism could be that CS leach calcium from the bones, which leads to an increased risk of fracture.(Van Staa et al., 2000; Vestergaard et al., 2008) However, measuring bone density as a marker can often be costly and inaccessible for study. Therefore, the unobservable change in the risk of fracture due to CS use is what I am interested in modeling. That is – how can I model the population-level hazard when the effect of the exposure, which is assumed to have some lag, is also changing over time?

Another use of my method could be to evaluate the presence of spurious or unexpected associations, specifically those that return an estimated lag-time that is infeasible biologically. For example, an estimated time-to-null hazard of one day [for CVD outcomes following smoking cessation] is highly improbable, and could imply some confounding by indication in the analytic approach. Alternately, should the estimated lag be infinite, one may need to reassess the biologic model that assumes the CVD hazard associated with smoking could even return to the level of never smokers. Such concepts and limitations are discussed in more detail in Chapters 3 and 5.

In the remainder of Chapter 1, I introduce examples of lag effects and provide some background on previous approaches used to account for lag-time in analysis. To serve as inspiration for the novel methods developed here, I also provide brief introductions to pharmacokinetic and longitudinal data models, and give an overview on profile likelihood estimation.

1.1 LAG OF EFFECT

In the context of time-variant primary exposures, mixed effects and survival analysis models have been used to estimate the association between an exposure and some time-to-event outcome of interest. However, when the underlying exposure quantity is unknown, or the exposure may not have an immediate action mechanism, researchers make assumptions to attempt to account for the delay. Therefore, most clinical and analytic approaches only try to account for the delay in the estimated effect but fail to estimate it.

Analysis of administrative and longitudinal data typically requires assumptions be made regarding the causal and temporal relationships between exposures and outcomes. As long as a risk factor occurs with enough time prior to the event of interest, the mechanism of effect can be estimated with certainty and minimal bias.(Rothman, 1981) Difficulties may arise when one must allow an amount of time for an exposure to fully turn "on".

The field of epidemiologic study is ripe with examples of delayed and lagged effects, both in the realm of pharmaceutical interventions, as well as, with non-therapeutic exposures such as environmental pollutants.(Langholz et al., 1999; Thomas, 1983) To narrow the focus of this methodology, I am specifically interested in dichotomous events and outcomes whose risk due to protracted exposure is as-

sumed to plateau, or stabilize, given "enough" time exposed. That is, those who have been exposed for an extended period of time are considered to be at a maximum hazard associated with the underlying exposure mechanism.

Additional complications arise when exposures are not consistent over time, making classification decisions difficult during population-model building, and can often lead to the exclusion of subjects in transition-states. These individuals present a rich source of untapped information about the exposure. Restricted analyses may not paint the full picture of the time-to-effect, and the results may only be generalized to the populations represented by the restricted sample. For example, the benefits of weight loss in overweight/obese individuals with regards to reducing the risk of CVD and/or type 2 diabetes may not be applicable to weight-cyclers, or "yo-yo dieters", as studies have found a majority of men and women are unable to maintain their reduction in body weight.(Strohacker et al., 2009)

In an analysis of Nurses Health Study data, Giovannucci et al. (1995) found a protective effect of regular aspirin use on the risk of colorectal cancer. While the association was only statistically significant in a restricted sample of consecutive reports [of aspirin use], the unrestricted analysis still indicated that consumption of two or more aspirin tablets per week would lower the risk of colorectal cancer. Similar results were seen in the Health Professionals Follow-Up Study, whereby the Giovannucci et al. (1994) concluded that any extended period of aspirin use was associated with a reduced risk of colorectal cancer and adenomas.

Other examples include the declining risk of cardiovascular events following smoking cessation, the change in the likelihood of fracture due to corticosteroid intake or discontinuation, and clinical improvements in depression symptoms after the initiation of treatment. In all of these examples, the true impact of the exposure

can only be properly estimated, if the amount of time-to-effect, or lag, is correctly specified.

1.2 PREVIOUS WORK THAT CONSIDERS LAGGED EFFECTS

Rothman's Induction and Latent Periods

Rothman (1981) explored the differences in terminology for time between various states of the disease process, specifically causal mechanisms leading to disease onset and then disease detection. The take-away from the article follows the notion that the events must happen in a particular order. Failing to take into account the period lengths may result in non-differential misclassification and underestimation of the effect of interest. The proposed solution is to look at models and analyses under different assumptions of the empirical induction period, and to select the corresponding lag that results in an effect estimate furthest from the null.

Rothman's theory has been disputed(Salvan et al., 1995; Richardson et al., 2011), and more recent work has focused on maximizing the likelihood/partial likelihood function, or selecting a lag-adjusted model based on Akaike's Information Criterion(Akaike, 1974). The latter method allows for comparisons across non-nested models by penalizing the likelihood for the number of parameters estimated. This can be useful when trying to account for nuisance parameters or estimates of correlation structure that may have problems with model overfitting.

Time windows of susceptibility

One methodology that relates back to Rothman's work is the use of time windows to mark individuals as "exposed" within particular intervals following exposure. These have been referred to as the "time windows of susceptibility" or "sliding win-

dows". Applications of this method have been primarily focused on occupational and environmental exposures, and the windows are selected by comparing the deviance statistics for models under different fixed intervals of exposure.(Finkelstein, 1991; Hauptmann et al., 2000a)

A particular drawback of this approach has been the need to select the windows a priori, and the method does not account for protracted exposures or non-linear latency functions over time. While Hauptmann et al. extended the approach to look at both window width and position, the application was restricted to a case-control study design which resulted in bias due to the retrospective assessment of the exposure history.(Hauptmann et al., 2000a)

Splines and weighted cumulative exposures

One of the more common and published approaches for dealing with lagged effects relies on weighting past exposure events or integrating over the entire exposure history in order calculate the "etiologically relevant" exposure metric. (Langholz et al., 1999; Abrahamowicz et al., 1992, 1996; Hauptmann et al., 2000b; Rachet et al., 2003; Sylvestre & Abrahamowicz, 2009) The particular usefulness of this method comes from its ability to account for varying exposure intensities and durations, as well as, estimating a clinically meaningful measure for the exposure-response relationship.

The cubic splines model for cumulative exposure weighting of lagged effects was described by Abrahamowicz et al. (1996) in an application to lupus nephritis. Here, the exposure history is broken into segments based on a pre-selected number of knots and a differentiable order of polynomial spline functions. One strength of this approach is the ability to test for the type of exposure-hazard relationship,

since the lower order functions of the predictor are nested within the higher order model. Similar to Rothman, the authors utilized AIC to assess the model's fit compared to conventional time-varying dose and duration models. Meanwhile, the confidence bounds for the predictors were derived by maximizing the partial likelihood. This is the method that Rachet et al. (2003) employed for estimating the distribution of the lag-time associated with reduction in the hazard of heart attack following smoking cessation.

Distributed lag linear and non-linear models

The distributed lag modeling framework is inspired by time-series regression in economics and has been described in the context of generalized additive models(Zanobetti et al., 2000) and non-linear models (Gasparrini et al., 2010). The frameworks allow for various lag-parameterized basis functions to define the exposure's effective amount over time. The idea of these models is to calculate multiple parameters for each lag-period exposure, with a final coefficient summarizing the overall effect of a unit change in the weighted average of exposure.

1.3 THE PHARMACOKINETIC ONE-COMPARTMENT MODEL

The one-compartment model (**OCM**) is a pharmacokinetic formulation that quantifies the amount or concentration of a desired drug within the plasma, over time. Given a known rate of elimination [of the drug from the system], one can compute the effective amount of the agent in the compartment of choice.(Winter, 2004) For an Intravenous (IV)-administered drug, it is possible to determine the amount of time needed to reach a steady state level, as well as, the relative amount of steady state concentration at any given time. This "concentration" model can be used as a

general framework for describing the behaviors of a latent exposure, dictated by a rate of growth or decay towards or away from a steady state plateau, over time.

The formulas required to calculate the concentration at a given time are broken down into three possible time-frames: 1) At the beginning of infusion, 2) During steady state, and 3) After discontinuing infusion. The single generalized formula that can be used to calculate concentration at time, t, follows(Wijnand, 1988):

$$C_p^t = \begin{cases} \frac{k_0}{k_e V}\left[1 - e^{-k_e t}\right] & \text{if } t \leq D \\ \frac{k_0}{k_e V}\left[1 - e^{-k_e D}\right] e^{-k_e (t-D)} & \text{if } t > D \end{cases} \tag{1.1}$$

Where the infusion begins at time 0, D denotes the end-time, $t - D$ the time elapsed since ending the infusion [for the second condition], k_0 and k_e are the infusion and elimination rates, and V represents the volume of the administered infusion. The first condition is equivalent to the second by replacing D with t, when t is less than or equal to D. By the definition of steady state, where the concentration accumulates and eliminates at the same rate, the total concentration is equal to:

$$C_p^{ss} = \frac{k_0}{k_e V} \tag{1.2}$$

Thus, the relative concentration at time t vs. the steady state (ss) level can be described by:

$$\frac{C_p^t}{C_p^{ss}} = \frac{k_0}{k_e V}\left[1 - e^{-k_e D}\right] * e^{-k_e (t-D)} \times \frac{k_e V}{k_0} = \left[1 - e^{-k_e D}\right] * e^{-k_e (t-D)} \tag{1.3}$$

That is, at any given time point, the achieved proportion of the steady state value can be modeled via an exponential curve with a known constant elimination rate. For as long as the infusion continues, this function is monotonically increas-

ing, and then monotonically decreasing once the infusion is stopped. Another strength of this formulation is the ease of calculating the elimination rate, given a known amount of time required to get to half of the steady state concentration. Specifically, one can prove that the elimination rate is just a function of the half-life, and conversely, that if the elimination rate is known, one can estimate the half-life of the drug in the compartment of choice.

$$\frac{1}{2} = 1 - e^{-k_e t_h} \longrightarrow \frac{1}{2} = e^{-k_e t_h} \longrightarrow \ln(\frac{1}{2}) = \ln(1) - \ln(2) = -\ln(2) = -k_e t_h \Longrightarrow$$

$$k_e = \frac{\ln 2}{t_h} \Longleftrightarrow t_h = \frac{\ln 2}{k_e}$$

The OCM structure assumes that the curve will increase over time for approximately 4-5 half-lives, before entering steady state, at which point, continuation of the medication does not appreciably change the level of concentration in the system. Once the medication is stopped, it should take approximately 5 half-lives to return back to zero or "normal", and starting another infusion or dose should result in a concentration that is equivalent to the sum of the two curves.

These concepts can be stretched to a more abstract formulation, noting that equation (1.3) defines the relative concentration in the plasma, over time, versus the maximum concentration achievable at steady state. Let me define this concentration ratio from as C_{ratio}.

In equation (1.3) the assumption stands that time begins to increment at the start of infusion, $t = 0$. For a population or sample where the start times are not all indexed by time=0, the formula is revised to three segments separated by the start and stop times of the infusion. Prior to initialization, the relative concentration is zero. Once the infusion is started, this ratio will begin to rise towards 1, and after

discontinuation, the ratio is expected to decline back to zero. Both the increase and decrease depend upon the rate of elimination. Functionally, this can be written as:

$$C_{ratio} = \begin{cases} 0 & \text{if } t \leq b \\ 1 - e^{-k_e(t-b)} & \text{if } b < t \leq f \\ \left[1 - e^{-k_e(f-b)}\right] * e^{-k_e(t-f)} & \text{if } t > f \end{cases} \tag{1.4}$$

Note, that this set of equations (1.4) generalizes the drug exposure start time as b, rather than 0. This shift is important, as well as the specifications that $b < f$, for $b \epsilon T$ and $f \epsilon T$, where T is the range of surveillance times, for which data is available and/or collected. The quantity, $f - b$, is equivalent to the total amount of time exposed, or D from equation (1.3).

When multiple IV infusions are given, the rate of elimination does not change, thus the total concentration, or effective dose, in the compartment becomes a simple sum of the individual concentrations.(Bourne, 2010) I will take advantage of this mechanism in chapter 2.

1.4 LONGITUDINAL MODELS

The question of interest for this dissertation lies in the modeling of a lagged hazard, or a risk over time conditional on surviving up to that time. There are several regression approaches that work in this context, including, but not limited to, Cox proportional hazards (CPH) and pooled logistic regression (PLR) models. Both models have been shown to work for analyses of risk over time, with benefits and costs to each approach.(Cupples et al., 1988; D'Agostino et al., 1990; Ngwa et al., 2016) I offer a brief overview of these methods, below, to prepare for their use throughout the remainder of the dissertation.

1.4.1 Cox Proportional Hazards Regression

Cox proportional hazards (CPH) regression models are one of the most common forms of regression used in analyzing exposure-response associations in the context of time-to-event outcomes.(Cox, 1972; Therneau & Grambsch, 2000; Hosmer et al., 2008; Kleinbaum & Klein, 2011) Here, time is considered to be a part of the outcome and the interest lies in determining differences in survival due to some exposure by looking at the relative hazards over time. The linear model is fit for the hazard at time t, $h(t|\boldsymbol{X}(t))$, given the data $\boldsymbol{X}(t)$ and an unknown baseline hazard function, $h_0(t)$.

$$h(t|\boldsymbol{X}(t)) = h_0(t) \exp\left[x_{1j}\beta_1 + ... + x_{qj}\beta_q\right]$$

The semi-parametric nature of the Cox model implies that the baseline hazard does not need to be specified or estimated. The parameters for each of the exposures in the model are assumed to be proportional across time. Using the extension proposed by Andersen & Gill (1982), it is possible to update the values for each subject's time-dependent exposures, such that the conditional form of the equation satisfies the proportional hazards assumption.

A typical maximum likelihood estimation approach first requires specification of the likelihood and log-likelihood functions. For the CPH model, the likelihood becomes a product of the unique event time hazards. To keep the derivations and steps generalize-able, I consider the case where more than one event may occur at a particular time. To handle these ties, I utilize the likelihood and log-likelihood formulations proposed by Breslow (1974) throughout the dissertation. The numerator represents the sum of exponential risk of event for all individuals with an event at that time, while the denominator is the sum of the exponential risk of event across all individuals that have survived up to that time multiplied by the

number of events at that time.

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{Y}) = \prod_{k=1}^{K} \frac{\sum_{j \epsilon R(t_k, Y_j=1)} exp(\boldsymbol{X_j \beta})}{\left[\sum_{j \epsilon R(t_k)} exp(\boldsymbol{X_j \beta}) \right]^{m_k}}$$
$$\ell(\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{Y}) = \sum_{k=1}^{K} \left[\sum_{j \epsilon R(t_k, Y_j=1)} \left(\sum_{q} x_{qj} \beta_q \right) - m_k \ln \left(\sum_{j \epsilon R(t_k)} e^{\sum_q x_{qj} \beta_q} \right) \right] \tag{1.5}$$

where t_k is the kth unique event time index, with m_k equal to the total number of events at time t_k, $R(t_k)$ represents the set of subjects at risk in time t_k, and with the likelihood function's numerator taking the sum of exponential risk scores for all subjects with events at time t_k. Specifically, I define the risk score for subject i, as the sum of the product of the covariates and their respective predicted coefficients:

$$\hat{r_i} = \sum_{q} x_{qi} \hat{\beta}_q = x_{1i} \hat{\beta}_1 + ... + x_{qi} \hat{\beta}_q$$

Due to the unique semi-discrete nature of the CPH estimation process, where risks are summed and component likelihoods calculated by strata-time slices, tied events are likely to occur and need to be handled appropriately. As already mentioned, Breslow's approach will be used throughout the dissertation, which has been shown to be less conservative than the approach proposed by Efron (1977). However, the simplicity of Breslow's formula, that treats each event in a given time as equally-likely, allows for faster computations, a preferable quality for my method.(Hertz-Picciotto & Rockhill, 1997)

1.4.2 Pooled Logistic Regression

An alternative method to the CPH is the pooled logistic regression (PLR) model, which has been shown to work for repeated measures study designs. Specifically, interval-sliced data for subjects with time-varying covariates can be pooled to estimate the conditional odds of an event.(Cupples et al., 1988; D'Agostino et al., 1990) The primary difference, here, is in the interpretation of the effect measures as the conditional odds of event having survived up to that time. As long as the interval considered for the repeated measures is small and the events are relatively rare, the PLR models provide reasonably comparable odds ratio estimates to the CPH hazard ratio of the effect and it's standard error.(Green & Symons, 1983)

My methodology can handle both types of analytic models, though my main focus remains on time-to-event outcomes. In chapter 2, I outline the equations that relate to the pooled logistic analyses, and in chapter 3, I touch upon the differences and, potential, limitations of my method applied to this modeling framework.

For completeness, below, I have shared the PLR likelihood and log-likelihood functions that are maximized during the process of estimating the effect parameters.

$$\begin{aligned} \mathcal{L}(\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{Y}) &= \prod_{i=1}^{n} \prod_{t=1}^{T} p_{it}^{Y_{it}} (1 - p_{it})^{(1-Y_{it})} \\ \ell(\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{Y}) &= \sum_{i=1}^{n} \sum_{t=1}^{T} Y_{it} ln p_{it} + (1 - Y_{it}) ln(1 - p_{it}) \end{aligned} \tag{1.6}$$

$$p_{it} = \frac{exp(\beta_0 + x_{1it}\beta_1 + ... + x_{qit}\beta_q)}{1 + exp(\beta_0 + x_{1it}\beta_1 + ... + x_{qit}\beta_q)} \tag{1.7}$$

Let i denote the subject at time t, and the individual's time-specific values are defined as follows: p_{it} is the probability of event, $\boldsymbol{X}$ is the data for q exposures,

taking values $x_{1it}, ..., x_{qit}$, β_q is the qth covariate's true association coefficient, and Y_{it} takes the value of 1 for an event and 0, otherwise, for subject i at time t.

1.4.3 Profile Likelihood Estimation

The profile likelihood [also referred to as the profile log-likelihood method (PLL)] approach can be thought of as the marginal likelihood of a model across levels of a pre-specified parameter.(Cole et al., 2014; Venzon & Moolgavkar, 1988; Murphy & Van Der Vaart, 2000; Sprott, 2000; Cox & Reid, 1992) In this estimation technique, the parameter of interest is fixed, while the other parameters are estimated via traditional maximum likelihood and regression methods. Graphing the likelihood or log-likelihood for the fully adjusted model against the fixed parameter, provides a visual representation of the likelihood's behavior attributable to the parameter of interest. If the likelihood is unimodal and the log-likelihood looks like an inverted "U", then the resulting curve's maximum should occur at the value of the parameter that would be found using maximum likelihood estimation.

One reason to define this likelihood profile is to determine confidence bounds for the parameter, by looking at which points of the curve cross the horizontal line located at one chi-square's distance below the maximum. Figure 1.1 illustrates how this may look. The horizontal line indicates one chi-square distance from the maxima, and the confidence interval for λ's estimate is defined by the values at which the horizontal line intersects the profile curve.[1]

Since the log-likelihood can be used for purposes of maximization, rather than the full likelihood, I utilize the form of the profile log-likelihood (PLL) for most of the estimation algorithms proposed in this text.

[1]Image taken from: https://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture8/fig4new.png

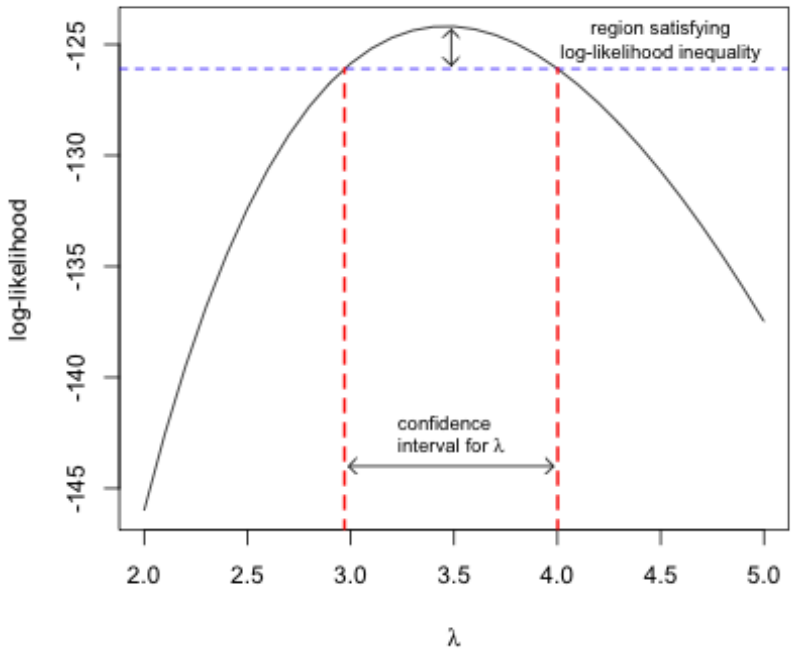


Figure 1.1: Profile Likelihood

1.5 DISSERTATION SECTIONS

In chapter 2, the properties of the OCM are described to serve as a structure for lagged latent exposure profiles over time. The transition from OCM to effective exposures models is fairly straightforward by recognizing the cumulative exponential function can be parameterized using lag in terms of the elimination rate. I further step away from the OCM by extending the exposure's formulation from a single parameter to two parameters, such that the exposure's effect curve has separate rates of incline and decline. With each of these defined exposures, I introduce a profile likelihood-based algorithm for estimating the lag and hazard concurrently, with corresponding estimates of uncertainty. I derive both the single- and two-parameter approaches for CPH and PLR analytic models of the outcome.

Chapter 3 presents the results of Monte Carlo simulation studies in which the

statistical performances of the single- and two-parameter approaches are explored for the concurrent estimation under known inputs for the lag and effect size. The simulation studies are based on a range of scenarios that correspond to real-life clinical examples. In Chapter 4, I examine the lagged relationship between smoking and CVD hazard amongst the Black Women's Health Study (BWHS) cohort. This applied chapter explores smoking as both a binary exposure and in terms of packs per day over time, accounting for multiple time-varying potential confounders. I compare the effective exposure analyses to conventional exposure variables of smoking, such as current vs. past vs. never smoking, or cumulative years smoked.

In Chapter 5, I summarize the results from chapters 2-4, discuss the strengths and limitations of my proposed methods, and describe theoretical and applied research perspectives.

CHAPTER 2

Effective Exposure Derivation

The goal of this chapter is derive and describe a novel methodology for estimation of lagged effects, both in terms of the latent period and overall effect size of a series of exposures. I start by drawing the parallels between the pharmacokinetic one-compartment model (OCM), described in the Introduction [Chapter 1], and the effective exposure distribution, over time, as parameterized by the half-life of the effect. In order to use this formulation, I describe the relevant assumptions needed for estimating the hazard and lag parameters concurrently.

After stating the methodologies behind the single-parameter single-exposure model, I further extend the algorithmic estimation approach to more complex exposures – specifically, the two-parameter effective exposure model, and the protracted exposures models. The latter functions as an extension to either the one- or two-lag parameter variants of effective exposure. For the purpose of this chapter I only focus on the multiple dosing (protracted exposures) using a single lag.

This chapter is meant to be read as a full description of the methods developed for the dissertation. In Chapter 3, I compare multiple models using simulations and outline the relative strengths and weaknesses of the algorithms.

2.1 TRANSITION TO LONGITUDINAL OBSERVATIONAL DATA

As described in the introduction to OCM (section 1.3), given a known elimination rate, k_e, one can calculate the relative exposure level, or concentration over time. Recalling from (1.3), the volume and clearance parameters become irrelevant when modeling the ratio curve as it plateaus to 1 (approaches steady state) over sufficient time. The primary interest becomes the rate at which the curve inclines toward or

declines from steady state, which is driven by the lag parameter, or the half-life.

In this model, it is assumed that after an extended period of time, the hazard, or concentration of risk, due to an exposure should stabilize at some steady state-like effect size. The attributed risk, in this context, does not require an assumption of causality, but does depend on the existence of some underlying action mechanism by which the likelihood of the outcome changes relative to a threshold of the cumulative exposure over time. That means that one could model the underlying, or latent, hazard of an exposure on some outcome, using a time-varying quantity as a proxy for the weighted cumulative exposure. This latent quantity can be considered in terms of a measure that is relative to the maximum level of effect at any given time.

Let me define the term "**Effective Exposure**" (EE), or $E_{it}(\lambda)$, *as the relative amount of an exposure necessary to impose some effect, and for which the effect will eventually reach plateau, or steady state*. This could also be referred to as the point at which the hazard ratio associated with the exposure attains maximum. It will be used to represent the latent (unobserved) time-varying association of an event in relation to the individual's lifetime history of an exposure of interest.

The λ is used as the lag parameter, implying that the underlying effective exposure changes over time based on a decay rate[1]. As the pharmacokinetic elimination rate is constant with respect to volume and dose, the assumption stands that the lag parameter of an effect curve should not vary across individuals, time, or concentration of exposure.

This rate parameter can also be converted to a half-life, h, by the properties of the OCM, which lends more intuitively to interpretation. This also implies that

[1] Analogous to the elimination rate, k_e from the OCM

it may be possible to estimate λ or h, given total exposure profiles over time for individuals in a cohort. For example, one could estimate that the absolute cardiovascular disease (CVD) hazard associated with smoking would decrease 50% after h (half-life) number of years following complete cessation. The change is relative to a starting point [hazard], and thus should be considered in terms of the individual's effective exposure and not in terms of the time required to reduce the hazard ratio by 50%.

The term "dose" will be used to refer to the level at which steady state, or the maximum hazard, occurs. The lag estimation always ties back to a single unit of the dose being used in the model. This means that for each version of the exposure used (binary vs. continuous), careful attention must be paid to the interpretation of both the lag and effect size parameters. For example, Chapter 4 will refer to current smokers vs. not, when estimating a general CVD hazard due to smoking. It also explores the CVD hazard associated with packs per day smoked.

The latter assumes that the risk of CVD associated with 2 packs/day of smoking plateaus at twice the effective exposure of a 1-pack/day smoker. Additionally, the time needed for the 2 packs/day smoker's hazard to return to the level of the 1 pack/day smoker would be equal to the half-life years for the 1 pack/day unit risk. Discussion of the nuances behind these interpretations continues throughout this text.

The last term to define before deriving the models is that of the risk "profile" or "trajectory". The two words are used interchangeably to imply the history of exposure for an individual. In the context of time-varying exposures, an interesting trajectory might include that of the "on-again off-again" smoker, such that the CVD hazard associated with smoking is not monotonic throughout their lifetime, but

rather fluctuates based on the individual's set of known smoking periods. Thus, I refer to strictly increasing or decreasing effective exposure as a monotonic risk profile, the former applying to individuals who smoke throughout the course of a study and the latter illustrating the CVD hazard decreasing over time following successful cessation of smoking.

2.2 ONE PARAMETER EFFECTIVE EXPOSURE

The first version of the Effective Exposure model will be called "One Parameter Effective Exposure" (**OPEE**). It assumes the form closest to the OCM, such that it depends on a single parameter to define the rates of the incline towards steady state and decline back to zero.

2.2.1 Exposure Specification

Without loss of generalizability, the following derivation will be described in terms of a binary (yes vs. no) exposure. Let D equal 1 for a single exposure event that starts at time $t = b$ and ends at time $t = f$ for subject i, and λ denotes the rate parameterization of the lag. I can represent subject i's effective exposure $E_{it}(\lambda)$, at time t for a given lag, λ, by the following:

$$\begin{aligned} E_{it}\left(\lambda, b, f\right) &= D\left(1 - e^{-\lambda(t-b)}\right) * I\left(t\epsilon\left[b, f\right]\right) + D\left(1 - e^{-\lambda(f-b)}\right) e^{-\lambda(t-f)} * I\left(t > f\right) \\ &= D * \left[e^{-\lambda * \max(0, t-f)} - e^{-\lambda * \max(0, t-b)}\right] = D\left[e^{-\lambda z_2} - e^{-\lambda z_1}\right] \end{aligned} \tag{2.1}$$

where

$$z_1 = \max(0, t-b) = \begin{cases} t-b & \text{if } t > b \\ 0 & \text{otherwise} \end{cases}$$
$$z_2 = \max(0, t-f) = \begin{cases} t-f & \text{if } t > f \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

Figure 2.1 shows how the EE curve approaches steady state for different lags. Specifically, the shortest lag, in terms of half-life in days, quickly rises to steady state, versus the 900 days, which does not even reach half of the total steady state height upon exposure discontinuation[2].

Given a lag parameter and known start/stop times of exposure, $E_{it}(\lambda)$ represents the particular height or latent exposure level at any point in time for an individual. By exploiting this specification, I developed an iterative algorithm to calculate the effective exposure [for each subject at each measurement time] and use the corresponding value as the primary exposure metric in either the Cox proportional hazards (CPH) or pooled logistic regression (PLR) model framework. The odds ratio estimated by the PLR model is conditional on a subject's survival up to that point in time, thereby it is an approximation of the hazard ratio under appropriate conditions.(Ngwa et al., 2016; Cupples et al., 1988; D'Agostino et al., 1990) I will therefore refer to the maximum hazard as the effect size estimated by the EE approach. Details regarding the algorithm can be found in section 2.2.3.

Parameterizing EE by the half-life instead of the lag parameter produces an

[2]The 90-day half-life is used as a base case for the simulations in Chapter 3, where an explanation is provided for this selection. Meanwhile, 900 days represents the maximum follow-up time for the base case scenario, thus considered the upper bound for estimate-able half-lives in this type of study design.

equation that looks like:

$$E_{it}(h) = D\left(e^{-z_2 \log 2/h} - e^{-z_1 \log 2/h}\right) \tag{2.3}$$

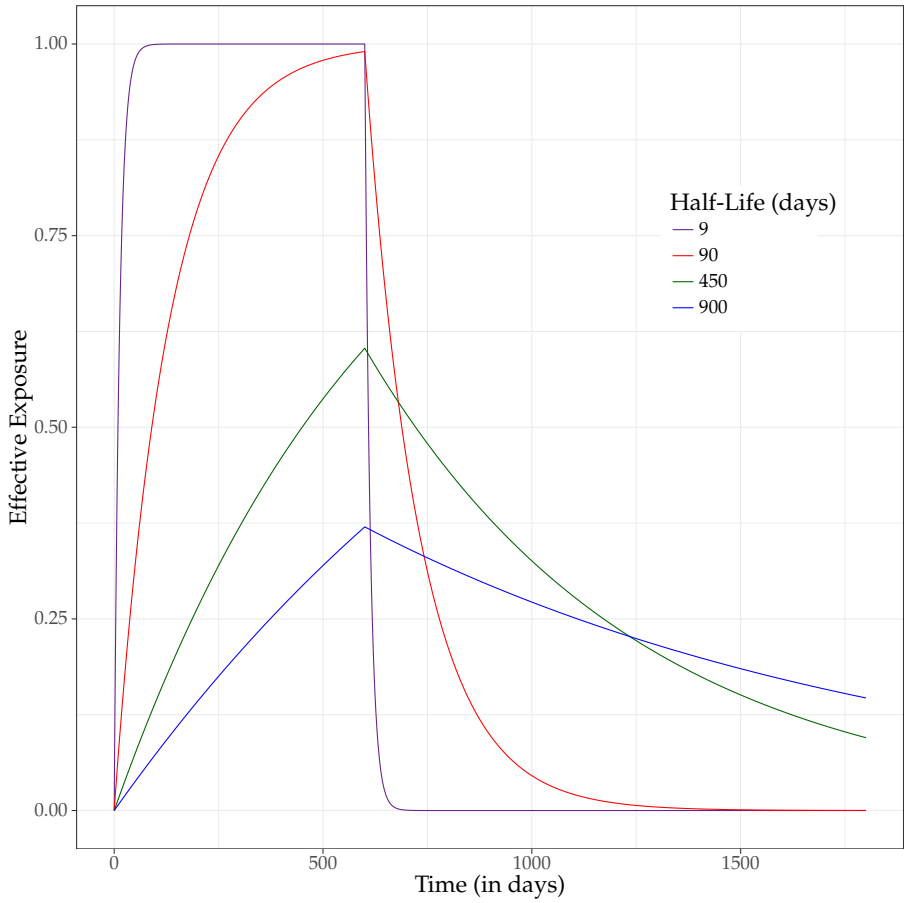


Figure 2.1: Effective exposure curves for different lag parameters, using the half-life definition in days.

2.2.2 Estimating Equations

2.2.2.1 Cox Proportional Hazards

To estimate the lag-time of an effect, I use the log-likelihood equation for CPH from equation (1.5) with Breslow's handling of ties to calculate the partial log-likelihood.

In a univariate model, I maximize the following with respect to λ and β:

$$\ell(\beta, \lambda, Y) = \sum_{k=1}^{K} \left[\sum_{j \epsilon R(t_k, Y_j=1)} \beta E_j(\lambda) - m_k \ln \left(\sum_{j \epsilon R(t_k)} e^{\beta E_j(\lambda)} \right) \right] \tag{2.4}$$

where j is the subject-time index for a particular risk set R. $R(t_k)$ denotes all subjects at risk at event time t_k, and $R(t_k, Y_j = 1)$ further restricts the risk set to sum across all m_k individuals with events at time t_k. $E_j(\lambda)$ is the calculated effective exposure for individual j in the risk set at t_k, as defined in chapter 1. $E_{it}(\lambda)$ and $E_j(\lambda)$ can be used interchangeably depending on the indexing - in this case j is the conditional index for individuals in the risk set R at time t_k, which could also be represented by the combined indices it_k.

Some important notes about my use of Cox's partial log-likelihood function:

1. The sum from $k = 1$ to K requires that risk sets be defined by both unique stop-time and strata, when assuming the baseline hazard also differs across strata.

2. The observed likelihood and log-likelihood in a multivariate model setting can be calculated by substituting $\hat{\beta} E_j(\hat{\lambda})$ with an individual's risk score from the multivariate model with Q total covariates and their corresponding estimated parameters, $\hat{\gamma_1}, ..., \hat{\gamma_q}$. – i.e. $\hat{r_j} = \hat{\beta} E_j(\hat{\lambda}) + \hat{\gamma_1} x_1 + ... + \hat{\gamma_q} x_q$

3. Further, I will assume that all of Andersen and Gill's conditions are met for the use of the Cox model with time-dependent covariates.(Andersen & Gill, 1982) That is, I assume that all of the hazards for the parameters in the model are proportional over time.

The predicted beta [and gamma parameters] in note 2 come from the CPH

model that is fit after fixing the lag parameter. Setting λ or h allows for the use of standard maximum likelihood estimation techniques, which do not require full explanation here. In summary, the model assumes the input effective exposure is the true exposure measure for which an estimate of effect is needed.

2.2.2.2 Pooled Logistic Regression

The equations required for a PLR analysis differ primarily in the construction of the likelihood and number of measurement intervals used. Here, the focus is on the predicted probability of an event at time t, rather than the hazard. Recalling (1.7), the univariate form of the logistic predicted probability of event for subject i at time t given a known lag of the effect λ can be written as:

$$\hat{p_{it}} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 E_{it}(\hat{\lambda})}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 E_{it}(\hat{\lambda})}} \tag{2.5}$$

As described in the CPH paragraphs, by fixing the lag parameter, estimation of $\hat{\beta}$ is conditional on the lag specification. Thus, standard methodologies for logistic regression would apply to the effect size parameter estimation.

2.2.2.3 Confidence Intervals

In order to obtain a standard error for the lag estimate, I can take the second derivative of the partial log-likelihood with respect to the effect size and lag parameters, to approximate the Fisher's Information matrix.(Hastie et al., 2009) To do this, and to be able to use the delta method, certain conditions and regularity assumptions need to be stated.

Regularity Conditions and Necessary Assumptions

1. The first derivative with respect to λ of the log of the EE exists and is finite
2. Let EE be expressed as a function of x and λ, such that x is the measured set of variables contributing to EE (x is the set $\{D, f, b, i, t\}$).
   (a) $f(x; \lambda)$ has bounded support in x and bounds do not depend on λ
   (b) $f(x; \lambda)$ has infinite support and is continuously differentiable
3. Y, the binary outcome of interest is independent and identically distributed across all subjects and risk sets.
4. EE is a smooth function.
5. The log-likelihood's first and second derivatives exist
6. The lag is normally distributed in the population
7. The likelihood is unimodal

Condition 1 is necessary with either condition 2a or 2b. Since $E_{it}(\lambda)$ has been defined as a function of exponentials, I am able to continue on the basis of conditions 1 and 2b. Even though the graph in figure 2.1 of EE appears to be non-differentiable at the point of discontinuation (i.e. change-point), the equation does not assume continuity through the change-point, therefore regularity holds as long as the piecewise elements are continuously differentiable.

The delta method allows me to use the inverse of Fisher's Information to approximate the asymptotically normal standard errors for the β and λ parameters. Each component of the Information Matrix has been derived in full and is presented in Appendix D. The formulas presented, below, are the first and second

derivatives of $E_{it}(\lambda)$ with respect to λ, which are used for the first and second derivative calculations of the log-likelihood with respect to λ.

$$\begin{aligned}\frac{\partial E_{it}(\lambda)}{\partial \lambda} &= D\left[z_1 e^{-\lambda z_1} - z_2 e^{-\lambda z_2}\right] \\ \frac{\partial^2 E_{it}(\lambda)}{\partial \lambda^2} &= D\left[z_2^2 e^{-\lambda z_2} - z_1^2 e^{-\lambda z_1}\right]\end{aligned} \tag{2.6}$$

Note that I can parameterize using the half-life, h, instead of λ, by imposing an additional derivative: $\left[\lambda = \frac{\log 2}{h} \longrightarrow \frac{\partial \lambda}{\partial h} = \frac{-\log 2}{h^2}\right]$

$$\begin{aligned}\frac{\partial E_{it}(h)}{\partial h} &= \frac{D \log 2}{h^2}\left[z_2 e^{-z_2 \log 2/h} - z_1 e^{-z_1 \log 2/h}\right] \\ \frac{\partial^2 E_{it}(h)}{\partial h^2} &= \frac{D\,(\log 2)^2}{h^4}\left(z_2^2 e^{-z_2 \log 2/h} - z_1^2 e^{-z_1 \log 2/h}\right) \\ &\quad - \frac{2D \log 2}{h^3}\left(z_2 e^{-z_2 \log 2/h} - z_1 e^{-z_1 \log 2/h}\right)\end{aligned} \tag{2.7}$$

2.2.3 Estimation Algorithm

All of the formulas, functions, and algorithms have been programmed using R version 3.2.3.(R Core Team, 2017) Package and function dependencies are outlined in my program documentation, though not everything will be included in the appendices of the dissertation.

The optimization methods I developed mimics the profile likelihood method in that a partial likelihood is computed by fixing a single parameter and maximizing across the rest. I fit the data simultaneously across several values of λ to iteratively search for the $\hat{\lambda}$ that maximizes the log-likelihood. This "guess-and-check" method may seem cumbersome, but it should be, theoretically, more efficient than a full profile likelihood fitting procedure. This will be especially important when considering a two-parameter space for the lag. Chapter 5 will cover more of the

strength and limitations comparing these two methods.

Proper data preparation eases the computational burden of the estimation process, by only requiring a single calculation to assign EE for each individual at each time slice. Part of the programming developed with this dissertation includes functions designed to create the necessary time and dosing vectors/variables (D, z_1, and z_2) that inform the $E_{it}(\lambda)$ calculation. The documentation is available in the Appendix B – for this particular function, please refer to "makeDVecs()".

Depending on the underlying event model, pooled vs. cox, the algorithm fits a lag-dependent OPEE against the outcome, in a univariate or multivariate setting, and returns the estimated effect $\hat{\beta}$ and model fit (log-likelihood and/or AIC). I utilize the "survival" package (version 2.38) in R to estimate $\hat{\beta}$ and other model coefficients.(Therneau, 2015)

The coxph() object outputs the predicted risk score (predicted probability of survival) for the final model fit, which can then be used to calculate the components of the Information Matrix – i.e. return estimates for the standard error in the lag parameter, and an adjusted standard error for the beta parameter, as well. For algorithms that employ the PLR models to estimate $\hat{\beta}$, I have chosen to utilize the "speedglm" package, which is efficient for generalized linear models fit to large data matrices.(Enea et al., 2015)

The equations presented in this section, so far, have only focused on the decay-rate parameterization for the lag, λ. Given the relationship between λ and h, the derivatives with respect to lambda, $\frac{\partial E_{it}(\lambda)}{\partial \lambda}$, and half-life, $\frac{\partial E_{it}(h)}{\partial h}$, actually have different values and magnitudes upon evaluation, due to differences in the measurement units[3]. To get estimates for the variability of the half-life parame-

[3] $E_{it}(h) = E_{it}\left(\frac{\log 2}{\lambda}\right)$

terized lag, I derived the respective formulae for substitution into the likelihood, log-likelihood, Score and Information Matrix functions appropriately. These are shown in Appendix D on page 181.

In the OPEE half-life searching algorithm, I track the number of iterations, the difference in the log-likelihood values between iterations, and which values have already been fit. For each step, if the difference between likelihoods falls below the algorithm's tolerance *or* the number of iterations exceeds the maximum iterations threshold, the loop is broken and the last maximum point is returned. The detailed steps are described by flowcharts in Appendix A. The overview flow for estimating the OPEE lag parameter is shown in figure 2.2.

This algorithm, as well as, the one described for extension to the two-parameter model, requires that the user/analyst provide an initial guess for the half-life or lag parameter, h_0. The first set of models fit use the initial guess and additional guesses by fixing the lower half-life to half of h_0's magnitude and the upper to twice the initial h_0's magnitude. Here, the magnitude refers to the value for the half-life that resides in an ordered list of three points: $H1 < H2 < H3$.

The first step of the algorithm compares the likelihoods arranged in order of the fixed half-life parameters, $H1 = \frac{h_0}{2}$, $H2 = h_0$, $H3 = 2h_0$, to determine which direction to travel based on the half-life that maximizes the likelihood. For example, if in the initial set of points, $\boldsymbol{H}$, the maximum likelihood corresponds to $H3$, then a new value is fit at twice the magnitude of the upper bound.

Once the maximum likelilhood has been centered, the upper bound side is tightened to fit a model for the half-life value that is equidistant from the maximum as on the lower-bound side, i.e., a new model is fit for $H3 = H1 + H2$. Following this step is a sequence of fits that narrows towards the maximum point

in the center or between maximized points, until one of the threshold criteria are met.

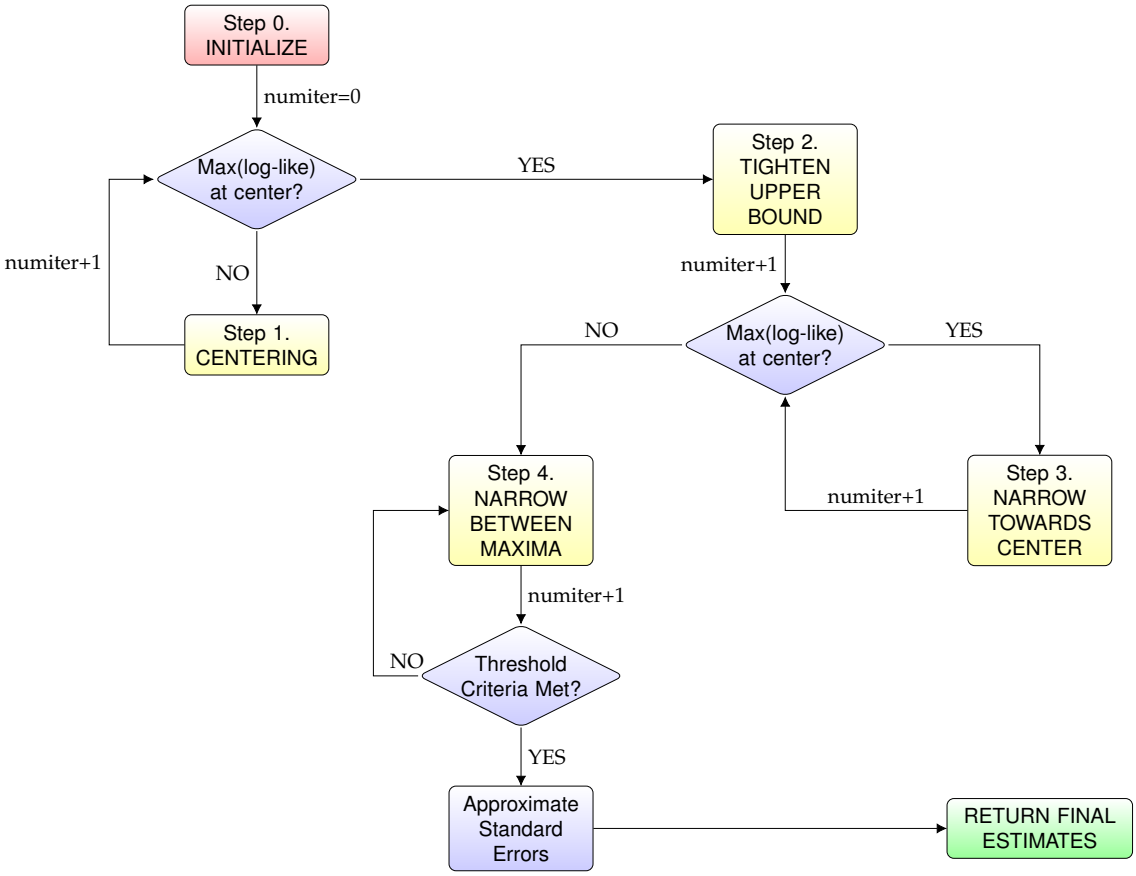


Figure 2.2: OPEE Algorithm Flowchart

Following determination of the half-life, the magnitude of association is estimated by fitting one last model, and the final AIC is adjusted to include one extra parameter – i.e., use the log-likelihood to re-calculate the AIC based on 2-parameters for a crude CPH model with estimated $\hat{\beta}$ and $\hat{h}$. The final estimates for effect size and lag combine with the individual's predicted probabilities to approximate the inverse of the negative Information matrix [Fisher's approximation of the covariance matrix]. The square root of the diagonals, or variances, reflects the estimated standard errors and the 95% confidence intervals are constructed using

the standard normal $Z = 1.96$. It is important to recalculate the $\hat{\beta}$ standard error as the output from the model fitting procedures does not account for the dependence of $\hat{\beta}$ on the lag.

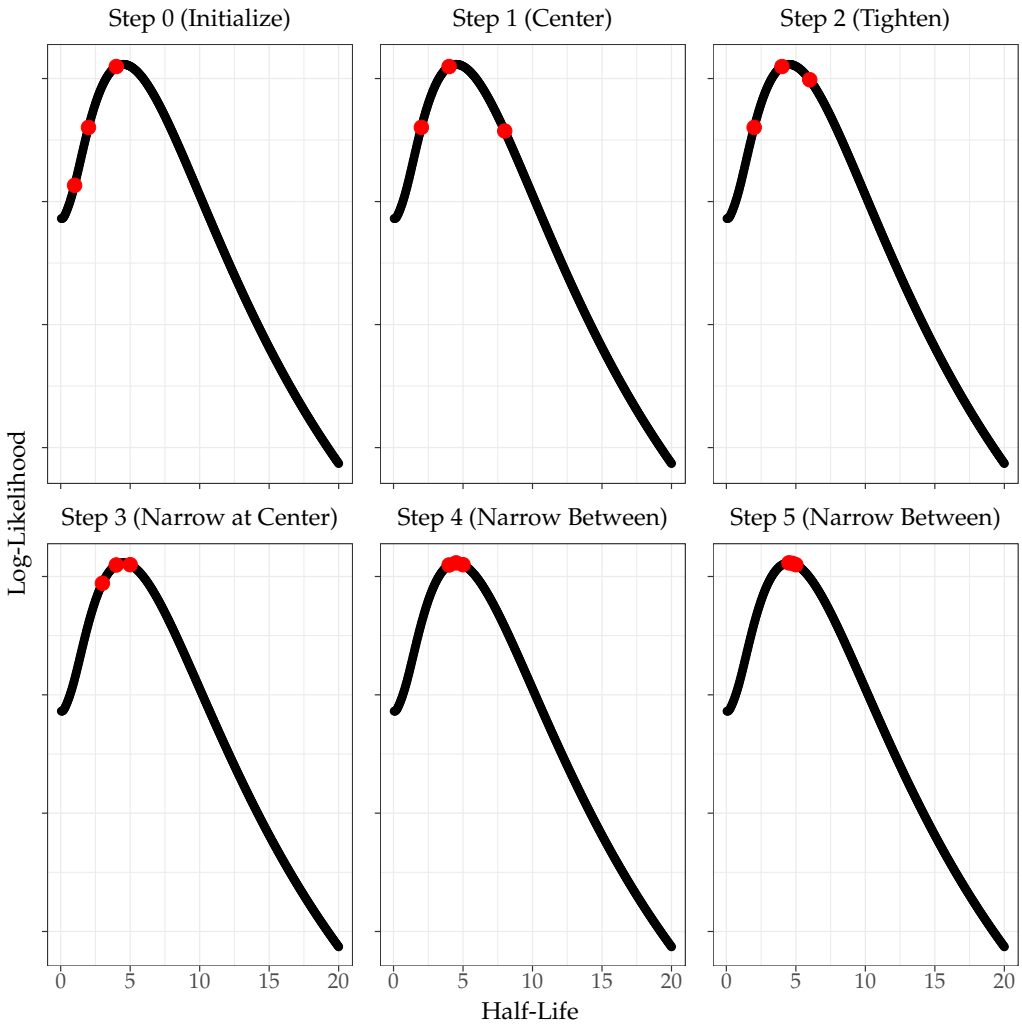


Figure 2.3: Example of OPEE Algorithm Steps on BWHS Binary Smoking Profile Log-Likelihood

A stepwise example of the OPEE algorithm can be seen in figure 2.3. The black log-likelihood curve [backdrop in all six panels] comes from fitting the CPH model of CVD due to the OPEE of dichotomous smoking on a restricted subset of the Black Women's Health Study (BWHS) sample. The x-axis represents fixed values

for the OPEE half-life parameter ranging from 0.05 to 20 years, in increments of 0.05 years amounting to 400 total model fit points. The y-axis represents the profile log-likelihood fit for the corresponding half-life, after adjustment for all the covariates and confounders (described in Chapter 4). In this example, the values for the log-likelihood have been omitted though the scale represents increasing log-likelihood in the upward direction; i.e. the maximum pictured on the black curve is the true profile log-likelihood maximum at a half-life of 4.5 years (as seen in the results table D.2).

I start by initializing a single half-life value that comes from a clinically-relevant range. Here, I start with 3 years, given the a priori belief that smoking's risk on cardiovascular disease should to return to normal between 2 and 5 years.(Rachet et al., 2003; Rosenberg et al., 1990) The following steps describe the example of the OPEE algorithm in Figure 2.3 moving along the top panels, left to right, and then left to right across the bottom panels. H denotes the vector of half-life parameters considered and L the corresponding log-likelihood values that are being compared. The index of each vector is denoted by the ordered value, while the superscript$^{(i)}$ implies the iteration step. The notation matches what can be seen in the flowcharts in appendix A.

- Step 0 - Initialize h_0, create vector of starting points, and calculate the corresponding maximum log-likelihoods:

$$H^{(0)} = \left[H1^{(0)} = \frac{h_0}{2}, \quad H2^{(0)} = h_0, \quad H3^{(0)} = 2h_0 \right]$$

$$L(H)^{(0)} = \left[L1^{(0)} = \ell\left(H1^{(0)}\right), \quad L2^{(0)} = \ell\left(H2^{(0)}\right), \quad L3^{(0)} = \ell\left(H3^{(0)}\right) \right]$$

$$\texttt{like1} \leftarrow L2^{(0)}$$

- Step 1 - Center and expand in the direction of the maximum, calculate the newest index likelihood, and determine the tolerance.

$$\texttt{like2} = \max\left(L(H)^{(0)}\right) = L3^{(0)}, \quad tol = |\texttt{like1} - \texttt{like2}|, \quad \texttt{like1} \leftarrow \texttt{like2}$$

$$H1^{(1)} = H2^{(0)}, \quad H2^{(1)} = H3^{(0)}, \quad H3^{(1)} = 2H3^{(0)}$$

$$L1^{(1)} = L2^{(0)}, \quad L2^{(1)} = L3^{(0)}, \quad L3^{(1)} = \ell\left(H3^{(1)}\right)$$

$$\max\left(L(H)^{(1)}\right) = L2^{(1)} = L3^{(0)} = \texttt{like1}$$

- Step 2 - Maximum log-likelihood at center. Tighten upper-bound to equal distance from center as lower-bound:

$$H1^{(2)} = H1^{(1)}, \quad H2^{(2)} = H2^{(1)}, \quad H3^{(2)} = H1^{(1)} + H2^{(1)}$$

$$L1^{(2)} = L1^{(2)}, \quad L2^{(2)} = L2^{(1)}, \quad L3^{(2)} = \ell\left(H3^{(2)}\right)$$

$$\texttt{like2} = L3^{(2)}, \quad tol = |\texttt{like1} - \texttt{like2}|$$

$$\max\left(L(H)^{(2)}\right) = L2^{(2)} = \texttt{like1}$$

- Step 3 - Maximum log-likelihood still at center. Narrow bounds towards center by half the distance:

$$d^{(3)} = H2^{(2)} - H1^{(2)} = H3^{(2)} - H2^{(2)}$$

$$H1^{(3)} = H2^{(2)} - \frac{d^{(3)}}{2}, \quad H2^{(3)} = H2^{(2)}, \quad H3^{(3)} = H2^{(2)} + \frac{d^{(3)}}{2}$$

$$L1^{(3)} = \ell\left(H1^{(3)}\right), \quad L2^{(3)} = L2^{(2)}, \quad L3^{(3)} = \ell\left(H3^{(3)}\right)$$

$$\texttt{like2} = \max\left(L(H)^{(3)}\right) = L3^{(3)}, \quad tol = |\texttt{like1} - \texttt{like2}|, \quad \texttt{like1} \leftarrow \texttt{like2}$$

- Step 4 - Maximum no longer at center (after step 3), transition to search between maximae:

$$H1^{(4)} = H2^{(3)}, \quad H2^{(4)} = \frac{H2^{(3)} + H3^{(3)}}{2}, \quad H3^{(4)} = H3^{(3)}$$

$$L1^{(4)} = L2^{(3)}, \quad L2^{(4)} = \ell\left(H2^{(4)}\right), \quad L3^{(4)} = L3^{(3)}$$

$$\texttt{like2} = \max\left(L(H)^{(4)}\right) = L2^{(4)}, \quad tol = |\texttt{like1} - \texttt{like2}|, \quad \texttt{like1} \leftarrow \texttt{like2}$$

- Step 5 - Tolerance threshold not met, continue searching between maximae.

$$H1^{(5)} = H2^{(4)}, \quad H2^{(5)} = \frac{H2^{(4)} + H3^{(4)}}{2}, \quad H3^{(5)} = H3^{(4)}$$

$$L1^{(5)} = L2^{(4)}, \quad L2^{(5)} = \ell\left(H2^{(5)}\right), \quad L3^{(5)} = L3^{(4)}$$

$$\max\left(L(H)^{(5)}\right) = L1^{(5)} = L2^{(4)} = \texttt{like1}$$

To avoid calculating a "zero" tolerance when the maximum remains the same in these steps, set the comparison to be between the first and second maximum log-likelihood values.

$$\texttt{like2} = \text{second}\max\left(L(H)^{(5)}\right) = L2^{(5)}, \quad tol = |\texttt{like1} - \texttt{like2}|$$

Threshold for tolerance reached. Stop algorithm and compute standard error approximations.

2.3 TWO PARAMETER EFFECTIVE EXPOSURE

With certain exposures, it may not be appropriate to assume that the effective exposure accumulates at the same rate as it goes away. For example, upon exposure to lead, the exposure distributes to various compartments in the body, including in the bones, where it accumulates and is stored as a source of "continual internal exposure".(Gulson et al., 1995; Flora et al., 2012) While this exposure can be measured in the blood, the true quantity that persists (to affect us) as our bones demineralize is unknown or requires high-level expensive technology to be measured.

One recommended approach to lowering the levels of lead in the body (i.e. treating lead poisoining) is chelation therapy.(Centers for Disease Control and Prevention, 2015) The chelating agent attaches to heavy metals like lead, flushing them out of the system. In this particular situation, it may be possible to lower the impact of lead through treatment, but I would not expect the time-to-reduction to be the same as the rate associated with increasing disease hazards from initial lead exposure. For this type of an exposure, I introduce a two-parameter effective exposure (**TPEE**) model as an extension of the OPEE, where the lags differ for the incline and decline of the EE curve.

Figure 2.4 illustrates what a curve with differing rates of accumulation and decay, or incline and decline, may look like, compared to the single-lag effective exposure. The blue solid line represents an EE curve with a single half-life parameter of 90 days, while the red dashed line represents the EE curve under the same incline with a decline parameter half-life of 900 days. Both curves assume continuous exposure for 180 days prior to discontinuation.

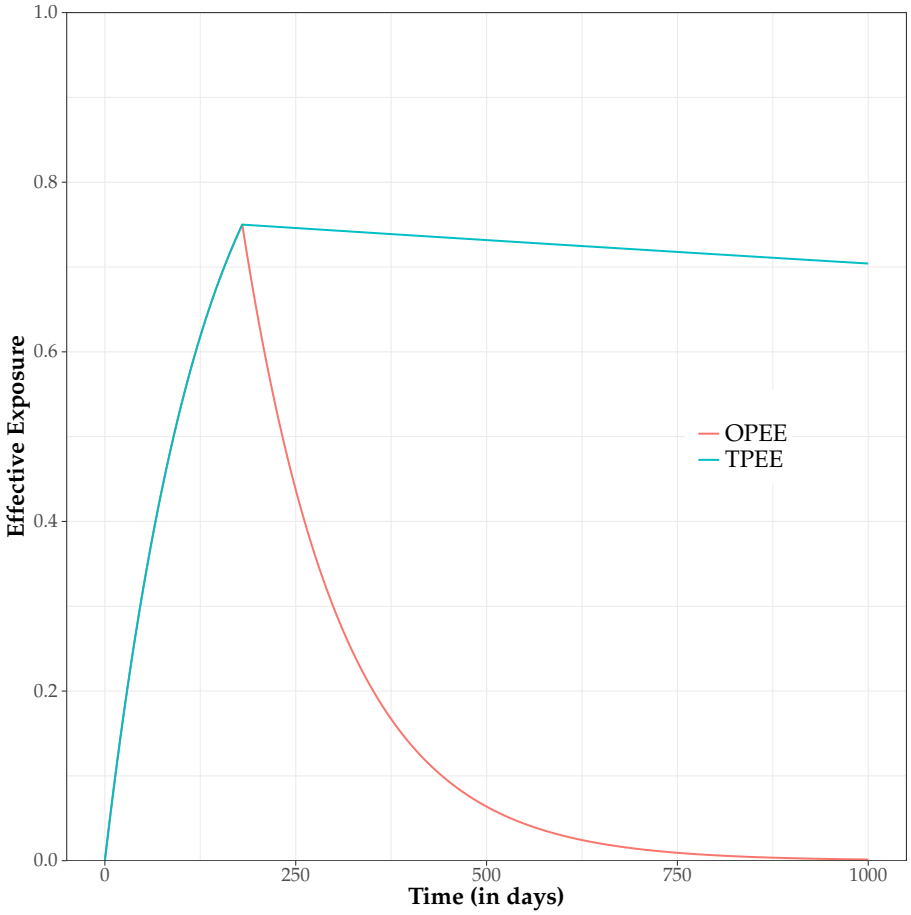


Figure 2.4: Comparison of the effective exposure curves over time for the one- and two-parameter approaches.

2.3.1 Exposure Specification

The OCM function is unique, in that it assumes a piecewise structure dependent on change-points. This assumption also implies that after an infusion is discontinued the accumulated total – from which the decay starts – is considered fixed.

Let Λ denote a set of two lag parameters, λ_1 and λ_2. Reverting back to the original form of (1.4), since $(f - b) = (t - b) - (t - f) = z_1 - z_2$, the TPEE equation follows:

$$E_{it}\left(\Lambda = (\lambda_1, \lambda_2)\right) = D\left[1 - e^{-\lambda_1(z_1 - z_2)}\right] e^{-\lambda_2 z_2} \tag{2.8}$$

Where the definitions of z_1, z_2, and D remain the same, but now λ_1 is the parametric lag for the incline and λ_2 stands in for the decline lag. Parameterizing by the half-life parameter produces an equation that looks like:

$$E_{it}(h_1, h_2) = D\left[1 - e^{-(z_1 - z_2)\log 2/h_1}\right] e^{-z_2 \log 2/h_2} \tag{2.9}$$

2.3.2 Estimating Equations

In the OPEE framework, the likelihood and log-likelihood equations contain the EE functions $E_{it}(\lambda)$ or $E_j(\lambda)$. These can also be written as $E_{it}(\Lambda)$ and $E_j(\Lambda)$, which I have denoted as the EE function of the lag parameters λ_1 and λ_2 in equation (2.8).

The overall forms of the likelihood and log-likelihood equations (2.4) and (1.6) do not change. This is because $E_{it}(\Lambda)$ still represents the total Effective Exposure at time t for subject i. Substituting $E_{it}(\Lambda)$ for $E_{it}(\lambda)$, however, does require the calculation of new first and second derivatives with respect to both parameters and their combination. Since the effective exposure is assumed to be a quantity that is independent of the other model covariates, once the underlying equations are appropriately adjusted, the calculation of the likelihood, with respect to the other variables in the model, remains unchanged.

Full derivations of the Score and Information Matrix are provided in Appendix D (starting on page 181) for both CPH and PLR likelihood functions. This includes framework for the lag, Λ, and half-life, $\boldsymbol{H}$, parameterizations. Again, medical researchers may prefer the half-life parameterization's estimate interpretation and corresponding variability measure.

2.3.3 Estimation Algorithm

The estimation algorithm for the TPEE approach relies on a grid search across combinations of the incline and decline parameters. Figure 2.5 shows the big picture steps of the TPEE algorithm, which is similar to the one presented to the OPEE algorithm.

In the first few steps the user specifies a single initialized value and the process starts by halving and doubling this value for the first set of comparisons. This 3-value vector is assumed to be the same for both incline and decline, leading to 9 total combinations corresponding to a set of incline and decline coordinates. The full algorithm is described by flowcharts in Appendix A.

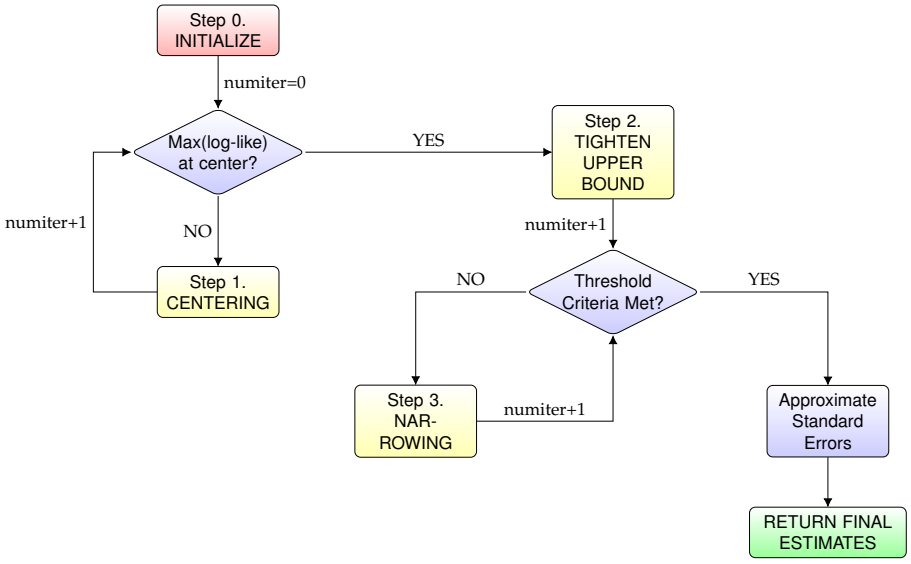


Figure 2.5: TPEE Algorithm Flowchart

Depending on the pair of half-lives that produces the maximum log-likelihood in the set, each parameter's vector is expanded in the direction of the maximum by halving or doubling – again, this is similar to the OPEE step 1 "centering". Upon centering the maximum log-likelihood at the 5th index pair, the algorithm con-

tracts both parameter upper bounds to be equidistant to the center (i.e. "tightening"). New CPH/PLR models are fit for the corresponding incline-decline EE for which the likelihood has not yet been calculated.

The last sequence of steps iterates model fits for EE based on the first and second maxima of each parameter, holding the other parameter constant (i.e. "narrowing"). At this point, should the likelihood surface at a set of coordinates reach a ridge[4] the algorithm is stopped and the last maximum location is returned as the final estimated pair of half-lives. While not explicitly described here, each step of the algorithm also checks the tolerance and number of iterations, breaking the loop and returning the last maximum likelihood coordinates when either threshold are met.

Following the algorithmic search, I use the final pair of likelihood-maximizing half-lives to estimate the maximum hazard parameter, and calculate the 95% confidence intervals using the normally-approximated standard errors for the effect and lag parameter estimates. The updated Score and Information Matrix equations for approximating these standard errors can be found in the Appendix D.

Figure 2.6 provides a visual example of the steps in the TPEE algorithm using data from the BWHS restricted sample binary smoking's profile log-likelihood surface. The colored background is meant to show the contour of the 3-dimensional surface, with the lines representing the PLL's joint 90, 95, and 99% confidence bounds. The lighter shading indicates larger values of log-likelihood, such that a peak occurs in the center of the contour bounds.

Step 0 starts with initial values of 1.5, 3, and 6 years for both half-life lag parameters. The top right quadrant of points are carried over to the left bottom quadrant

[4]This can be seen when two or more coordinates produce the exact same model fit.

in step 1, and these points remain the same in step 2 as the upper bounds are tightened. Step 3 searches for the true maximum log-likelihood somewhere in between step 2's top-right quadrant, and steps 4 and 5 narrow between maximae even further.

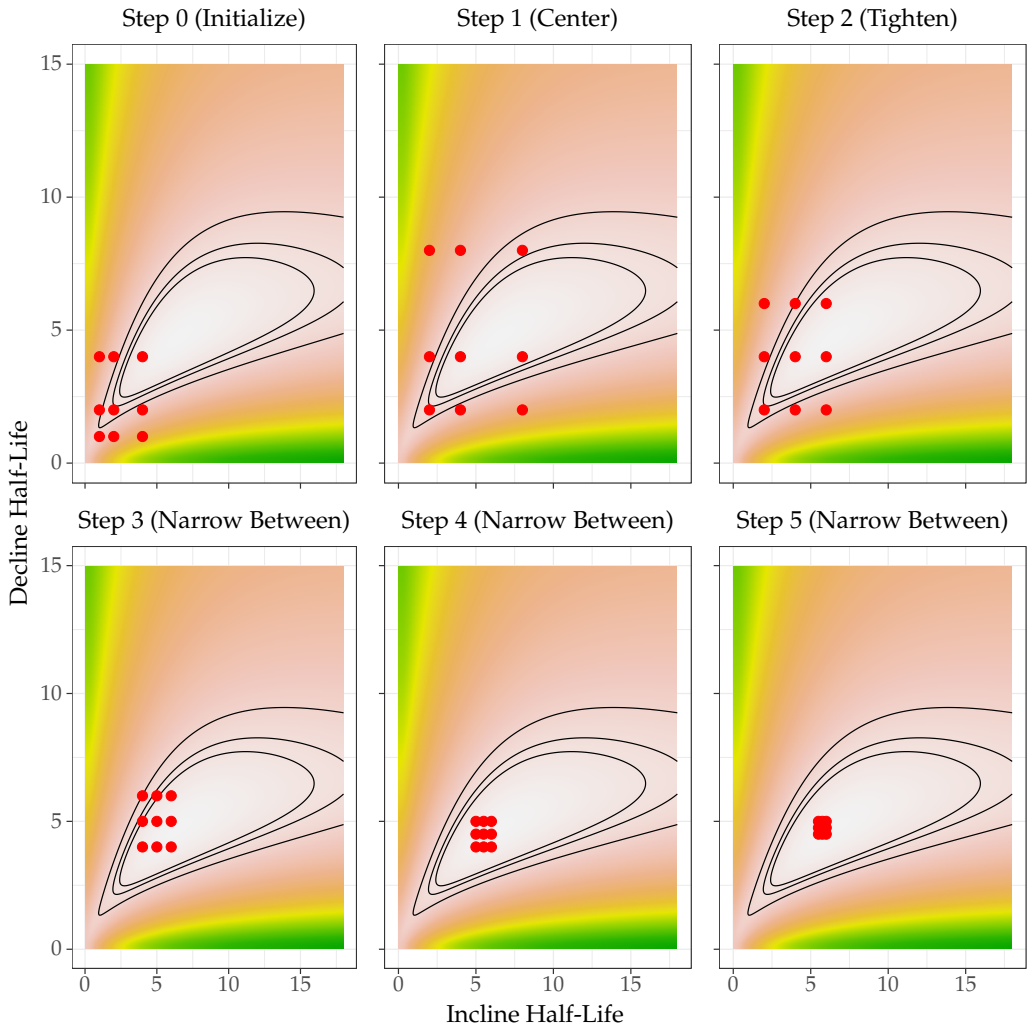


Figure 2.6: TPEE Algorithm Steps on BWHS Binary Smoking Profile Log-Likelihood Contour

2.4 MULTIPLE DOSING SCHEME

One of the main strengths of my proposed methodology is its ability to handle changing exposures over time. This is slightly more nuanced than the simple protracted exposures approaches, as my method allows for subjects to discontinue and/or start new regimens that may represent the same underlying action mechanism.

The term regimen is used to describe the particular exposure [instance] for the set of parameters that dictate start and stop times, and level of the dose. Let z_3 and z_5 denote the time since start of a second and third regimen, and z_4 and z_6 denote the time since stopping the second and third regimens, respectively.

Figure 2.7 illustrates the life-trajectory for changing regimens[5] of the effective exposure. Specifically, it demonstrates what the sum of all the effective exposure curves looks like over time, where the declining effect due to regimen 1 still contributes to an increased risk while the subject is on regimen 2 and has not, yet, reached steady state. Since the maximization procedure assumes a single set of parameters for the effect's lag, the summation of the individual regimen's effective exposure curves are still differentiable.

The figure assumes a single parameter effective exposure model. As discussed in the description for the TPEE formulation, since $E_{it}(\lambda)$ is a function of the lag, the algorithms can still be used. The multiple regimens are "simply " summed within specification for a given subject's exposure total at time t.

The solid lines in figure 2.7 represent a single set of dosing regimens for an individual. In this example, I use an OPEE half-life of 5.85 years, so simplify interpretation, assume that the doses represent the packs per day smoked. This subject

[5] as denoted by the vertical dashed lines

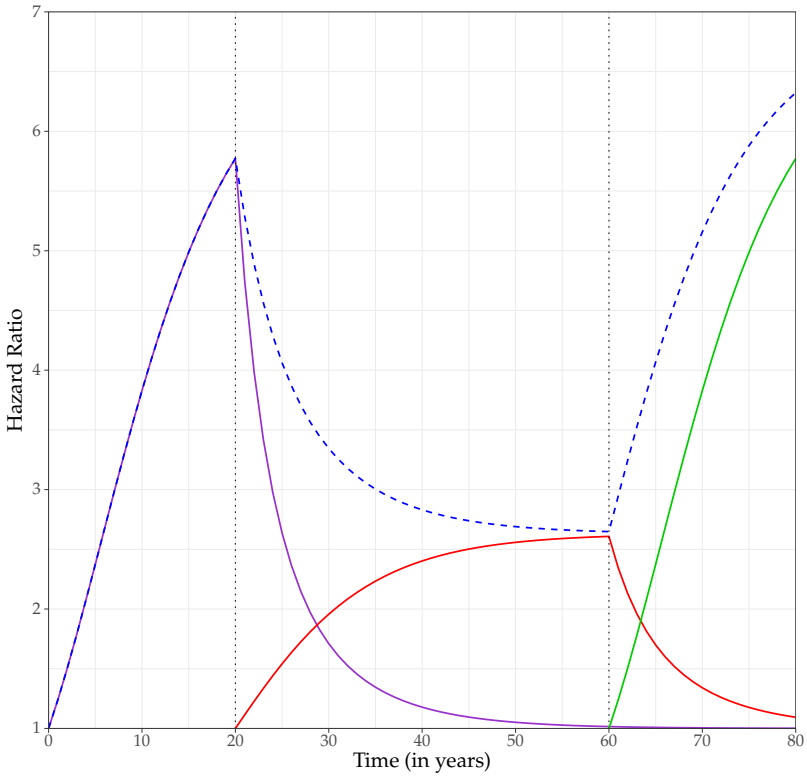


Figure 2.7: Demonstration of the accumulation of protracted exposures under OPEE model with a half-life of 5.85 years, a plateau hazard ratio for 1-unit effective exposure of 2.63, and different steady state dose levels.

smoked 2 packs per day from the start of follow-up for 20 years, and again from 60 years until the end of follow-up. From years 20 to 60, the subject reduced smoking to 1 pack per day. The dashed blue line represents the hazard ratio compared to a never smoker over time, with the black vertical dashed lines representing the change-points in dosing levels. The figure represents a situation where there is a single lag-parameter defined by a half-life of 5.85 years, and the CVD hazard from smoking 1 pack per day, for an extended period of time, is 2.63 times the hazard of never smokers.

The risk profile shown in figure 2.7 can be denoted by equation (2.10). Specifi-

cally, $E_{it}(\lambda)$ becomes a sum function of the individual dosing exposures. From the example in the figure, the subject is exposed from years 0 to 20 and 60 to 80 at a dose level of 2, while being exposed at a dose level of 1 from time 20 to 60. For the figure's example, the EE at time t can be calculated as:

$$\begin{aligned} E_{it}^{(tot)}(\lambda) = & D_1 \left(1 - e^{-(z_1 - z_2)\lambda}\right) e^{-z_2\lambda} \\ & + D_2 \left(1 - e^{-(z_3 - z_4)\lambda}\right) e^{-z_4\lambda} \\ & + D_3 \left(1 - e^{-(z_5 - z_6)\lambda}\right) e^{-z_6\lambda} \end{aligned} \tag{2.10}$$

where

$$z_1 = \begin{cases} t & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases} \qquad z_4 = z_5 = \begin{cases} t - 60 & \text{if } t > 60 \\ 0 & \text{otherwise} \end{cases}$$

$$z_2 = z_3 = \begin{cases} t - 20 & \text{if } t > 20 \\ 0 & \text{otherwise} \end{cases} \qquad z_6 = \begin{cases} t - 80 & \text{if } t > 80 \\ 0 & \text{otherwise} \end{cases}$$

Using the piecewise notation, it becomes clear that the first and second derivatives can be readily calculated for each individual exposure occurrence. Let $E_{it}^{(1)}(\lambda)$, $E_{it}^{(2)}(\lambda)$, $E_{it}^{(3)}(\lambda)$ be the EE components for each of the exposure events/period.

$$\begin{aligned} E_{it}^{(1)}(\lambda) &= D_1 \left(1 - e^{-(z_1 - z_2)\lambda}\right) e^{-z_2\lambda} \\ E_{it}^{(2)}(\lambda) &= D_2 \left(1 - e^{-(z_3 - z_4)\lambda}\right) e^{-z_4\lambda} \\ E_{it}^{(3)}(\lambda) &= D_3 \left(1 - e^{-(z_5 - z_6)\lambda}\right) e^{-z_6\lambda} \\ E_{it}^{(tot)}(\lambda) &= E_{it}^{(1)}(\lambda) + E_{it}^{(2)}(\lambda) + E_{it}^{(3)}(\lambda) \end{aligned} \tag{2.11}$$

As mentioned previously, the λ can be interchanged with $\log 2/h$, and $E_{it}^{(tot)}(h)$

represents the longitudinal function of EE based on this parameterization. To transition from OPEE to TPEE, I can substitute the single parameter with (λ_1, λ_2) or (h_1, h_2) in equation (2.11), updating the piecewise component exposures that feed into $E_{it}^{(tot)}$.

2.4.1 Estimation and Algorithm Modifications

Assuming that the hazards are additive, and that the maximum hazard plateaus at a single-unit of the EE, then the formulation for the likelihood functions and estimation algorithms stay the same. Specifically, the likelihood depends on the quantity of $E_{it}(\Lambda)$, that is computed for a given sequence of exposures under fixed lag parameter(s). To account for the added components, I have developed a series of functions that calculate the accumulated $E_{it}(\Lambda)$ for each subject at each event or interval time. The details are presented in Appendix B under the "C1fun.h()" module.

2.5 INTERPRETATION PARADIGM

The remainder of this chapter provides insight on the semantics required for proper and intelligent interpretation of the resulting estimates.

In figure 2.8 the solid lines represent the hazard over time for two women who successfully quit smoking after 30 years of prior exposure. These are monotonic trajectories reflecting the decline in CVD hazard from the 30-year 2 packs/day and 1 pack/day smokers, i.e. women with different dosing levels of the EE. The corresponding y-axis is the left-hand "Hazard Ratio" (HR), on which one can see that the hazard changes 2-fold or by half at the 5.85-year half-life.

Meanwhile, the corresponding dashed lines are for the same individuals, con-

sidering the Effective Exposures over time – as shown by the right-hand y-axis. Here, the individual's HR over time compared to unexposed is not parallel to the EE curve, implying the rates are different for the two scales. One important note is that the floor of the HR axis is at 1, while the EE bottoms-out at 0.

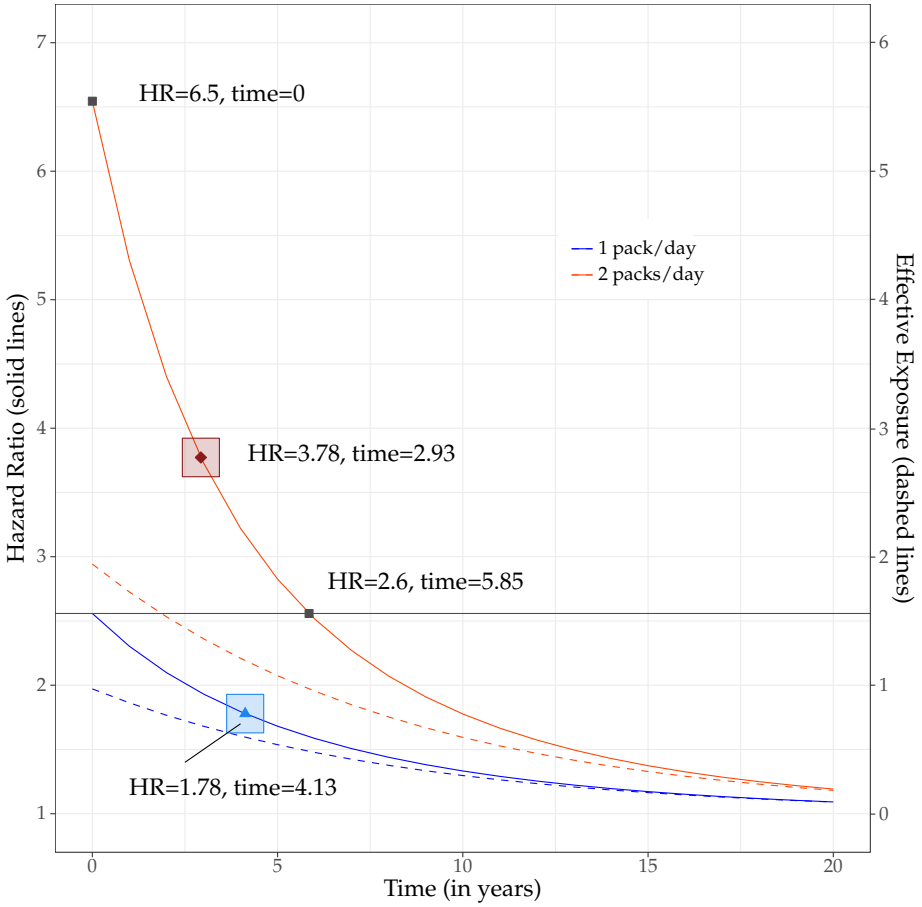


Figure 2.8: Comparisons of the rate of change in risk as measured on the Hazard Ratio vs. Effective Exposure scales.

Given the desire to understand change in excess hazard, as estimated by my models, I have added another function, specifically for calculating an individual's predicted time-to-risk reduction. The "solve.time()" function (documentation in Appendix B) allows the user input the model's parameters, the individual's

amount of time exposed (i.e. the starting hazard), and the desired proportional decrease on the excess hazard scale (i.e. HR-1).

The following example is taken out of context for the purposes of providing an illustration of the interpretations that come out of these computations. The complete analysis for these numbers and this corresponding paragraph can be found in Chapter 4 (on page 113).

For a consistent 2 packs/day [dose of exposure] smoker of 30 years [time exposed], after accounting for other risk factors of CVD [model-based estimates], the CVD hazard associated with a woman's smoking exposure is 6.5 times [HR at start] that of her counterfactual never smoker. Using the same adjusted model, a woman who smoked 1 pack/day for 30 years is at 2.6 times the never smoker's hazard of CVD. A 50% reduction [desired HR reduction] in excess hazard for these same 2- and 1-pack/day smokers, would take 2.9 and 4.1 years [calculated] following complete and successful cessation. The corresponding ending hazard ratios would be 3.8 and 1.8 [HR at end], respectively.

Alternatively, after 5.85 years [estimated half-life, table D.2] of complete and successful quitting of smoking, the 2-pack/day smoker's CVD hazard is expected to reach the 1 pack/day smoker's hazard, i.e., after the half-life number of years, the risk is reduced by 50%. In terms of reduction in hazard ratio or excess hazard, this implies that the hazard ratio of CVD for a 2 pack/day smoker compared to a never smoker reaches the hazard ratio for the 30-year 1 pack/day smoker compared to a never smoker, after 5.85 years of no smoking exposure.

All of the points described in the preceding paragraphs have been annotated in figure 2.8.

CHAPTER 3
Simulation Study

The purpose of this chapter is to demonstrate the strengths and limitations of the effective exposure estimation methods under known lag and effect size parameters, by simulating data representative of real-world examples. I explore a wide range of scenarios to identify the type of data that would lead to unbiased and robust estimation of the parameters of interest.

In these simulation studies, I also compare several analytic methodologies in terms of coverage probabilities and estimation bias. The aim of this chapter is to show that the model performance improves with information content – i.e. the more information, the better. In particular, the information necessary for estimating the half-life revolves around the proportion of subjects and subject-time spent "in transition" between steady states. Meanwhile, information content for the hazard ratio estimate comes from the proportion of subjects and subject-time spent at "maximum risk" (or at the hazard's plateau).

I will start by describing the types of scenarios considered, including one that is based on real data from the Black Women's Health Study (BWHS). I will then move through the data generation processes and analytic approaches considered. The "Results" section will focus on patterns and primary findings from the simulations performed, with some concluding remarks regarding the strengths and limitations of the simulation study.

3.1 SCENARIO SPECIFICATIONS

The initial setup for the 1-parameter simulation is *loosely* based on the association between corticosteroids (CS) use and risk of fracture. CS use has been shown to

leach calcium from the bones and with an increase in the risk of fracture within the first 3 months of treatment.(Mitra, 2011) Those on oral CS over an extended period of time are at roughly 1.5 times the risk of fracture as those not taking CS.(Van Staa et al., 2000) Following discontinuation and after accounting for duration of use and dose, excess risk decreases towards the baseline risk over the course of a year.(Van Staa et al., 2000; Vestergaard et al., 2008)

3.1.1 Base Case: One-Parameter Effective Exposure

The base case simulation (BC1) represents an "optimal" study design in which one would expect the proposed methods to work consistently well. For the sake of simplicity, I have chosen to look at the one-parameter effective exposure (OPEE) model with the half-life parameterization of the lag. Following the CS example, I set the half-life to 3 months or 90 days. Recalling that the OCM and OPEE models assume steady state is reached after 4-5 half-lives, a group of individuals starting at maximum hazard would be expected to return to their baseline hazard around 360-450 days or roughly one year. Meanwhile, newly exposed individuals would reach half of the prolonged exposure hazard after 90 days.

BC1 includes subjects with one of three possible exposure trajectories: those who are never exposed serving as the controls ("ctrl"), one group that initiates use at baseline, and another that discontinues use at baseline after having been exposed for at least two years[1]. The latter two are referred to as the "up" and "down" groups, respectively.

The sample is large with 10,000 subjects in each group (N=30,000 total), and subjects are followed for a period of 900 days. Equivalent to just under 3 years, the

[1]Exact time is set to 900 days to keep durations consistent

900 days mark, or 10 half-lives, also implies that both transitioning populations, in this sample, will achieve steady state within the follow-up time. This time-frame is quite standard for administrative data studies in which event times may be known to the day, but not hour, of occurrence.

I consider a simple binary or dichotomous exposure scheme, with an underlying relative effect of a 50% increased hazard, or 1.5 times the hazard for lifetime-exposed versus unexposed individuals. At 90 days, the *up* group [2] is considered to have reached 50% of the maximum hazard, which translates to a 1.22-fold hazard of event compared to never exposed individuals. Additionally, I set the study-wide prevalence of the outcome to 10%, meaning that controls in the study are also at risk for event.

Figure 3.1 shows four panels of simulation scenarios, of which the upper left-hand corner is the BC1 set of trajectories. Similar to figure 2.8, the left-y-axis and solid lines represent the EE curve, while the right-y-axis and dashed lines reflect the hazard ratio over time. On day 90, both risk curves (EE and HR), for the *up* and *down*, groups intersect. After 90 days, those in the *up* group surpass the *down* group in EE.

3.1.1.1 Sample Sizes

In order to understand the effect of sample size on the estimation algorithm performance, I consider the base case scenario with 1,000 (N1k), 10,000 (N10k) and 100,000 (N100k) total participants. Since the study design for these simulations requires balanced groups, the single remaining individual, following the thirds split, is put into the control group.

[2] individuals that became exposed at the start of follow-up

For the smaller samples N1k and N10k, I would expect the estimation of both lag and HR parameters to be more biased with additionally worse coverage for the lag parameter(s), compared to the BC1 scenario. Meanwhile, the N100k scenarios should improve in both performance metrics, versus to the BC1 scenario. The latter set is, naturally, more computationally expensive, but corresponds to the size of administrative data. None of these sample sizes is comparable to clinical study data, but looking at the smaller sets may shed light onto the generalizability of these methods to the standard prospective cohort study design.

3.1.1.2 Dosing levels

A natural question that arises from a dose-dependent scenario is whether the binary model will appropriately identify the lag. My initial hypothesis is that the half-life may be underestimated, or negatively biased, when performing the OPEE and TPEE algorithms with the binary exposure. In a balanced design the resulting effect size estimate should be a weighted average of the different effect levels.

The set of simulations (DoseMods) that aims to account for differences in dosing of exposures is considered as a variation on the base case. The three groups remain the same, though half of the *up* and *down* groups are assigned a dose-level of 2. Specifically, 5,000 of the *up* group subjects are assumed to plateau at twice the risk level, while 5,000 of the *down* group subjects start the study at twice the hazard. This translates to an hazard ratio of 2.25 compared to the controls, which can be seen in the upper right corner panel of figure 3.1.

3.1.1.3 *Adding a Fourth Group at Steady State Risk*

To be able to differentiate between those who remain exposed and those on the decline, I also consider a set of simulations (FourG) with a fourth group that does not discontinue exposure ("on"). The hypothesis here is that the increased amount of information pertaining to continued exposure would allow for more accurate estimation of both parameters. Specifically, I hypothesize that the estimated hazard plateau will be less biased.

Figure 3.1's bottom left corner shows the same set of trajectories as the base case (BC1, upper-left) with the addition of the constant *on* group.

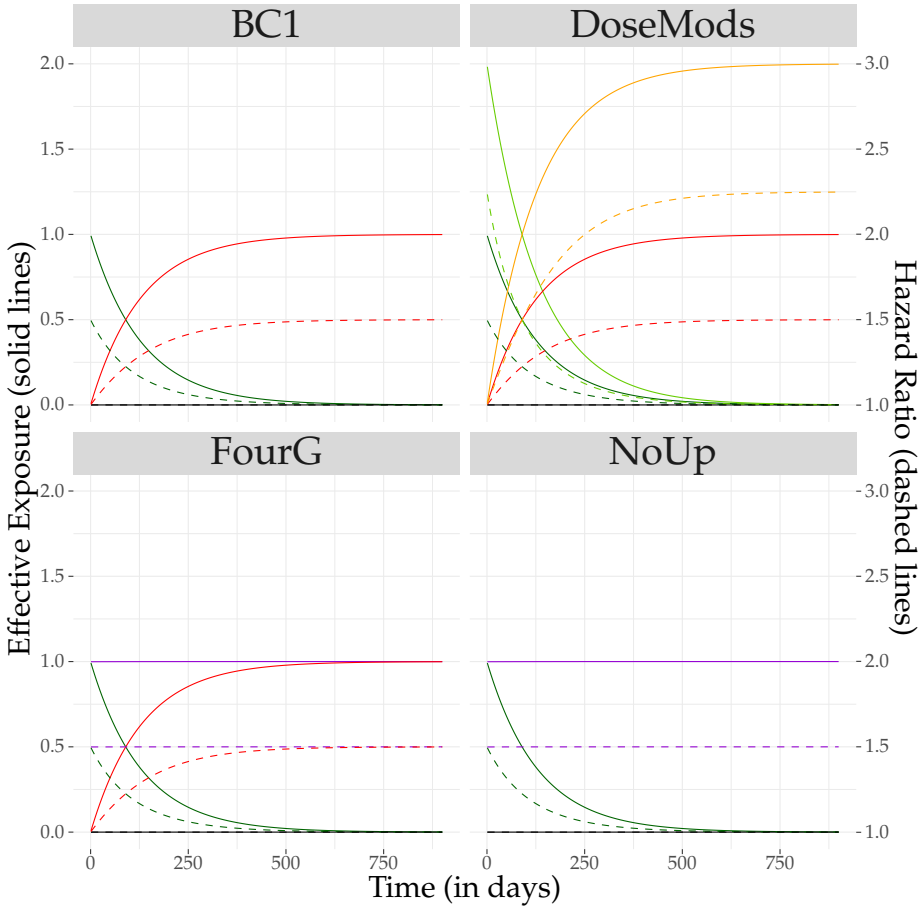


Figure 3.1: Plots for One-Parameter Base Case Simulation Scenarios

3.1.1.4 *No Initiators*

The remaining panel (bottom right) in figure 3.1 demonstrates a set of monotonic trajectories that does not include the *up* group (NoUp), but does have individuals who remain exposed throughout follow-up (*on* group). When considering studies that focus on discontinuation of exposure, such as quitting smoking or coming off of oral CS, to control for confounding a sample may exclude subjects that initiate, so as to focus the analyses on the benefits of cessation.

For example, in chapter 4, I discuss the "restricted" sample of BWHS participants, which is comprised of women who are smokers and non-smokers throughout the study, in addition to a group of women who successfully stopped smoking [3]. Part of the goal for this analysis is to mirror previous approaches that restrict the sample, like Rachet et al. (2003). I hypothesize that in this case the estimated time to reduction might be shorter (decline estimate negatively biased), and the incline estimate will have large bias – due to the lack of information about the upwards transition.

In addition to the breadth of applications this variation can generalize, it also represents a situation where the TPEE algorithm should fail. This is due to the fact that there is little information about the incline parameter – since no one is transitioning upward in risk.

3.1.1.5 *High Risk*

Not much explanation is needed for the high-risk variation (HR5) of the base case. Here the input hazard ratio is a 5-fold risk of event for those are steady state risk. I hypothesize that this stronger effect size provides "more information" and allows

[3] As determined by those with complete follow-up questionnaires absent of self-reports of smoking after a minimum of 10-years of previous smoking reported at baseline

for more precise estimation of the risk and associated effect half-life.

3.1.1.6 Null or Nearly Null Risk

To ensure that my model does not artificially induce an association when there is none, I considered a scenario where there is no true relationship between the exposure and outcome (HR1). The null EE curve is still generated with an imposed half-life of 90 days, even though this underlying exposure measure is not meaningful. This is because the probability of event due to the exposure is considered null, which implies that there should not really be a lag associated with this effect.

I hypothesize that my estimation procedures may have difficulty discriminating between exposed and unexposed individuals, as the likelihood surface may be flat. In the situation that either the OPEE or TPEE algorithms returns implausible lags and/or effect size estimates, I also add some fixed half-life models into the analyses to determine whether imposing any lag-of-effect biases the estimate of effect size.

I further extend my simulations to the 10% and 20% increased risk scenarios, or HR=1.1 and 1.2, respectively. The goal of this is to determine where the OPEE model "breaks", to provide insight on the capabilities of my algorithms in detecting nearly null associations that are lagged.

3.1.1.7 Variations on Half-Life

Shorter

Given that the parameter space for half-life is bounded by 0, I selected a half-life=1 day as a simulation variant (Half1) to look at null lag in the base case set of trajectories. Using a small half-life should result in similar effect size estimates as

a standard "current dose" model. Meanwhile, those exposed in the past (*down* group) should not have events attributable to the lagged exposure, because the effect would subside almost immediately. This can be seen by looking at the upper right corner of figure 3.2. The step-like function demonstrates how the *down* exposure group returns to the level of the controls instantaneously, and the *up* group may as well be classified as exposed from the start.

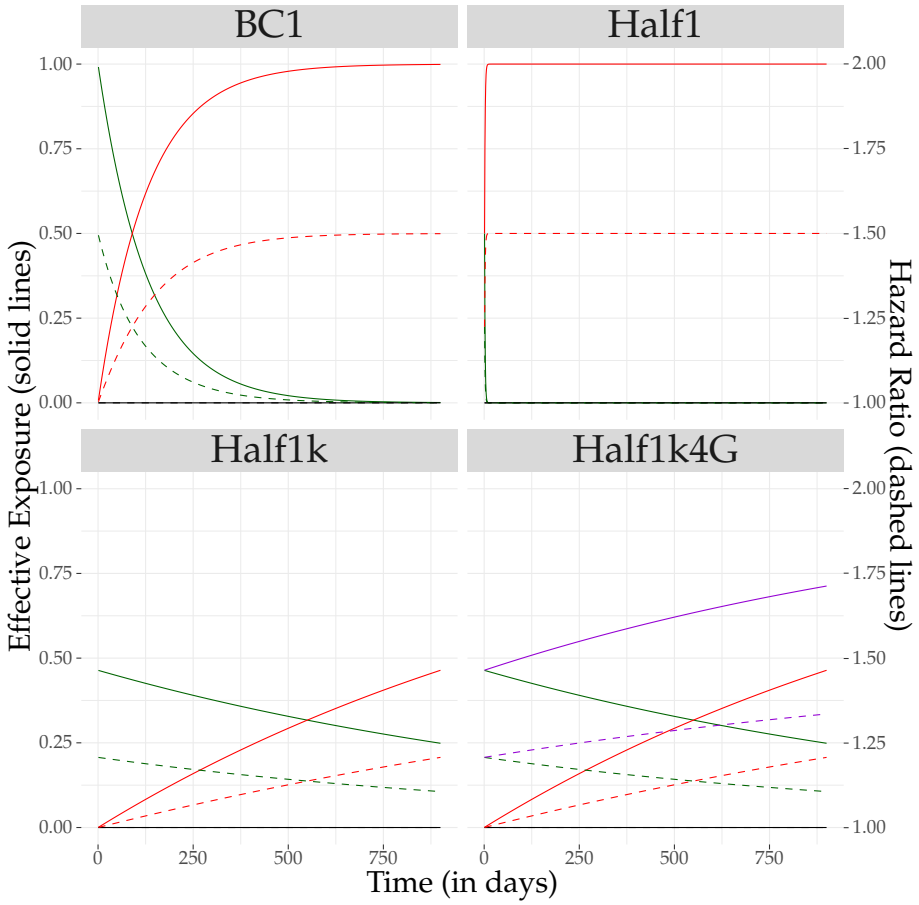


Figure 3.2: Plots for One-Parameter Half-Life Variations to the Base-case Simulation Scenario

The close proximity of this half-life to the likelihood's edge implies the normality of the likelihood may be violated, which should make estimation of the

asymptotic standard errors difficult. Due to this, I also consider half-lives of 10 and 30 days (Half10, Half10.4G and Half30.4G) to determine the breaking point of both estimation algorithms and associated standard error approximations. Note: The Half10 scenario represents the base case risk profiles with a 10-day half-life, while the Half10.4G represents the FourG scenario profiles with a single 10-day half-life parameter. Meanwhile, the scenario using a single half-life of 30 days is only considered for the four risk profiles extension.

Longer

In a situation where no one reaches steady state, I hypothesize that the true maximum hazard associated with the EE may not readily identifiable. My algorithms may be better at estimating the steady state hazard than the conventional exposure models, however, the minimal change in risk during the follow-up period (less transitioning information) may bias the estimates for half-life (or lag parameter) and the hazard ratio.

One particular branch off of the BC1 scenario that I consider is one with a half-life of effect of 1,000 days (Half1k). In this case, the individuals in my study never reach steady state, because the study period is shorter than a single half-life. As no one in the sample has reached the maximum hazard being estimated, the information, or lack thereof, may hinder my ability to estimate the lag parameter.

The bottom left panel of figure 3.2 shows the linear-like curves for the EE and HR over time. Keeping regimen timings (start and stop times of exposures) the same as in the BC1 scenario, the *down* group has only been exposed for 900 days prior to the study, meaning that these individuals have not even reached 50% increased hazard. Meanwhile, the individuals that start exposure at the beginning

of the study [*up* group] only cross the exposure level of the *down* group at roughly 550 days.

To account for the fact that the Half1k variation does not have anyone close to plateau throughout the study, I added the fourth *on* group in a new set of simulations (Half1k4G). The bottom right panel of the same figure 3.2 shows that the *on* group, which starts the study at the same EE level as the *down* group, continues to climb towards steady state throughout the follow-up period. This fourth group provides better information about the total obtainable risk associated with exposure and the resulting estimates should be less biased for both half-life and HR parameters than the 3-risk profiles variant (Half1k).

I also consider the single half-life of 450 days (Half450.4G) with four risk profiles (not pictured). This is to allow me to assess the algorithms' ability to detect a lag that is half of the follow-up time. That is, within the study period, individuals that are *up* or *down* should transition from null to 1.36 and from 1.36 to 1.08 times the hazard of the *ctrl* group, respectively. The fourth *on* group starts at the same hazard as the *down* individuals, but continues towards the 1.5-fold plateau throughout the study (ending at the hazard associated with an EE level after 4 half-lives of transition).

3.1.2 Two-Parameter Effective Exposures

All of the scenarios described so far have assumed the single half-life parameter model for EE. Naturally, the reader would like to see variations that address the two-parameter effective exposure (TPEE) models.

To simplify comparisons, I consider variants of the BC1 scenario, changing only the decline parameter and keeping the incline half-life at 90 days. This way, I can

look at a situation where the decline is either faster or slower than the half-life, with appropriate control and comparison. The data generation schemes and other parameters (sample size, HR, prevalence, risk profiles) are kept the same as the base case. The base case scenario in terms of the TPEE structure could be though of as having an incline=90 days and decline=90 days (i.e. [90,90]).

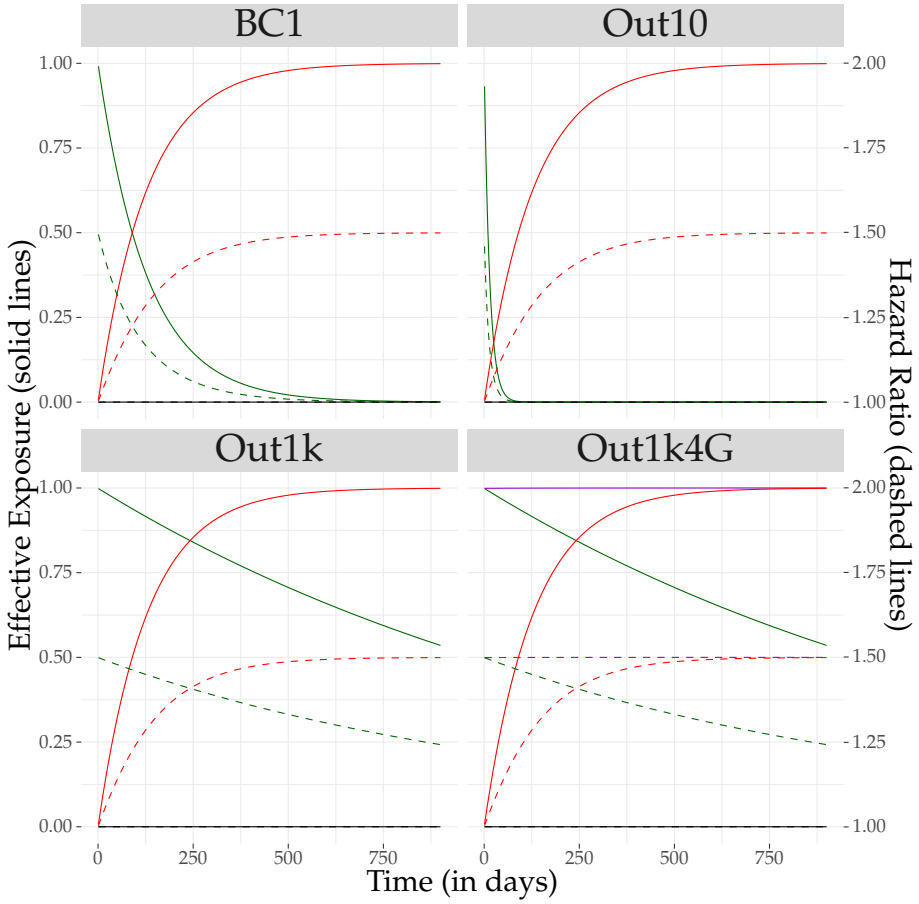


Figure 3.3: Plots for Two-Parameter Simulation Scenarios

Figure 3.3 demonstrates the two-parameter scenarios compared to BC1. The upper right corner shows a decline half-life of 10 days (Out10), while the two bottom panels reflect a decline half-life of 1,000 days. The bottom right (Out1k4G) differs from the bottom left (Out1k) in that a fourth *on* group is added to the simu-

lations [4].

Aside from assessing the TPEE algorithm's performance in these scenarios, I also am interested in the OPEE algorithm's ability to converge given the analytic and underlying models do not match. I hypothesize that the TPEE algorithm may have similar bounds in the estimation of short and long decline half-lives as with the short and long OPEE half-life models.

To parallel the OPEE half-lives considered, I also simulate four risk profile scenarios for the 10-day decline (Out10.4G), and declines of 30- and 450-days (Out30.4G and Out450.4G, respectively).

3.1.3 Multivariate Real Data

To, more closely, investigate the role of protracted exposures in a "real-world" setting, I utilize multivariate time-varying data from the BWHS cohort.[5] In summary, this dataset includes women with monotonic and complex trajectories of self-reported smoking patterns over time, along with time-varying information regarding potential confounders and risk factors of CVD. The goal is to use the *real* predictors and exposure profiles for smoking and covariates, to simulate the event of interest while controlling the half-life parameter and maximum hazard associated with prolonged smoking exposure.

Specifically, this dataset has real trajectories of dichotomous smoking exposure, and only includes right-censoring for incident cancer diagnoses, loss to follow-up, and death. Deaths are not further classified as CVD or non-CVD in this set, to allow for random assignment of events across all individuals and their available

[4]Similar to the base case FourG and Half1k4G scenarios goal is to improve identifiability and discrimination of the trajectories, thereby reducing bias in the estimation of the lag parameters.

[5]Outlined in more detail in Chapter 4.

follow-up times. This is different from the samples used in Chapter 4, because the true analytic samples censor cases of CVD following the event. Meanwhile, the set of true CVD cases are allowed to continue through the study without censoring, unless cancer, death, or loss to follow-up occur. This decision, to treat all participants the same, provides a dataset on which I can simulate CVD-like outcomes for all available follow-up trajectories of smoking's EE. Both scenarios described, below, consider the OPEE of binary smoking as the underlying CVD hazard model.

Figure 3.4 shows a small sampling of participant trajectories classified by study-wide (or lifetime in the study) exposure variation over time. The control group is not included in the figure, but makes up a substantial portion (60%) of the true cohort and scenario data.

3.1.3.1 Monotonic Trajectories

To disentangle the information-gain coming from covariate adjustment versus information gained by including more individuals in fluctuating [i.e. in states of transition], I first consider a subset of participants with unidirectional smoking profiles, similar to the "restricted" BWHS subset in Chapter 4. The scenario, here, is referred to as the Multivariate with Monotonic Trajectories set of simulations (MVmono).

The BWHS set is restricted to participants that either smoked throughout the study (Smokers Throughout), never smoked prior to and throughout the study (Nonsmokers Throughout), or smoked in the past and quit at some point during or prior to the study with no return to smoking (Successful Cessators). Both the smokers and quitters are further restricted to those with at least 10 years of smoking exposure at entry to the study (1995). The top two panels of figure 3.4 (Smokers

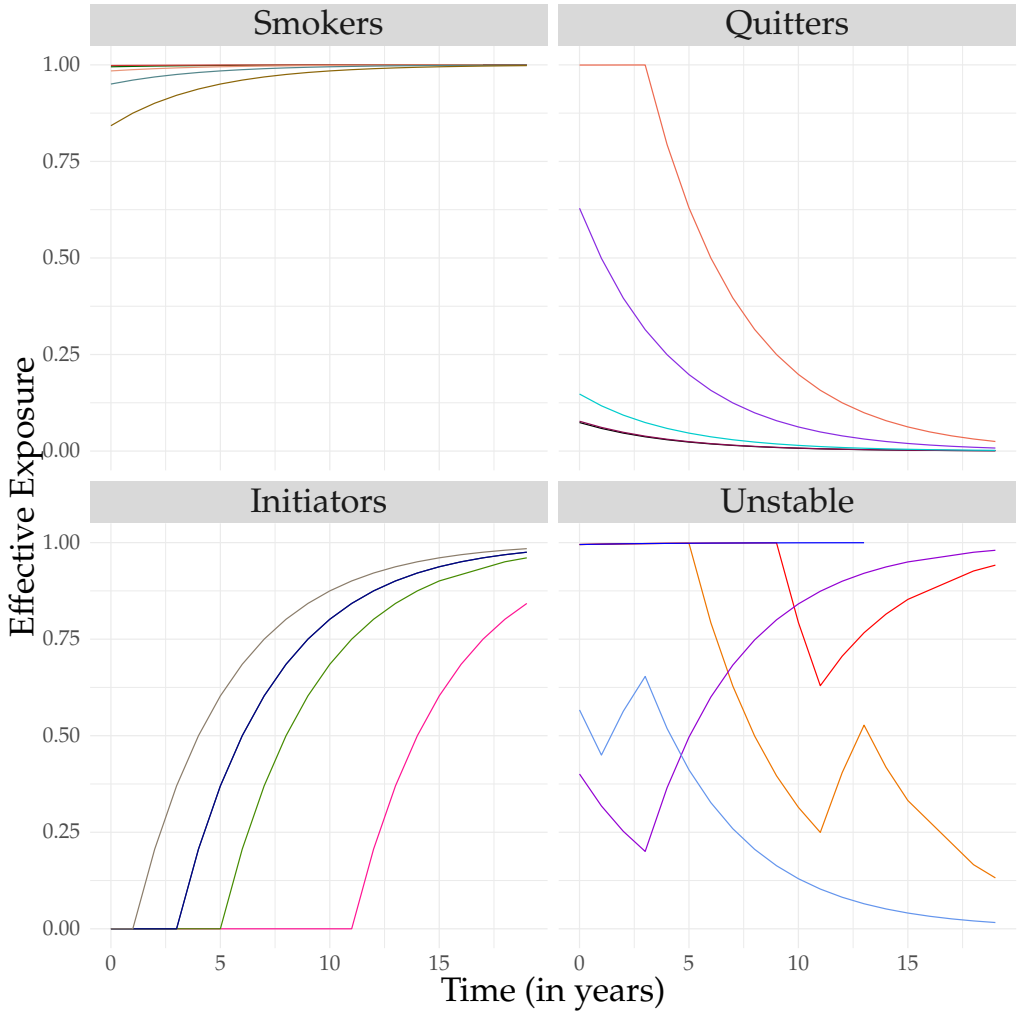


Figure 3.4: Example Trajectories from Black Women's Health Study Data Imposing a One-Parameter Half-Life of 3 years

and Quitters) show EE, or the annual probability of event, for several participants under the imposed OPEE half-life of 3 years and corresponding 3-fold hazard of CVD for extended exposure compared to never smokers.

As alluded to in the next paragraph, and as is consistent with the "NoUp" base case variant, I hypothesize that the lack of information about smoking initiation will make estimation more difficult when using the TPEE algorithm, particularly with regards to the incline parameter.

3.1.3.2 Multi-Trajectories

I call the final set of simulations the Multivariate Multi-trajectory scenario (MV-multi). The hypothesis underlying this scenario is that the OPEE and TPEE algorithms will have less bias and more coverage of the true lag and hazard ratio parameters compared to the MVmono scenario. Returning to figure 3.4, the two bottom panels represent the additional types of subjects and risk profiles that are now included in this simulation scenario.

3.2 SIMULATION METHODOLOGY

3.2.1 Univariate Scenarios

For each simulation I start by specifying inputs for exposure trajectories [6], [population] hazard ratio for steady state, and a population prevalence. The base case simulation scenario, its variants, and the two-parameter simulation studies all follow the same data generation schema where population-based exposure profiles are specified a priori. Specifically, the probability of event at time t is calculated for a given exposure profile on a daily basis. This allows me to control the granularity of the data at roughly 1/100th of a half-life (recalling the base case half-life is 90 days, so technically 1/90th). These probabilities are assumed to apply to all individuals with the same risk profile, and events are assigned using a logistic model of event at the daily discrete time points.

The discrepancy between the data generation model (odds) and the estimation model (hazard) is not of concern, because the odds approximates the hazard when the interval is short and the rate of events in the interval is sufficiently small.(Green

[6] i.e. half-life parameter(s), time since start of exposure, time since discontinuation, and dosing level

& Symons, 1983). Thus, when the overall prevalence of event is divided evenly across all study time-points, the daily prevalence is roughly 0.01% (recalling the 10% study-wide prevalence).

At each time point, the probability is compared to a random number drawn from the Uniform(0,1) distribution, with events set to "True" when p is less than the corresponding random number. This process is done per subject by iterating through the ordered sequence of time points until the first event occurs for an individual, or until the last time point (900 days). Subjects are then right-censored at event, but no other censoring situations or missing data are assumed to occur. All non-event times for each individual are retained until the full set is split into one observation per subject per unique event-time (study-wide event times).

Using the half-life parameterization of lag, the true parameters are set to reflect a 50% increase in hazard (or a HR of 1.5) for those at the steady state EE. The base case underlying EE curve assumes that new users will reach approximately half of the steady state's hazard after 90 days of constant exposure. Meanwhile, those who start at steady state and discontinue exposure, will fall to approximately half their hazard after 90 days. The two-parameter scenarios are slightly more complicated to interpret, so I keep all the inclines at 90 days implying that all the *down* subjects start at the same hazard (at time=0) as they would in BC1.

The total number of events and subjects, simulation seed used, and computation times are collected. Additionally, for each model fit, I retain the log-likelihood and AIC, the estimated HR, and an estimate for the lag parameter(s) with corresponding standard errors for all estimated parameters. This resulting data is combined to assess the bias and coverage probabilities of the new techniques.

3.2.2 Multivariate Scenarios

As described in the scenario specification, the multivariate scenarios are based on real data from the BWHS cohort. Thus, the covariance structures across time and potential confounders for each individual are preserved, assuming that this cohort provides an adequate random sample of the population from which it arose.

Prior to simulating the outcomes on the full set, I fit a pooled logistic regression (PLR) model on the true analytic sample. To adjust the estimates for each covariate's relationship with CVD appropriately, I used the categorical smoking specification (current vs. never and past vs. never) as the primary exposure in a multivariate model with additional adjustment for age [7]. The corresponding betas for the intercept, age, and pre-specified covariates are saved for use (described later). These estimates represent part of the log-linear model for the yearly probability of event.

Unlike the CPH model, the PLR model provides an estimate for the intercept and age-parameter. By adjusting for age, I am able to account for the time-to-event, while still being able to compute a predicted probability of event given a subject's covariate structure at each time point. To compute the predicted probability of event due to a known half-life of effect, I assume the OPEE model and set the true lag parameter to a 3-year half-life with a corresponding 3-fold odds of event for those at steady state risk for CVD due to smoking.

The following formula describes this simulated probability:

$$\text{odds}_{it} = -9.34 + 1.099E_{it}(h) + 0.04\text{Age}_{it} + 0.02\text{BMI}_{it} + 0.25\text{FamHxCVD}_{i}$$

[7] The models used in Chapter 4 are age-stratified Cox Proportional Hazards, thus, age is not considered a covariate in those models and coefficients for the baseline hazard by age (or age and intercept) are not available.

$$
\begin{aligned}
&- 0.07\text{Statins}_{it} + 0.16\text{HighChol}_{it} + 0.64\text{T2D}_{it} + 0.84\text{HTN}_{it} \\
&+ 0.16\text{Meno}_{it}^{\text{Pre}} + 0.33\text{Meno}_{it}^{\text{Post,Age@meno<45}} + 0.12\text{Meno}_{it}^{\text{Post,Age@meno 45-49}} \\
&- 0.25\text{Exercise}_{it}^{\text{<1hr/wk}} - 0.39\text{Exercise}_{it}^{\text{1+hr/wk}} \\
&- 0.09\text{Alcohol}_{it}^{\text{Current}} + 0.005\text{Alcohol}_{it}^{\text{Past}}
\end{aligned}
$$

$$
p_{it} = \frac{exp(\text{odds}_{it})}{1 + exp(\text{odds}_{it})}
$$

Similar to the univariate scenario's event-time specifications, each individual's probability of event is compared to a random univariate value, iteratively from the first time point to the last, or until an event is deemed to occur. In this case, however, not all subjects have the same available amount of follow-up time, which is reflected by the censoring structure already imposed on the data prior to simulations.

3.3 ANALYTIC VARIATIONS

The analytic model predominately used in the analyses is the CPH regression. As mentioned in Chapter 2, the "survival" package functions are used for fitting the time-dependent variable models. For the "Interval-Based" analyses, described later, I also use Pooled Logistic Regression via the "speedglm" package in R.

One major reason for doing these simulations has been to compare my method to conventional approaches used in epidemiologic research in the presence of lagged effects. Thus, I consider some conventional exposure metrics that are used in the application in chapter 4. Additionally, I want to understand how well the OPEE and TPEE algorithms estimate the correct lag and effect parameters in various scenarios.

Standard time-varying classifications of smoking exposure may include a categorical status (current vs. past vs. never), current indicator (current vs. not current), ever indicator (ever vs. never), and some combinations using a pack-years variable. The latter has been accepted as a standard for incorporating dosing (intensity) over time, however, the majority of my simulations are based on binary exposures, so no comparable metric was created within my study. Therefore, only the first three exposure measures are analyzed within each simulation scenario.

Of the three metric models, the current and ever models use a single parameter while the categorical model includes two parameters, one for current and on for past exposed individuals compared to the never exposed. The simple risk profiles would assign *up* and *on* individuals as currently exposed, and the *down* group as past exposed. All three risk profiles would contribute to the ever indicator.

These models are compared to the OPEE and TPEE by looking at Akaike's Information Criterion (AIC), which applies a penalty for an increasing number of estimated parameters. One and two parameters need to be added to the overall number of model parameters in the AIC calculation for the OPEE and TPEE models, respectively. My hypothesis is that the OPEE and TPEE methods will produce "better fit" results than the conventional exposure measures, as determined by minimizing the AIC.

I consider the Monte Carlo estimate of the percent bias, proposed by Koehler et al. (2009), as a way to compare the bias of the univariate and multivariate model estimates.

$$\hat{\varphi}_R^b = \frac{1}{R}\sum_{r=1}^{R}\frac{\hat{\theta}^r - \theta}{\theta} \times 100$$

To determine how well the variability approximations (i.e. standard errors) cover the true parameter of interest, I considered the coverage probability (CP)

metric. This can be understood as the proportion of simulations whose nominal 95% confidence intervals contain the true half-life or β parameter. Ideally, the CP should be around 95%. Koehler et al. (2009) presents a Monte Carlo estimate of CP as:

$$CP = \frac{1}{R}\sum_{r=1}^{R} I\left[\hat{\theta}_r - 1.96se(\hat{\theta}_r) \leq \theta \leq \hat{\theta}_r + 1.96se(\hat{\theta}_r)\right]$$

In both metric formulas, θ denotes true value for the particular parameter of interest, while R is the number of simulations. The $\hat{\theta}_r$ refers to the estimated parameter for the r simulation, and $se(\hat{\theta}_r)$ is the corresponding approximated standard error. All of my scenarios included 1,000 simulations, i.e. R=1,000. However, for the tables presented in the text of this chapter, I have chosen to report the "non-failed" CP and % bias (explained in section 3.3.1.2). This means that each scenario's R may vary in the tables, though the complete simulation results across all 1,000 runs can be found in appendix C.

3.3.1 Effective Exposure Algorithms

All simulation scenarios are analyzed using both the OPEE and TPEE algorithms, regardless of the underlying data generation model. The goal is to compare the performance of each algorithm's estimation under varying truths. I hypothesize that the OPEE approach will fail to identify the correct lag under scenarios with two lag parameters, and that the TPEE algorithm will converge to similar estimates for each lag parameter in an OPEE-simulated dataset.

For the purposes of assessing coverage in the new methods, the standard errors used for the hazard in each model are re-calculated using the Information matrix-derived variance. In this situation, the resulting standard errors are typically wider than those from the original model fitting process, but this allows the variability to

be adequately adjusted for the additional parameter(s).

As the algorithms require an input for the initial half-life "guess", all univariate scenario algorithms are initialized with a starting half-life of 60 days, regardless of the true half-life. When an algorithm fails (see below), the simulated data is subjected to a second half-life search starting at a 100-day initialization. If the second-initialized-algorithm fails to converge, no additional algorithms are applied and the original initialized half-life results are kept for analysis.

3.3.1.1 Initialization

To account for potential bias from invalid initialization, I performed a small exploratory analysis of different initialization half-lives using the BC1 scenario simulations. These initial values considered are the half-life at 60-, 70-, 80-, 90-, 100- and 110-days. As the first set of models compared during the initialization step includes the marginal likelihoods for half and twice the initial value, testing a wider range of input values becomes irrelevant. This is because the first step expands using a similar mechanism of halving (or multiplying by 2) to get into the approximate range of the maximum likelihood. Thus, any variation in the results due to initialization should depend on the location of the center when step 2 starts, i.e. for a maximum occurring at 70 days (for a given simulated dataset), initializing the half-life at 80 days may be problematic if the likelihood is not symmetric. I hypothesize that a minor shift of the initial guess can move the profile likelihood points for the algorithm around a problem ridge area, such that other initializations produce consistent estimates for the half-life and HR parameters.

3.3.1.2 Algorithmic Failure

When I refer to "failed" and "non-failed" simulations, I am specifically focusing on the OPEE and TPEE algorithms that struggle to converge during the estimation process. This is based on non-estimate-able values for the half-life standard error and/or infinite beta parameter estimates. In some situations this occurs due to a flattening of the likelihood surface or failure of normality near a bound, while other failures could be considered to come from the random variations of the simulated data.

To make the results comparable, I have chosen to display only the "non-failed" simulation results in the tables here. Therefore, the corresponding "# Fail" column is meant to orient the reader with regards to the total number of failures for a particular parameter estimate out of the 1,000 simulations performed in that scenario.

3.3.2 Interval-Based Analyses

To account for situations in which the specific event time is unknown and assumptions are made about timing of exposure, I consider the impact of interval slicing on the estimation of the lag and risk parameters. Unlike the original analytic methods for the simulations, the interval-based analyses are expected to shed some light on the effect of repeated measures data assumptions. What I mean is that a typical cohort dataset, like the BWHS, will collect information at pre-specified intervals, which are assumed to be the start and stop points for both exposures and outcomes. For example, in Chapter 4, a limitation may be the fact that each outcome is only reflected in the year of the event, despite variability in the timing of events within that year. Therefore, the resulting estimate for the half-life is expected to be biased, since an individual with the event in March, who may have been exposed

for 3 months, is calculated as exposed for 12 months, since the outcome is set to the year-end time.

Using the base case scenario's simulated samples, I split each dataset into secondary sets with varying interval lengths (10, 50, 100, 300, 900). The interval sets are then analyzed using both PLR and CPH. By comparing analytic model estimates of the hazard ratio[8] and lag parameters for the various intervals, I may be able to make recommendations about the granularity of data needed for optimal performance of my methods.

There are two dataset structures that I consider – structure (1) which assumes that information is collected at equally spaced intervals and that all events occur at the end of the interval; structure (2) which assumes that information is collected at equally spaced intervals, yet the cases' event times within the interval are known.

In Structure 1 the Cox Proportional Hazards models (*cph1*) assigns all events to the same time in the interval, thereby the EE estimate for cases in an interval would be biased. For example, subject A is coming *down* from exposure, and has an event at 28 days. The true EE for A's event time is 87%, but if A is assigned as an event at the end of a 100-day interval then the corresponding EE will be 68%. Meanwhile, the non-case B who follows the same trajectory as A will be compared using these same EE levels, which reflect the interval-end EE. Similarly, the Pooled Logistic Regression (*plr1*) will have true EE assigned to non-cases in the interval, while the *up* group cases may be overestimated and *down* group cases underestimated at the true time of the event in the interval. The results for *cph1* vs *plr1* should not differ dramatically, as the time components in the CPH risk sets are identical to those in each interval pool. As the interval length increases, the CPH and PLR ratio

[8] recalling that the odds ratio is an approximation of the hazard ratio here

estimates should diverge [9].

In Structure 2 the Cox model (*cph2*) risk sets are based on event timings, leading to correct EE assignment for cases, but biased EEs for non-cases in the risk set. For example, when subject A and B are both coming *down* from exposure, A's EE at event time 28 is 87%, but a 100-day interval means that B's EE at time 100 of 68% is used for non-case comparison. Realistically, the two individuals should be compared using the same EE at day 28. Therefore, a 10-day interval (B at 30 days EE=86%) should result in smaller misclassification bias than the 100-day intervals. The problem persists using the pooled logistic regression model (*plr2*), but both of structure 2's analytic models should out-perform structure 1's models in identifying the correct half-life parameter.

3.4 RESULTS

3.4.1 Conventional Exposure Metrics

The conventional measures model tables (tables 3.1, and C.2) have several shorthand notations worth explaining. PvsN and CvsN come from the time-varying categorical three-level model and represent the estimated HR for the Past or Current group compared to the Never group. Current refers to the model where those in the *on* or *up* groups are considered to be at risk, and Ever refers to the model where those in all groups except *ctrl* are at risk – i.e. the *down* group is treated as being at the same hazard as *on* and *up*. The estimates presented for the conventional measures in the interval analyses in the Appendix (table C.10) reflect the hazard ratio (HR) in the CPH and odds ratio (OR) in the PLR models.

[9] based on the same concept that odds and risk can approximate one another as long as the intervals remain short enough

Table 3.1: Hazard Ratio Estimates for Conventional Metrics Across All Simulations Of Selected Scenarios

Simulation	Input	CvsN		Current		Ever	
Scenario	HR	Mean	% Bias	Mean	% Bias	Mean	% Bias
BC1	1.5	1.42	-5.48	1.37	-8.58	1.24	-17.21
N1k	1.5	1.46	-2.48	1.40	-6.94	1.28	-14.79
N10k	1.5	1.42	-5.18	1.37	-8.36	1.24	-17.02
N100k	1.5	1.42	-5.56	1.37	-8.67	1.24	-17.23
FourG	1.5	1.46	-2.75	1.41	-5.94	1.33	-11.53
NoUp	1.5	1.50	0.16	1.45	-3.17	1.28	-14.42
DoseMods	1.5	1.71	14.28	1.38	-7.99	1.41	-6.08
Half10	1.5	1.49	-0.52	1.49	-0.91	1.25	-16.85
Half10.4G	1.5	1.50	-0.26	1.49	-0.65	1.33	-11.28
Half1k	1.5	1.11	-25.99	1.03	-31.21	1.13	-24.58
Half1k4G	1.5	1.19	-20.47	1.11	-26.08	1.18	-21.37
MVmono	3	2.96	-1.47	2.86	-4.82	1.51	-49.69
MVmulti	3	2.75	-8.18	2.59	-13.52	1.63	-45.82
Out10	1.5	1.42	-5.46	1.41	-5.82	1.21	-19.25
Out10.4G	1.5	1.46	-2.73	1.45	-3.11	1.31	-12.90
Out1k	1.5	1.42	-5.48	1.20	-19.71	1.39	-7.47
Out1k4G	1.5	1.46	-2.75	1.24	-17.40	1.43	-4.99

Input HR: The hazard ratio used in data simulation for the effective exposure at steady-state risk
CvsN: Estimated HR for Current Exposure compared to Never exposed in categorical model
Current: Estimate HR for Current vs. Not Current exposed risk ratio
Ever: Estimate HR for Ever vs. Never exposed risk ratio

Table 3.1 shows a selection of the simulation scenarios performed and the remaining scenarios considered can be found in the Appendix table C.2. In all simulation scenarios and models of the categorical (current vs. past vs. never) exposure, the past users effect size estimate was lower than the simulated hazard ratio (data not shown).

As expected, all the conventional metrics in the multivariate simulations showed an underestimation of the true association measure. For the multivariate scenarios, the conventional metrics for current exposure were the least biased in both categorical and dichotomous models, with slightly more distance from the truth (CvsN % Bias -8.2 vs. -1.5, Current % Bias -13.5 vs. -4.8) in MVmulti than MVmono. This was not the case for the "Ever" models, likely because more individuals were con-

tributing to the "past" or "ever" exposure categories in the MVmulti than MVmono scenarios.

Adding a fourth group to the scenario (FourG vs. BC1, Half10.4G vs. Half10, Half1k4G vs. Half1k, Out10.4G vs. Out10, Out1k4G vs. Out1k) reduced the bias for all conventional measures. This is consistent with the information hypothesis that having individuals who are *on* or exposed and close to steady state improves the estimation of the hazard level. I discuss the variations observed in the OPEE and TPEE algorithm performance between the 3- and 4-risk trajectory simulations later (section 3.4.6).

3.4.2 Sample Size Variations

The conventional measure results for N10k and N100k sample sizes were identical to the base case sample. It is interesting to note that all of these metrics were the least biased for the smallest sample size scenario (N1k). Meanwhile, decreasing sample size reduces coverage, increases bias, and is more likely to fail in convergence, specifically for the half-life parameter.

Table 3.2 shows the OPEE performance of the HR and single-parameter half-life estimates. The base case scenario corresponds to the 30,000 subject sample size rows.

In this table, the "# Fail" column reflects convergence issues specific to the parameter – for the hazard ratio, failure is denoted by an infinite estimate or negative variance component. Meanwhile, the half-life estimate's failure comes from a negative approximated variance component. The latter two points are the same and discussed in more detail later. As mentioned previously, the number in this column reflect the failures *out of* the 1,000 total simulations performed for that particular

Table 3.2: OPEE Performance by Sample Size

Parameter	Sample Size	Mean	Median	Coverage Probability	% Bias	# Fail
Hazard Ratio	1,000	2.4e+66	1.66	95.3	1.6e+68	0
	10,000	1.52	1.52	95.2	1.63	1
	30,000	1.51	1.51	95.9	0.47	1
	100,000	1.50	1.50	94	0.07	0
Half-Life	1,000	777.63	90.00	85.4	764.03	5
	10,000	95.59	90.00	90.6	6.21	1
	30,000	92.22	90.00	92.9	2.47	1
	100,000	91.30	90.00	94.9	1.44	0

scenario.

While samples size seems to affect the estimation performance with respect to the half-life parameter, the coverage of the hazard ratio by the normally approximated 95% confidence intervals was consistently favorable regardless of the number of subjects in the simulation sample. This implies that my method can still deliver appropriate estimates of the hazard ratio, even when the sample size may be too small to estimate the half-life of the effective exposure. Increasing the sample size may improve performance, however, the 100,000 sample size required 3-fold more computational time to perform the OPEE and TPEE algorithms compared to the base case scenario of 30,000 individuals.

Using the TPEE algorithm in the N1k and N10k scenarios led to large overestimation of the HR, while the base case sample size also saw slight overestimation of hazard from TPEE (in Appendix C table C.6). The TPEE analyses on the N100k scenario only failed for 1 simulation, compared to 74 in the base case sample size, and overall had minimally biased estimates of the hazard ratio.

3.4.3 One-Parameter Half-Life Variations

In table 3.3 the half-life=1 day scenario appears to have consistently low bias in the hazard ratio estimation with a 95% coverage probability in the 95% CI bounds. Additionally, this short half-life had the best coverage probability for OPEE half-life across all scenarios evaluated by the OPEE model.(Table C.7) Given that good estimation of the half-life drives less biased estimation of the hazard ratio, these results are not surprising. This is further confirmed by the large bias seen for both large half-life (1,000 days, single half-life) scenarios, regardless of the number of trajectory groups (table 3.8). The large half-life models tended to fail more often than for other scenarios, in both OPEE and TPEE algorithms.

Table 3.3: Simulation Results by One-Parameter Half-Life in Three Profile Simulations

	True	OPEE				TPEE			
Parameter	Half-Life	Mean	CP	% Bias	# Fail	Mean	CP	% Bias	# Fail
Hazard Ratio	1 day	1.51	95.2	0.35	0	1.52	92.2	1.59	129
	10 days	1.51	94.7	0.34	0	1.52	94.1	1.36	49
	90 days	1.51	95.9	0.47	1	1.60	97.4	6.44	22
	1,000 days	1.5e+87	86.8	9.8e+88	60	3.1e+203	60.3	2.1e+205	456
Incline Half-Life	1 day	3.77	98.4	277.08	4	6.56	100	555.84	139
	10 days	11.85	86.1	18.53	2	15.33	85.4	53.27	77
	90 days	92.22	92.9	2.47	1	123.38	90.5	37.09	42
	1,000 days	4006.82	80.4	300.68	57	26270.03	58.7	2527	456
Decline Half-Life	1 day					24.07	92.7	2306.7	767
	10 days					28.02	94.5	180.2	561
	90 days					98.43	88.9	9.4	71
	1,000 days					2752.38	93.5	175.2	507

CP: Coverage Probability

Table 3.4 demonstrates the single-parameter half-life variations for the four risk-trajectories. The sweet spot for OPEE performance appears at the 90 days half-life, where coverage of the hazard ratio and single lag parameter is the maximized and percent bias minimized. The results for these same scenarios using the TPEE algorithm (table C.6) also show the least % bias for all three parameter estimates

in the 90 days simulation scenario, though coverage is slightly higher and failures occur less often for a half-life of 450 days. This is understandable given that the *down* group spends more time in transition under the 450-day half-life (than in the 90-day half-life), thus providing more information for the decline parameter.

Table 3.4: OPEE Performance by One-Parameter Half-Life in Four Profile Simulations

Parameter	True Half-Life	Mean	Median	Coverage Probability	% Bias	# Fail
Hazard Ratio	10 days	1.50	1.50	94.8	0.2	0
	30 days	1.50	1.50	95	0.2	0
	90 days	1.50	1.50	95.7	0.2	0
	450 days	1.53	1.51	95.4	1.9	0
	1,000 days	5.4e+20	1.50	88.2	3.6e+22	95
Half-Life	10 days	11.83	9.38	85.5	18.3	2
	30 days	31.30	30	90.7	4.3	0
	90 days	91.89	90	92.3	2.1	0
	450 days	491.93	450	92.2	9.3	0
	1,000 days	2838.39	900	86.1	183.8	95

3.4.4 Two Half-Life Parameters Variations

The OPEE algorithm applied to an underlying TPEE scenario with four risk trajectories showed better coverage of the true Decline than Incline half-life when the decline was shorter than the incline (Table 3.5). The 1,000 day decline scenario had equally poor coverage of both parameters in the OPEE context (7-8%).

Generally, none of the TPEE scenarios failed the OPEE algorithm's convergence or normal approximation of standard errors, and all of the four group estimated hazard ratios had small bias. The OPEE algorithm tended to underestimate the longer half-life parameter and overestimate the shorter, as would be expected.

All longer decline simulation scenarios (Out450.4G, Out1k, and Out1k4G) correctly selected the minimum AIC model as the TPEE over 90% of the time vs.

Table 3.5: OPEE Performance by Two-Parameter Effective Exposure Four Group Simulations

Parameter	True Decline	Mean	Median	Coverage Probability	% Bias	# Fail
Hazard Ratio	10 days	1.49	1.49	93.5	-0.48	0
	30 days	1.50	1.50	95	-0.05	0
	90 days	1.50	1.50	95.7	0.20	0
	450 days	1.48	1.48	93.5	-1.28	0
	1,000 days	1.52	1.50	95.8	1.02	0
Incline and Decline	10 days	45.02	41.25	39.7 59.7	-50 350.2	0
	30 days	58.79	56.25	57.9 77.5	-34.7 96	0
	90 days	91.89	90.00	92.3	2.1	0
	450 days	226.86	225.00	24.2 10.7	152.1 -49.6	0
	1,000 days	387.94	360.00	7.8 7.2	331.1 -61.2	0

The two values in the Coverage Probability and % Bias columns reflect the measure's performance in relation to the incline and decline true half-life parameters. The left-hand value for each denotes the incline. The base case (90,90) scenario has only one unique value half-life, thus only one value is reported in the table.

OPEE, Categorical, Current vs. Not-Current, and Ever vs. Never (data can be found in appendix table C.4). The two-parameter scenarios with longer decline (Half=(90,1000), Out1k and Out1k4G) had mean estimates bias for the past category (PvN) in the categorical and ever models, closer to the true HR=1.5 than in the base case scenario variations and the two-parameter scenarios with the shorter decline lags (table 3.1).

Table 3.6 demonstrates the TPEE performance for the true underlying TPEE simulations in the scenarios with four risk trajectories. As the decline parameter increases, the rate of failure in estimating any parameter decreases. Similar to the single half-life simulation 4-group scenarios, coverage and bias of the hazard ratio were good (close to 95% and 0%, respectively) regardless of the underlying TPEE model. Failure was more common for the decline parameter than for the incline parameter.

Table 3.6: TPEE Performance by Two-Parameter Effective Exposure Four Group Simulations

Parameter	True Decline	Mean	Median	Coverage Probability	% Bias	# Fail
Hazard Ratio	10 days	1.51	1.51	93.9	0.66	15
	30 days	1.51	1.51	95.3	0.72	7
	90 days	1.51	1.51	96	0.62	4
	450 days	1.51	1.51	95.5	0.39	2
	1,000 days	1.50	1.50	95.2	0.32	0
Incline Half-Life	10 days	88.93	86.25	88.4	-1.19	21
	30 days	90.38	86.25	90.4	0.43	15
	90 days	95.05	90	91.7	5.61	4
	450 days	95.60	90	92.4	6.22	2
	1,000 days	95.00	90	92.3	5.56	0
Decline Half-Life	10 days	29.02	19.69	94.8	190.21	460
	30 days	43.70	36.56	93.6	45.68	216
	90 days	97.96	90	91.9	8.85	28
	450 days	482.72	480	94	7.27	2
	1,000 days	1221.44	990	91.3	22.14	0

Interestingly, the smaller decline parameter scenarios tended to have better estimation of the incline parameter based on the smaller bias, though the incline's coverage could be considered lower (I might call it "ball park" across all decline variations, CP=88.4 for Out10.4G vs. CP=91.3 for FourG).

3.4.5 Hazard Ratio Variations

The hazard ratio estimates for conventional measures in the null scenario, shown in Table 3.7, all produced an average null HR estimate for the effect of the exposure. The 10% and 20% increased-risk scenarios also had minimally-biased[10] results for the current exposure compared to non- or never-exposed individuals risk, in the dichotomous and categorical exposure models, respectively.

It stands to note that the HR1 and HR10p scenarios were unable to estimate

[10] towards-the-null

Table 3.7: Conventional vs. Fixed Half-Life Models in Null and Nearly Null HR Scenarios

		Categorical		Current	Ever vs.	Fixed Half-Life (days)			
Scenario	Statistic	PvsN	CvsN	vs. Not	Never	1	10	100	1,000
	Mean	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
HR=1	% Bias	0.072	0.071	0.046	0.069	0.044	0.034	0.034	0.498
	minAIC	586		0	0	17	62	161	174
	Mean	1.02	1.09	1.08	1.05	1.08	1.08	1.10	1.13
HR=1.1	% Bias	-7.7	-1.3	-2.05	-4.5	-2.01	-1.68	0.18	2.8
	minAIC	393		0	0	13	93	417	84
	Mean	1.03	1.17	1.15	1.10	1.15	1.16	1.20	1.26
HR=1.2	% Bias	-14.23	-2.52	-3.91	-8.39	-3.85	-3.22	0.32	5.14
	minAIC	200		0	0	9	105	674	12

HR: Hazard Ratio
PvsN: Estimate for Past Exposure compared to Never exposed in categorical model
CvsN: Estimate for Current Exposure compared to Never exposed in categorical model
minAIC: Number of times the model was selected by a minimum Akaike Information Criterion (AIC) out of 1,000 simulations

the true HR in both EE algorithms (Table C.3), while the HR20p had low bias and good coverage in the OPEE algorithm only. The median across OPEE and TPEE-estimated hazard ratios of the 1,000 simulations lands on the true parameter, but the mean hazard ratio estimates tend to explode for a true null or 1.1-fold hazard. The 1.2 HR appears to be a threshold for the algorithm's function of estimating the magnitude of association.

The over-estimation of the hazard, across the null-scenario simulations, is not an indicator for direction of the bias when using the OPEE or TPEE algorithms. This estimation-issue comes from the bias imposed during selection of the incorrect half-life parameter(s) by the algorithm. Thus, I consider what would happen if I imposed a half-life on my own, without the algorithm's application.

Considering fixed half-lives of 1, 10, 100, and 1,000 days in the null risk (HR1) scenario, the corresponding β parameter estimate was 0.0 on average (i.e. HR=1), across the 1,000 simulations. For this particular scenario (HR1), the categorical ex-

posure analytic model produced the minimum AIC in 586 of the 1,000 simulations, with the remaining "best fit" models selected equally across the fixed half-lives considered. In fact, neither of the standard dichotomous metrics ("Current" and "Ever") produced the minimum AIC statistic in any of the scenarios considered (Appendix Table C.4).

With increasing magnitude of risk, the minimum AIC-producing model – across conventional metrics and fixed half-lives – settles on the fixed half-life closest to the true lag parameter. Looking at table 3.7, the fixed half-life estimated HRs are the same, while the fixed half-life of 100 days is selected over the conventional models and other fixed half-lives 417 and 674 times out of 1,000 simulations, for the HR10p and HR20p scenarios, respectively. The estimated hazard ratios using the fixed (bounded) half-lives range from 1.08 to 1.13 and 1.15 to 1.26 for the true hazard ratios of 1.1 and 1.2. Comparatively, the conventional model hazard ratios all underestimate the relative effect of exposure.

3.4.6 Three vs. Four Risk Groups

Figure 3.5 shows a histogram of the true hazard ratio across the 3- and 4-group scenarios for a single half-life of 1,000 days. I see that the 4-profiles scenario has a narrower curve centered around 1.5. This implies that estimation of the true hazard ratio should be less biased in the 4-group runs than the 3-group runs.

Looking at table 3.8, it appears that adding a fourth group to the single-parameter half-life scenarios does not appreciably change the results using the OPEE algorithm, unlike the information hypothesis presented earlier. In the previous results (section 3.4.3) focused on the half-life parameter only, the largest single half-life of 1,000 days, produced infinite estimates of the hazard ratio. This is likely a result of

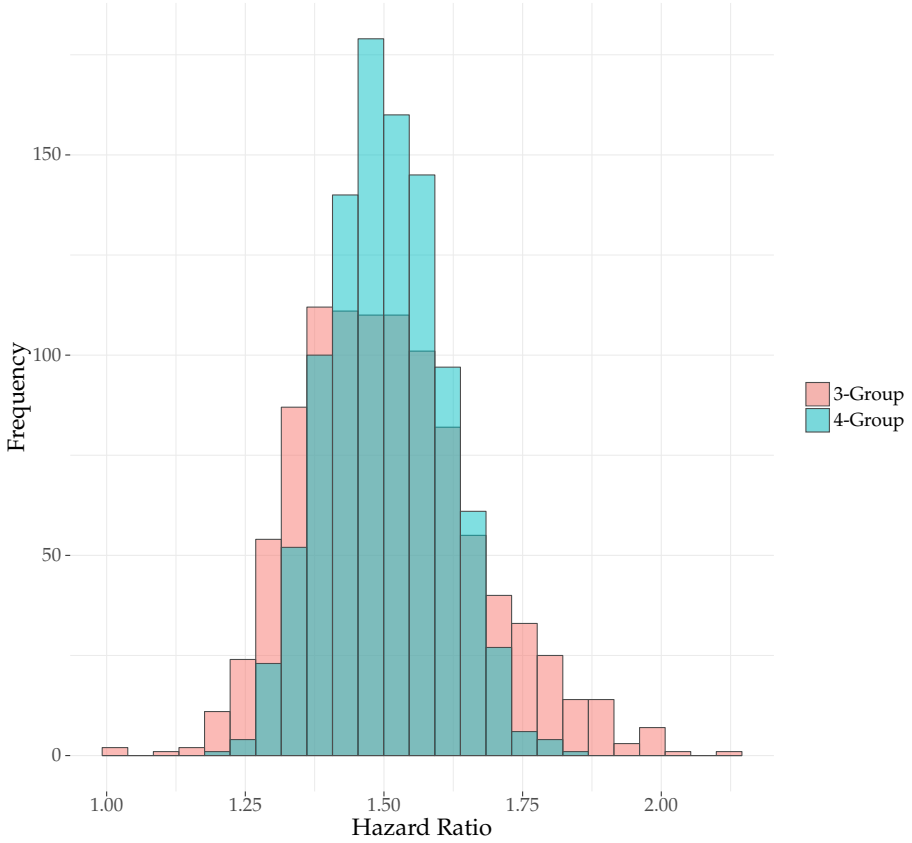


Figure 3.5: Histograms of steady state risk for half-life=1,000 days simulations, comparing the 3-group to 4-group designs.

the overestimation in the half-life parameter (failing 60 and 95 times for the 1,000 day 3 and 4 group scenarios, respectively). For this reason, the numbers presented in the table reflect the estimates of performance for the non-failed simulations, to circumvent the "Inf" cell-values. Looking at the non-failed scenarios only, the % bias is 301% and 184% for the half-life parameters in the 1,000 day 3- and 4- group scenarios, respectively. While removing the failed simulations does give estimates for the mean and % bias – i.e. non-infinite values – the estimated biases for the 1,000-day scenario HR parameters remain largely overestimated.

For the underlying TPEE scenarios, adding a fourth group generally tends to

Table 3.8: OPEE Performance by One-Parameter Half-Life Comparing 3- and 4- group simulations

Parameter	True Half-Life	# Risk Groups	Mean	Median	Coverage Probability	% Bias	# Fail
Hazard Ratio	10 days	3	1.51	1.51	94.7	0.34	0
		4	1.50	1.50	94.8	0.20	0
	90 days	3	1.51	1.51	95.9	0.47	1
		4	1.50	1.50	95.7	0.20	0
	1,000 days	3	1.5e+87	1.48	86.8	9.8e+88	60
		4	5.4e+20	1.50	88.2	3.6e+22	95
Half-Life	10 days	3	11.85	9.38	86.1	18.53	2
		4	11.83	9.38	85.5	18.28	2
	90 days	3	92.22	90	92.9	2.47	1
		4	91.89	90	92.3	2.10	0
	1,000 days	3	4006.8	840	80.4	300.7	57
		4	2838.4	900	86.1	183.8	95

reduce the likelihood of failures and bias of all parameters estimated. The decline of 1,000 days scenario coverage of the true hazard ratio was less than acceptable in the OPEE framework, but adding the fourth *on* group appreciably changed this (42% vs. 96%, appendix table C.7). Adding the fourth group minimally improved coverage for the decline parameter in the TPEE fits (Table 3.9), while all the 3-risk profile scenarios overestimated the incline parameter more than their 4-profiles counterparts.

3.4.7 Multivariate Scenarios

On average, the percent bias of the single half-life parameter estimates, across the 1,000 simulations, was 2.5% for the base case (BC1) scenario using the OPEE algorithm, while the MVmono and MVmulti estimates of percent bias were 1.19% and 0.26% for the same analytic approach, respectively. [11] Moving to the TPEE algorithm, the incline and decline percent biases were 37.1% and 9.4% in the base case.

[11] Recall that the MVmono and MVmulti models include adjustment for covariates

Table 3.9: TPEE Performance by Two-Parameter Half-Lives Comparing 3- and 4- group simulations

Parameter	True Decline	# Risk Groups	Mean	Median	Coverage Probability	% Bias	# Fail
Hazard Ratio	10 days	3	1.55	1.51	95.8	3.12	28
		4	1.51	1.51	93.9	0.66	15
	90 days	3	1.60	1.52	97.4	6.44	22
		4	1.51	1.51	96	0.62	4
	1,000 days	3	1.52	1.51	96.7	1.1	1
		4	1.50	1.50	95.2	0.3	0
Incline Half-Life	10 days	3	100.46	75	81.5	11.6	46
		4	88.93	86.25	89.3	-1.2	21
	90 days	3	123.38	90	90.5	37.1	42
		4	95.05	90	91.7	5.6	4
	1,000 days	3	101.71	90	91.7	13.0	1
		4	95	90	92.3	5.6	0
Decline Half-Life	10 days	3	37.09	23.44	94.3	270.9	458
		4	29.02	19.69	94.8	190.2	460
	90 days	3	98.43	90	88.9	9.4	71
		4	97.96	90	91.9	8.9	28
	1,000 days	3	1234.99	960	90	23.5	2
		4	1221.44	990	91.3	22.1	0

The percent bias, upon addition of covariates (MVmono), was 13% for the incline and 9.4% for the decline half-life parameters using the TPEE algorithm. The multivariate multi-trajectory (MVmulti) model simulations' percent bias stayed smallest (incline = 7.8% and decline = 6.4%).

The monotonic trajectories (MVmono) simulations had nearly the same performance across OPEE and TPEE algorithms as the multi-trajectory (MVmulti) sample sims (Table 3.10). However, the multi-trajectory set of simulations was better equipped (fewer failures, smaller bias) to inform the TPEE algorithms.

This was expected, since the monotonic trajectories sample had no individuals transitioning upwards[12]. By adding more information from the multi-trajectory in-

[12]Those smokers still climbing in risk would be expected to have almost reached the maximum

dividuals, coverage of the OPEE half-life parameter estimate improved minimally (93.8% in MVmono vs. 94.9% in MVmulti). The TPEE algorithm performance was not as similar between these scenarios, with better coverage of Incline and Decline parameters in the MVmulti simulations (Table 3.10), though the coverage of the hazard ratio estimates remained nearly identical.

Table 3.10: OPEE and TPEE Performance by Trajectory Variations

		OPEE				TPEE			
Scenario	Parameter	Mean	CP	% Bias	# Fail	Mean	CP	% Bias	# Fail
Base Case	HR	1.51	95.9	0.47	1	1.60	97.4	6.4	22
	Incline	92.22	92.9	2.5	1	123.38	90.5	37.1	42
	Decline					98.43	88.9	9.4	71
Four Groups	HR	1.50	95.7	0.2	0	1.51	96	0.62	4
	Incline	91.89	92.3	2.1	0	95.1	91.7	5.6	4
	Decline					97.96	91.9	8.9	28
No Up	HR	1.51	95.9	0.57	1	2.5e+130	99.1	1.7e+132	343
	Incline	97.81	91.6	8.7	1	2518.6	100	2698.5	386
	Decline					108.9	93	21	386
MV Mono	HR	2.98	94	-0.83	0	2.98	94.4	-0.80	9
	Incline	3.04	93.8	1.19	1	3.40	91.6	13.2	59
	Decline					3.28	82.6	9.4	60
MV Multi	HR	2.97	92.8	-0.90	0	2.97	93.6	-0.86	5
	Incline	3.01	94.9	0.26	0	3.24	92.6	7.8	20
	Decline					3.19	86.7	6.4	20

HR: Hazard Ratio; CP: Coverage Probability;
NoUp: Three trajectories of risk in base case scenario with "down", "ctrl", and "on" groups only
MV Mono: Monotonic trajectories from "Restricted" sample of BWHS participants used to simulate a 3-fold maximum hazard with a one parameter 3-year half-life.
MV Multi: Full sample of BWHS participants used to simulate a 3-fold maximum hazard with a one parameter 3-year half-life.

3.4.8 No Incline Variation

Similar to the MVmono, the NoUp scenario has no upwards-transitioning risk individuals. Compared to the base case and four group scenarios, the OPEE model in this scenario estimated a larger half-life, on average (97.8 vs. 92.2 vs. 91.9 days,

hazard plateau as inclusion in the sample at least 10-years of prior exposure at baseline

Table 3.10). The TPEE algorithm, here, continues to grossly overestimate the incline half-life even when restricting to non-failed simulations (Incline=2519 days). This is not as much of an issue for the decline parameter, though most three-group OPEE models and scenarios appeared to also have larger bias in the TPEE Incline than the Decline half-life estimates.

3.4.9 Dosing Variation

Referring to table 3.11, it is clear that both binary and dose-based OPEE and TPEE model algorithms lead to similar bias and slightly lower coverage probabilities in estimation of the lag parameter. However, the major pitfall of the binary-dose assumption for this scenario is the overestimation of the true hazard ratio with 0% coverage. This failure to estimate the true effect comes from improperly-weighting the risks associated with different levels of exposure, and is not a unique limitation to my methodology.(Copeland et al., 1977)

Table 3.11: Mean (Coverage Probability) of the OPEE and TPEE algorithm-estimated half-lives and risk ratios under Binary vs. Dose-based models in the context of 1,000 simulations of the Dosing variation scenario.

		Half-Life		
		Incline	Decline	Hazard Ratio
Dose Based	OPEE	91.20 (94.2%)		1.503 (95.4%)
	TPEE	100.01 (95.0%)	91.37 (89.8%)	1.515 (97.2%)
Binary	OPEE	90.91 (93.3%)		1.874 (0%)
	TPEE	103.99 (93.5%)	90.71 (87.9%)	1.911 (1.3%)

3.4.10 Initialization

Table 3.12 shows the parameter estimates by EE algorithm. Coverage and bias of both the hazard ratio and half-life estimates were nearly the same for all OPEE algorithms, regardless of initialization half-life. Only one simulation of the 1,000 sets failed the OPEE (initialized at 60 days), but this would not have been reflected in the final results for the BC1 simulations, because I ran all "failed" simulations with a secondary initial value (100 days).

The results for the TPEE algorithm by initial value were not as consistent. Roughly 20 simulations failed in terms of the hazard ratio estimation, but this was tied back to failures from estimating the half-life parameters. In particular, removing the failed simulations resulted in a coverage probability of the HR of 97.3 for both the 100 and 110 day initialized algorithms, and a % bias of the HR parameter of 7.1 and 60.6 for these two initial half-lives, respectively.

In general, the initialized half-life appears to affect some aspects of the EE model performance, though it was more egregious in the TPEE application.

3.4.11 Interval Analyses

Table 3.13 shows how both *plr1* and *cph1* tend to overestimate the half-life life in OPEE, with increasing bias as interval length increases. Recall that this structure miscalculates the true effective exposure for all individuals equally. Events are thus assigned an EE level associated with longer survival, which naturally means that the time-to-effect would appear longer. Meanwhile, structure 2 imposes different biases on EE for cases and non-cases, which underestimates the true OPEE lag at the 300 and 900 day intervals. More interval-analysis results can be found in the Appendix tables (C.10 and C.9).

Table 3.12: Base Case Algorithm Performance by Initialization

	Initial	OPEE				TPEE			
Parameter	Guess[1]	Mean	CP	% Bias	# Fail	Mean	CP	% Bias	# Fail
Hazard Ratio	60 days	1.51	95.9	0.47	0	1.61	97.4	7.2	22
	70 days	1.51	96.1	0.45	0	1.67	97.2	11.1	24
	80 days	1.51	96	0.45	0	1.63	97.3	8.9	19
	90 days	1.51	96	0.46	0	1.60	97.5	6.9	21
	100 days	1.51	95.9	0.44	0	1.61	97.3	7.1	24
	110 days	1.51	96.2	0.45	0	1.61	97.3	60.6	21
Incline Half-Life	60 days	92.22	92.9	2.47	1	126.27	90.3	40.3	42
	70 days	92.02	93.2	2.25	0	126.54	90.3	40.6	43
	80 days	91.91	93.2	2.13	0	127.57	90	41.8	41
	90 days	92.14	92.9	2.38	0	125.41	90.3	39.4	40
	100 days	91.87	92.9	2.08	0	122.58	90.4	36.2	45
	110 days	91.99	92.9	2.21	0	129.55	90	44	41
Decline Half-Life	60 days					97.99	89.1	8.9	69
	70 days					98.25	88.9	9.2	73
	80 days					98.30	89.4	9.2	69
	90 days					98.57	89.5	9.5	72
	100 days					99.00	89.7	10	74
	110 days					98.42	89.3	9.4	71

[1] Single initialized half-life value at the start of the algorithm's fitting process
OPEE: One-Parameter Effective Exposure Algorithm Results
TPEE: Two-Parameter Effective Exposure Algorithm Results
CP: Coverage Probability

The overestimation in structure 1 also occurred in *plr1* estimates of the hazard ratio, while *cph1* showed the best coverage of this estimate across all intervals and OPEE analytic approaches (*plr1* vs *plr2* vs *cph1* vs *cph2*). The method showing the best coverage of the OPEE half-life parameter was *cph2*, except for the 900-day interval length, because the risks sets at each event time are compared to the study-end EE for all non-cases. In this particular situation, where only baseline and study-end data contribute to the analysis, or where there is little information on time spent in transition, the both CPH models overestimate the hazard ratio.

Applying the TPEE algorithm to the interval datasets resulted in overestimation of HR, regardless of interval length (Appendix Table C.9). It is possible that this overestimation in the short interval lengths, specifically for cph1, *plr1*, and

plr2, is related to the overestimated half-lives for both parameters. Despite appearing to have low bias in estimating the hazard ratio for the 10-day interval length (HR=1.49), the *cph2* model failed in the TPEE algorithm for 425 of the thousand simulated samples. Failures rates of the TPEE algorithm increased with increasing interval length, except for the *cph1* models, though bias in estimation of both lag parameters became worse for all models with larger intervals.

In Table C.10 the Current vs. Never group's (categorical model) HR estimate appears to have the best coverage of the true HR, compared to the other conventional measures of exposure. The calculated EE in structure 1's CPH models (cph1) mirrors the behavior of the true EE estimates for coverage and bias in *cph2*. This is because the calculation applied to structure 1 only changes the exposure-level classification for the cases at their true event time in the interval. In fact, in the last two columns of table C.10, the coverage and bias appear the same for the true parameter fit (TrueEE) in both CPH models, until the interval length exceeds the half-life, at which point the estimates diverge.

3.4.12 Failures in OPEE/TPEE Algorithms

Failure of the OPEE algorithm was seen more often in the smallest sample size (N=1,000 participants failed 5 times in the OPEE algorithm), null hazard, and long single half-life [1,000 days] scenarios. The null lag models, where the underlying scenario had either a null HR for events or a half-life of one-day, "failed" in the majority of TPEE algorithms. Since the Half1 did not see these same rates of failure for the OPEE algorithms, it may be that the type of "failure" here comes from problems in the normal approximation of the half-lives' confidence bounds – specifically when approximating the decline bounds (per coverage probability of

21.6% across the 1,000 simulations, while the incline's coverage was 86% for these same simulations).

The large number of failures in the TPEE estimation of Out10 can be explained by the failed normality of the likelihood surface when the half-life estimate for the decline gets too small. By removing the [nearly half of the] simulations that failed, coverage of the decline parameter changes from 51% to 94%, while the incline half-life and ratio coverage probabilities improve by roughly 2% points. Additionally, for both failed and non-failed simulations, the minimum AIC model selection prefers the OPEE for the Out10 scenario (Table C.4).

It is possible that the underlying issue here (large failure rates in Out10) is actually due to the lack of information available from the short decline. Specifically, the *down* group drops in risk almost immediately, meaning the proportion of time spent in transition is 1/90th of the study period.

To resolve this, I could set the incline half-life to 1,000 days and the decline to 90. This is roughly the same magnitude as 90,10, but I would be able to keep the other parameter constant with the base case. Increasing both of the half-life parameters, while maintaining a larger incline half-life, could also be considered to try to have both half-lives within an estimate-able range for the study design.

Alternatively, I could stagger entry for the *down* group, to see if adding more *down* individuals to the time-dependent risk sets could improve estimation performance, regardless of the input incline and decline parameters. However, before attempting to methodologically address this combination, it is more important to understand the clinical parallel to this scenario. The biologic relationship, here, assumes the EE rises at a slower rate than it falls. An example could be the use of anti-psychotic medications that take a long time to activate symptom reduction,

but for which relapse can happen immediately.(Agid et al., 2006)

3.5 CONCLUSIONS

The simulation study presented here accounts for multiple scenarios and situations where the OPEE and TPEE algorithms may break. This includes multi-trajectory scheme and multivariate models. My methods are bounded by reducing sample size, decreasing the number of trajectories that contain information on rise and fall, hazard ratios that approach null (specifically less than a 1.2 hazard ratio for exposed), and when the half-life of the effect is nearly null or as long as the study length. Additionally, the time-varying estimation suffers when wide measurement intervals are imposed, especially when events are assigned a time-biased exposure level (i.e. structure 1).

Under a null model, the algorithms typically fail in some way. Thus, in situations where the OPEE and TPEE "blow up", one can fit some fixed half-life models and other conventional metrics for exposure with those results consistently pointing towards no association.

For reference, I have presented the results of exposure metrics that may be considered standard for a follow-up study of time-varying exposure and binary time-dependent outcomes. The categorical model was the only conventional metric model that was selected by AIC criteria across all the simulation scenarios (table C.4). This mostly occurred for scenarios where the simulations failed TPEE and/or OPEE. In the multivariate setting, adding "unstable" smokers (MVmulti) to the group reduced bias in the hazard ratio estimate for past-exposed individuals compared to the multivariate monotonic trajectories scenario.

The overarching theme of this chapter has been that more information content

improves estimation of both the lag and hazard ratio parameters. This is seen through the addition of a "fourth" group to any of the scenarios, where the inclusion of the *on* group[13], regardless of the other trajectories, improves estimation of the steady state's magnitude of association.

This "more info" concept is especially evident in the MVmulti scenario's correct AIC-selection of the OPEE model, even in the presence of OPEE and/or TPEE failure (all 1,000 simulations selected the OPEE model). The trajectories of "unstable" smoking participants (MVMulti) additionally lowered the bias and increased coverage of the half-life and HR estimates from the OPEE algorithm as compared to the MVmono scenario. The improvements were also seen in the conventional exposure analyses in the MVmulti vs. MVmono scenarios. This is because the time-varying indicators for smoking status also account for the changing exposures in each risk set that feeds into estimation of the latent EE trajectories.

While my dissertation does not focus on inference or hypothesis testing, I do believe that the Multivariate Multi-Trajectory scenario could hold a key to performance in that realm of statistics. Particularly, the increased complexity of this dataset, specifically addition of "on-again off-again" life-course smokers to the sample, should provide more robust estimation of the true lag and effect parameters. Naturally, adding subjects improves power, but this also benefits the knowledge-base of any exposure-response relationship in that the full cohort reflects a more representative sampling of the entire population.

For the interval-based analyses, almost all the OPEE model types and interval lengths had minimal bias in the estimation of the true HR. For the study-length interval (900 days), which can be paralleled to a case-control study design, OPEE

[13] recalling these individuals are assumed to be at or near steady state maximum risk

and TPEE are not recommended. However, this may be circumvented by creating pools or risk-sets artificially – presuming the EE information gets appropriately assigned for non-cases in those imposed intervals. This, in turn, creates an analytic dataset that looks like the original data, in which one observation is specific per subject per unique event time in the study. Given the known start and stop times for exposure, one can compute the EE at any time point even if it is not measured. Thus, preprocessing the data allows for analysis in a semi-continuous form that circumvents the problems associated with interval-based analyses. In particular, the function behaves as a time-transform for each subject's exposure history.

3.5.1 Limitations

As I did not explicitly model covariate relationships in this study, I am not able to comment on the impact of correlation between potential confounders and the EE. However, I could extend the study to look at this by using data from the multivariate scenarios to restrict the set of confounders in the model during the estimation process. This should provide insight regarding the ability of the model to estimate the lagged association when missing important confounders, such as hypertension and alcohol use.

Another limitation of this simulation study is that I did not explicitly investigate the effect of competing events. I tried to account for some of this in the multivariate scenario, by using real data that included right-censored individuals at loss to follow-up, death, or cancer.

Both TPEE and OPEE univariate models assume balanced study designs and do not account for random variation in the start and stop times of regimens. The random trajectories taken for the MVmulti scenario could be the contributor to the

lower percent bias seen in the half-life estimates of both EE algorithms. The two multivariate scenarios also present more of a population-based balance of subjects in each risk group –i.e. where the base case and its variants primarily included 2/3 exposed vs. 1/3 unexposed individuals, the BWHS underlying sample had proportionally far more non-smokers (3/4 and 3/5 for MVmono and MVmulti, respectively).

While I did not expect the OPEE algorithm to perform well under a true TPEE model scenario, the high failure rate of the TPEE algorithm applied to a true OPEE model scenario was not expected. Improvements can be made to the TPEE algorithm that would allow for better (more precise) estimation of a true OPEE model.

3.5.2 Strengths

By generating the data on a daily basis, I have already conditioned the time-to-event on the survival up to that time point at a fine gradient of possible survival times. Since time is relative to the half-life parameter, it would be just as feasible to generate events that occur in monthly or yearly units, as long as the granularity of the risk over time is preserved. To check the bounds on this conclusion, I could investigate a large half-life within the range of the study period. It would be good to know the threshold at which this estimation becomes problematic. For my purposes, I used a study-length equivalent to at least ten half-lives.

Both sets of MV scenarios had TPEE estimates of the two-lag parameters converging towards an OPEE model. A future step to consider would be the construction of a test to determine whether the TPEE or OPEE is more appropriate. The AIC presents a potential method for model selection, however, this is not a formal hypothesis test.

Table 3.13: Interval-Based OPEE Results. True Half-Life Parameter = 90 days and True Ratio = 1.5

								Amongst Non-Failed				
intlen	modtype	CP.Half	Half$_\mu$	Half$_{med}$	CP.Ratio	Ratio$_\mu$	# Failed	CP.Half	Half$_\mu$	Half$_{med}$	CP.Ratio	Ratio$_\mu$
10 days	cph1	94.30	95.79	93.75	95.10	1.50	0	94.30	95.79	93.75	95.10	1.50
	cph2	94.60	91.81	87.50	76.30	1.49	0	94.60	91.81	87.50	76.30	1.49
	plr1	93.10	95.95	93.75	84.50	1.51	0	93.10	95.95	93.75	84.50	1.51
	plr2	92.60	92.04	87.50	83.80	1.51	0	92.60	92.04	87.50	83.80	1.51
50 days	cph1	92.10	113.60	112.50	94.90	1.51	0	92.10	113.60	112.50	94.90	1.51
	cph2	96.30	93.37	93.75	66.80	1.49	0	96.30	93.37	93.75	66.80	1.49
	plr1	89.50	113.49	112.50	81.20	1.53	0	89.50	113.49	112.50	81.20	1.53
	plr2	92.50	93.30	93.75	83.90	1.51	0	92.50	93.30	93.75	83.90	1.51
100 days	cph1	78.30	135.66	131.25	94.00	1.52	0	78.30	135.66	131.25	94.00	1.52
	cph2	97.10	96.65	93.75	65.10	1.48	1	97.20	96.74	93.75	65.17	1.48
	plr1	74.30	135.06	131.25	78.20	1.54	0	74.30	135.06	131.25	78.20	1.54
	plr2	94.20	96.39	93.75	83.60	1.50	0	94.20	96.39	93.75	83.60	1.50
300 days	cph	26.70	215.05	212.50	93.60	1.55	1	26.73	215.26	212.50	93.69	1.55
	cph2	69.00	74.88	81.25	22.10	1.42	9	69.63	75.18	81.25	22.30	1.42
	plr	19.40	209.94	209.38	67.40	1.58	0	19.40	209.94	209.38	67.40	1.58
	plr2	68.10	74.45	81.25	65.00	1.44	0	68.10	74.45	81.25	65.00	1.44
900 days	cph1	55.30	360.54	350.00	92.50	1.52	3	55.47	361.58	350.00	92.78	1.52
	cph2	0.00	37.92	37.50	0.00	1.39	1000	NA	NA	NA	NA	NA
	plr1	32.20	356.43	350.00	70.70	1.60	17	32.76	362.37	350.00	71.92	1.60
	plr2	19.60	40.43	40.62	63.90	1.43	0	19.60	40.43	40.62	63.90	1.43

CP: Coverage Probability; μ: Mean across simulations; med: Median across simulations; CPH: Cox Proportional Hazards Regression; PLR: Pooled Logistic Regression
Structure 1 (cph1, plr1): Data split into time intervals at end of interval. Cases and Non-Cases assigned exposure at interval times.
Structure 2 (cph2, plr2): Data split into time intervals at end of interval. Cases assigned exposure associated with event timing and Non-Cases designated by exposure at interval times.

CHAPTER 4
Application to Real Data

4.1 BACKGROUND

4.1.1 Cigarette Smoking and Cardiovascular Diseases

Cigarette smoking has been shown to cause build-up of plaque in the arteries, also known as atherosclerosis, a precursor condition to more advanced forms of cardiovascular disease (CVD). Over a three-year period, the ARIC study found that disease progression of atherosclerosis, as measured by the intima-medial thickness of the carotid artery, was 50% increased amongst current smokers compared to non-smokers.(Howard et al., 1998) For one of the largely-studied CVDs, myocardial infarction (MI), smoking accounts for 36% of the population-attributable risk of a first MI.(Yusuf et al., 2004) Previous research has also shown that there is also a dose-response relationship of smoking and CVD, in that increasing the number of cigarettes smoked per day increases the risk of CVD.(Rosenberg et al., 1990; Tolstrup et al., 2014; Rogot & Murray, 1980; Teo et al., 2006) Due to these established relationships between CVD and smoking, clinicians and health policy have largely focused efforts on promoting smoking cessation.

Time-to-reduction in CVD hazard due to smoking has become particularly interesting for researchers. Rachet et al. (2003) explored this lagged association employing B-splines and the Cox proportional hazards (CPH) framework to estimate the time-to-return to baseline hazard for Framingham Heart Study participants who successfully quit. The investigators showed that the flexible modeling approach could estimate a distribution for the lag, assuming that lag of effect varied by subject, that was consistent with prior epidemiologic findings of roughly 3.4

years to reduction in risk of heart attack following complete cessation of smoking. The authors' technique was limited to individuals with successful smoking cessation, implying that the method could only generalize to situations in which the data is unidirectional. Additionally, Rachet et al. discussed the limitations of their results in the context of not knowing the true timing of exposure.

4.1.2 Aims

The goal of this chapter is to demonstrate an application of the Effective Exposure (EE) methodology to data from the Black Women's Health Study, a longitudinal cohort of approximately 59,000 African-American enrolled in 1995 and followed biennially.(Rosenberg et al., 1995) I will show that the underlying EE for smoking in relation to increased risk of CVD can be modeled using an exponential curve that plateaus at a maximum hazard level once the individual has smoked for a prolonged period of time.

In my analyses, I consider all events and exposures occurring between baseline and follow-up in 2015. Specifically, I estimate the increased hazard due to smoking exposure in the CPH analytic framework using time-to-first CVD event as my outcome. The OPEE and TPEE algorithms are applied to smoking as a binary ("on" vs "off") exposure, as well as, dosing based on the number of packs smoked per day. By looking at the profile likelihood surface for a range of plausible lags, I am able to evaluate the validity of my model assumptions and estimation performance.

To emphasize the usefulness of my approach, I compare my results to those of conventional exposure variables for smoking exposure and their estimated HR for CVD. By the end of the chapter, I hope to demonstrate the interpretability of the EE results, and will reflect on the strengths and limitations of my findings, in terms

of the method overall and by comparison to previous literature on the association between smoking (cessation, particularly) and several cardiovascular outcomes.

4.2 METHODS

4.2.1 Study Design

The Black Women's Health Study (BWHS) is a prospective cohort study comprised of 59,000 African American women from across the US, ages 21-69, who responded to a mailed 14-page questionnaire in 1995.(Rosenberg et al., 1995) As of 2013, biennial questionnaires had been completed with an overall follow-up of 88%, providing information on health events and various exposures, with some ascertained more frequently than others. The Institutional Review Board at Boston University granted approval for the BWHS and all subjects provided written informed consent.

4.2.2 Cardiovascular Disease – Outcome Specification

For these analyses, I consider several CVD conditions as the outcomes of interest, setting the first reported event of any one of these as the time-to-event. These include self-reported MI, stroke, congestive heart failure, and coronary artery bypass and grafting procedures. Participants were asked about one or more of these conditions at all questionnaires, along with the year of first diagnosis.

When the year of diagnosis is not known, the questionnaire cycle in which the CVD event was reported is taken as the year of the event. Additionally, if a subject was found to have died from an underlying CVD event, as depicted by an ICD10 "I" code on a death certificate, this is marked as a CVD death and considered an outcome in the year of death.

During the assignment of person-years of follow-up, the BWHS analyses typically assume all events to have occurred at the mid-point of the year. This accounts for potential misclassification of exposure for events that do not have a corresponding month of diagnosis, though it imposes some bias on those with a known exact time of event. For the purposes of my analyses, I assume all events occur at the end of the year of diagnosis. The implications of this assumption are discussed in more detail in the conclusions.

The BWHS is currently abstracting reports of CVD conditions with the hopes of creating a validated and confirmed case-set. Due to this ongoing process, my dataset is made up of both non-confirmed and confirmed cases, however, I have removed individual cases that have been disconfirmed. This is important to remember, as the clinical relevance of these results should be approached with caution. To quote D'Agostino et al. (1990), "the examples [here] are presented mainly for *comparison of the methods*; the reader should not view them for definitive substantive interpretation".

4.2.3 Smoking and Cigarettes/Day – Exposure Specifications

At baseline, in 1995, women were asked about smoking in terms of age at initiation, number of cigarettes smoked per day (on average) for early and more recent years of use, whether they quit and how long ago, and total duration of smoking prior to enrollment. This set of questions provides information for classifying individuals as current vs. past vs. never smokers, with only 38 women missing data on smoking at baseline [1]. During sensitivity analyses of past and current smokers (see section 4.2.6), only those with at least 10 years of prior smoking history at

[1] Additional had implausible years of smoking duration or missing years since quitting, which resulted in their exclusion from the analytic sample.

baseline are included.

In follow-up cycles, smoking was asked about in terms of cigarettes/day for every 2-year questionnaire, except in 2011. Smoking during pregnancy was asked in 1997, 1999, 2001, and 2003, while use of menthol cigarettes questions appeared in the 2003, 2005, 2007, 2009, and 2013 questionnaires.

The smoking data from the BWHS cohort has been cleaned for consistency of responses across time, as well as, within cycle for reported menthol cigarette use. Subjects with missing information during one or two follow-up periods, with consistent bordering responses, are assumed to continue the pattern of smoking as reported in the bounding cycles. For example, if a woman answered "non-smoker" in 1999 and "non-smoker" on the 2003 questionnaire, her missing 2001 cycle was set to non-smoker. Additionally, when cigarettes/day (frequency/intensity) was not provided for a given cycle, information was carried forward or back-filled from adjacent cycles [2].

For use in "conventional" analyses, the time-varying dataset includes variables of smoking status category (current vs. past vs. never), history of smoking (ever vs. never), current number of cigarettes or packs/day (continuous, on average for the 2-year period), and cumulative number of years smoked. When a cycle's smoking status for an individual is not known, the subject is excluded from the risk set for that particular year.

Women are classified into one of five risk trajectories, as defined by the time-varying smoking status across all follow-up periods: smokers throughout, non-smokers throughout, successful quitters, smoking initiators, and those with an unstable smoking status. The first two categories require a baseline report of never

[2]Provided the adjacent cycle include non-missing information on cigarettes smoked per day and the same binary smoking status as the one with missing data.

or current smoking, with no change to smoking or non-smoking status, respectively, throughout the 20 years of follow-up. Missing cycles do not change the risk profiles for these women, as transition from one status to missing and back is not considered to be a change in the exposure trajectory. The risk trajectories and the number of women, person-years, and cases that fall into each group are described in more detail in the results section 4.3 of this chapter.

4.2.4 Additional Covariates and Confounders

All models are adjusted for the following time-varying covariates: body mass index (BMI in kg/m^2, continuous); frequency of vigorous exercise (none vs. <1 hour/week vs. 1+ hours/week); current menopausal status (pre- vs. post- vs. unknown) and age at menopause (<45 vs. 45-<50 vs. 50+ years old for postmenopausal women only); alcohol consumption (current vs. past vs. never); history of diabetes (ever vs. never); history of high cholesterol and cholesterol medication/statins use (ever vs. never); and history of hypertension treated with medication (ever vs. never). Additionally, family history of cardiovascular disease (specifically, stroke and myocardial infarction), a fixed risk factor, is included in the models for all participants. The majority of these confounders were collected at each questionnaire cycle, though the carry-forward method is used to assign values at time points where this information can not be updated.

Subjects that did not report a history of any of the confounding conditions are assumed to be unaffected at baseline, and are set to "yes" at the first report of diabetes, high cholesterol, or hypertension with concurrent use of medications. Use of statins was collected on some questionnaires explicitly and as write-in responses on others. Once set to "at risk" for the condition/statin use, that history of "ever" is

carried forward for the remainder of follow-up.

4.2.5 Data Preparation

The BWHS data was first converted into an Andersen-Gill(Andersen & Gill, 1982) dataset in SAS 9.3(SAS Institute Inc., Cary, 2011), i.e. multiple rows per subject with time-varying exposures updated at each available questionnaire cycle with non-missing information. While each cycle included two years of follow-up, the observations were split into annual risk sets, to preserve the interval-based comparisons of Effective Exposure for cases and non-cases.

Having an appropriately structured dataset is key in the OPEE and TPEE frameworks, as well as, crucial for traditional time-varying covariate analyses in BWHS. The only consequence to pre-preparing the data into multiple observations per subject comes in the form of computational costs, as the results and precision are not affected.(Therneau & Grambsch, 2000) If the time measure for the time-to-event outcome is conditional on the interval, then this equates to the Cross-Sectional Pooling with time-adjustment model.(Ngwa et al., 2016) This CSP method is just a Cox proportional hazards model fit within annual intervals and strata of current age, assuming separate baseline hazards for each strata-year, rather than a single baseline hazard per individual (as might be used in a mixed effects modeling approach).

This same dataset structure can be used in a traditional Time-Dependent Covariate Model (TDCM), where the interval is not included as a stratification level, but rather each observation's start and stop times reflect the particular measures contributing to the risk set at the stop-time-event. Ngwa et al. (2016) showed that these two methods, time-adjusted CSP and TDCM are identical in performance

and estimation of longitudinal exposures and time-to-event outcomes. Age is left in the strata for all CPH models, to avoid issues with non-proportional hazards by age.

To account for tied event times, the Breslow(Breslow, 1974) method is used in all CPH model fits.

4.2.6 Restricted Sample Analyses

To understand the role of smoking cessation on risk of CVD, I considered a simplified or restricted subset that could be useful in comparison to the single-dosing simulation scenarios. This dataset also attempts to mimic the study-design used by Rachet et al. (2003) in the estimation of the lag distribution.

In particular, I restricted the individuals for this set to one of three risk trajectories: Non-smokers throughout, smokers throughout, and successful smoking cessators (quitters). The latter group included individuals who may have quit prior to the study entry, or at some point within the study, as long as there was no report of smoking for the remainder of the subject's follow-up and a minimum of two follow-up cycles as non-smokers. The successful cessation group, as well as, those who were smokers throughout, were further restricted to women with at least 10 years of smoking exposure at the baseline interview. This restriction did not account for differences in the intensity of smoking, or number of packs smoked per day[3]. By forcing all exposed individuals to have 10+ years of exposure at baseline, if the true EE half-life were between 2 and 5 years, then the first component of the multi-dosing EE should reflect a level where exposed individuals start within 2-5 half-lives of exposure.

[3]Though all subjects in both the restricted and full cohort datasets were required to not have missing information on packs/day

4.2.7 Conventional Analyses

The first set of exclusions applied to this full set removed missing or implausible information on baseline smoking status (N = 360). Individuals with prevalent CVD or prior reports of stroke, myocardial infarction, coronary bypass surgery or grafting, congestive heart failure, and other cardiovascular procedures at baseline are excluded from the analyses (N = 1,562). Additionally, subjects are excluded due to missing information on any of the covariates of interest (N = 4,064), with the exception of menopausal status. The total number of women excluded at baseline due to prevalent cancer or incidence of cancer within the first follow-up cycle is N = 1,685.

Previous research within the BWHS cohort has utilized PROC PHREG in SAS to conduct CPH regression with separate baseline hazards estimated for age and period combinations. As described in chapter 2, my algorithms all make use of R software, with several key packages loaded.R Core Team (2017) For the purposes of this dissertation, a similar approach is constructed within R 3.2.5 using the "survival" package(Therneau, 2015) to be able to compare estimates of various smoking exposure classifications on CVD hazard. These classifications include time-varying definitions of current/past/never smoking status, ever vs. never smoked, number of cigarettes or packs per day, and cumulative years of smoking.

4.2.8 Effective Exposure Approach

After performing analyses the "traditional" way, I apply the OPEE and TPEE models, using the profile likelihood (PLL) method and both of the lag-estimation algorithms. The goal of the PLL method is to fit all possible combinations of incline and decline parameters, to see the shape of the joint parameter likelihood sur-

face, over which the algorithms search for a maximum. This approach can be very computationally-intensive and expensive as it requires calculation of each subject's EE for a combination of incline and decline half-lives, followed by the fitting of the CPH model:

$$h(t) = h_0(t) \exp\left[\beta E_t(h_1, h_2) + \mathbf{\Gamma} \boldsymbol{X}\right]$$

Where $\boldsymbol{X}$ includes the fixed and time-varying covariates, and $\mathbf{\Gamma}$ denotes the coefficients representing the excess hazard due to each covariate. The CPH model fits include the log-likelihood, AIC, and effect estimates, corresponding to the maximum likelihood estimators conditional on the pair of lags. Each coordinate of the half-life combination is then used to plot the overall surface of the profile log-likelihoods.

I use the profile likelihood method to compare estimates of CVD hazard and smoking across possible half-lives from two-weeks ($\sim$0.05 years) to 20 years, in increments of 0.05 years. The maximum log-likelihoods and corresponding confidence bounds for both 1- and 2-lag-parameter models are compared to my estimation algorithm results.

4.2.8.1 Profile Likelihood

In the TPEE context, in order to ascertain the confidence bounds of a single parameter, I first subset all the likelihood fits by levels of a single parameter. The maximum likelihood from each fixed level is used to create the specific profile across that parameter. The points of this profile that have a log-likelihood less than 3 units away from the maximum provide the range of values for the 95% confidence interval, using the 2-degree of freedom Chi-Square (χ^2) statistic. I consider the 2-df chi-square instead of the 1-df, because the half-life parameters are not independent

and the joint distribution is more appropriate than the marginal.

As the fits across two parameters will contain the one parameter likelihoods, where the incline and decline parameters are equal, the marginal distribution (1-df) can be used to construct the OPEE confidence bounds.

4.2.8.2 Asymptotic Normality

I assume that the true lag of the EE smoking and CVD hazard is normally distributed, to calculate the 95% confidence intervals for the OPEE and TPEE lag and effect parameters using the Information Matrix-derived standard errors and a Z=1.96.

4.3 RESULTS

There were 323 cases that were censored due to loss to follow-up and 299 censored due to cancer incidence prior to becoming a CVD case. The full analytic cohort includes a total of 2,786 cases over 786,139 person-years for 51,303 women. In the full sample 3,705 subjects remained smokers throughout the study, 32,534 stayed non-smokers, and 9,433 and 153 subjects quit smoking successfully or started smoking (and did not quit) at some point during the study, respectively. The final group consisted of 5,478 subjects who had an "unstable" smoking status throughout the study, implying that they may have quit and returned to smoking, or one of several multiple-change-points trajectories over the course of follow-up.

In the restricted subset of quitters, smokers and non-smokers throughout, there were 2,396 total cases, of which all the non-smoking and smoker-throughout cases stayed in the analysis. The number of cases per person-years for "smokers" and "quitters" in the full analytic set is 749/148,836 and 352/47,665, respectively. For

the restricted analyses, the corresponding cases/person-years is 707/138,159 and 332/41,888, for the "smokers" and "quitters", respectively. In both samples analyzed, the number of cases and person-years remained the same for those who were never smokers throughout the study (1,301/512,257). The total number of person-years for the restricted analysis decreased to 692,636.

4.3.1 Participant Characteristics

Table 4.1 shows the population characteristics at baseline (1995) by baseline smoking status in terms of current vs. past vs. never smokers. All the numbers in the table are age-adjusted to the sample's age-distribution, and either represent the mean (standard deviation) or row proportion. The baseline characteristics are also presented for the restricted sample.

Smokers and past-smokers were more often drinkers (current or past), and older than never smokers at baseline. Nearly a fifth of all participants had high cholesterol at baseline, and just over a fifth were being treated for hypertension. The restricted sample included 4,772 and 7,231 participants that smoke or used to smoke at baseline, while the full sample included an additional 3,124 current smokers and 2,482 past smokers at baseline. The reference group at baseline, of never smokers, is based on 32,534 and 33,694 individuals for the restricted and full samples, respectively. Of the smokers at baseline in the restricted sample, 1,534 stopped smoking during some point in follow-up and were deemed to be successful at quitting (no return to smoking with at least 2 non-smoker follow-up cycles of data following cessation). The difference in the number of never smokers at baseline in the full and restricted samples comes from the 1,160 women who initiated smoking at some point during the study, of which 153 remained smokers

Table 4.1: Age-Adjusted Baseline Characteristics by Smoking Status in the Full and Restricted Black Women's Health Study Samples

	Baseline Smoking Exposure					
	Full Sample			**Restricted Sample**		
Baseline Characteristic[1]	Current	Past	Never	Current	Past	Never
N	7,896	9,713	33,694	4,772	7,231	32,534
Age (in years)[2]	40.2(9.3)	43.9(10.1)	36.3(10.1)	41.9(8.6)	45.6(9.6)	36.4(10.1)
Body Mass Index (kg/m^2)	27.7(6.4)	28.6(7.1)	27.8(6.6)	27.7(6.5)	28.4(6.9)	27.8(6.6)
Pack-Years of Exposure[3]	12.1(11.1)	8.0(10.5)	0.0(0.0)	13.8(10.9)	8.0(10.5)	0.0(0.0)
Family History of CVD, %[4]	41.2	39.70	38.20	39.6	40.5	38.4
History of Comorbidities						
High Cholesterol, %	18.10	19.7	19.10	17.8	20.2	19.40
Statins Use, %	0.8	0.8	0.7	0.8	0.7	0.7
Type 2 Diabetes, %	3.9	4.40	3.8	3.8	4.3	3.9
Treated Hypertension, %	23.1	22.2	22.5	23	22.7	22.8
Vigorous Activity Level						
None, %	41.1	30.6	31.7	42	30.2	31.8
<1 hour/week, %	16.10	16.10	16.8	17.2	16	16.8
$\geq$1 hour/week, %	42.8	53.2	51.5	40.80	53.7	51.4
Menopausal Status						
Premenopausal, %	76.7	78.10	78.7	76.10	77.60	78.10
Unknown or Dubious, %	6	6.2	6.2	6	6.4	6.3
Postmenopausal, %	17.3	15.7	15.1	17.90	16	15.6
Age at Menopause[5]						
<45, %	8.30	6.7	6.2	8.70	6.7	6.3
45-<50, %	4.10	3.7	3.3	4.3	3.7	3.4
50+, %	3.2	3.5	3.8	3.3	3.6	3.9
Unknown, %	1.7	1.9	1.9	1.6	1.9	1.9
Alcohol Use						
Non-Drinker, %	32.20	30.7	70.10	30.2	30.1	70.60
Current Drinker, %	47.4	39.4	17.3	45.8	38.5	16.8
Past Drinker, %	20.40	29.9	12.6	24	31.4	12.6

[1] Values are means(SD) or percentages and are standardized to the age distribution of the study population.
[2] Not age-adjusted.
[3] Cumulative Total of Packs per Day multiplied by Years of Smoking that quantity.
[4] CVD: Cardiovascular Disease; First degree relatives with stroke or myocardial infarction.
[5] Age at Menopause for Women classified as "Post-Menopausal" at Baseline.

throughout the remaining follow-up period.

Table 4.2: Black Women's Health Study Trajectories (1995-2015)

	Smoker Throughout	Non-Smoker Throughout	Successful Cessator	Smoking Initiator	Unstable Smoking
# Cases	352	1,301	749	8	376
# Women	3,705	32,534	9,433	153	5,478

4.3.2 Conventional Analyses

Table 4.3 shows the results from fitting multiple combinations of smoking variables on the complete set of participants from the BWHS sample (1995-2015).

The metric producing the best-fit model, as determined by the largest log-likelihood (-17516.91) and smallest AIC (35067.82), is the one that uses the time-dependent indicator of current smoking and the pack-years of smoking exposure. The estimated effect for a given current or past smoker requires a calculation using the linear combination of the coefficients for the current indicator and a one-unit change pack-years. The current smoker indicator is set to 0 for past smokers, though the effect of past smoking is thought to feed into the pack-years estimate. Meanwhile, both of the combination models assume that never smokers have a pack-years equivalent value of 0, which may violate the proportion hazards assumption for this effect estimate.

The exposure metric whose model performed second best, and which provides a more straight-forward interpretation, was the time-varying categorical model (log-likelihood=-17523.97, AIC=35081.94). For any given time-period and age, after adjusting for the covariates, those who are current smokers are at 2.32 (95% CI: 2.1-2.6) times the hazard of developing CVD compared to never smokers, while past smokers are at 1.23 (95% CI: 1.1-1.3) times the hazard compared to never

smokers.

Table 4.3: Smoking exposure and risk of cardiovascular disease: standard variable approaches comparison in full Black Women's Health Study Dataset (1995-2015)

Smoking Exposure Model		Hazard Ratio	95% Confidence Interval Lower	Upper	Log-Likelihood	AIC
Smoking Categories	Current	2.32	2.10	2.56	-17523.97	35081.94
	Past	1.23	1.12	1.34		
Ever vs. Never		1.54	1.43	1.67	-17588.86	35209.72
Current vs. Not		2.15	1.96	2.36	-17533.85	35099.71
Cumulative Years Smoked		1.02	1.02	1.02	-17546.45	35124.91
Current Smokers: Packs/Day		2.42	2.16	2.71	-17557.76	35147.52
Combination Model 1[1]	Current Smoker	1.86	1.67	2.06	-17516.91	35067.82
	Pack-Years	1.01	1.01	1.01		
Combination Model 2[2]	Ever Smoker	1.29	1.18	1.42	-17566.03	35166.06
	Pack-Years	1.01	1.01	1.01		

AIC: Akaike's Information Criterion
[1,2] Combination models 1 and 2 assume that never smokers have zero pack-years.

4.3.3 Effective Exposure Estimation

The maximum likelihood according to the profile log-likelihood surface is located at an incline half-life of 9.3 years (95% PLL CI: 4.2-20.0) and decline of 7.00 year (95% PLL CI: 4.2-11.7), with a corresponding hazard ratio of 2.8 (95% CI: 2.5-3.1) (Table 4.4). Thus, the CVD hazard for a lifetime smoker is expected to plateau at 2.8 times the never smoker's hazard. The confidence bounds for the half-life parameters in PLL were calculated using the 2-df chi-square distribution, and as the combination of half-lives considered did not extend beyond 20 years, the incline parameter's upper 95% CI is bounded by 20. Naturally, these results raise some questions – how could the detrimental effects of smoking take longer to accumulate than to dissipate? For now I will only focus on the decline parameter's interpretation.

After adjusting for confounders and covariates, a woman's CVD hazard associated with smoking is expected to decrease 50% after 7 years of successful and complete cessation. This is regardless of the starting risk. For a woman starting near the maximum hazard ratio of 2.8 compared to her never-smoker counterpart, her hazard ratio or excess hazard will reduce by 50% (to an HR = 1.89) after 4.8 years.

Using the packs/day classification, the incline and decline half-lives that maximize the likelihood were at 5.75 and 5.85 years, respectively (CI presented in table). As the OPEE model's packs/day maximum likelihood occurred at a single half-life of 5.85 years, it would appear that the packs/day dosing favors a single lag parameter. It is worth noting, however, that selecting to use the TPEE model does not change the HR estimate for CVD in relation to an effective exposure of smoking 1 pack/day.

The single-parameter lag where the profile is maximized for binary smoking falls at 7.1 years (95% PLL CI: 4.6-10.8), with a corresponding relative hazard of 2.5 (95% CI: 2.25-2.77). Here, the single-lag is very close to the decline half-life of the two-parameter profile. This demonstrates how the decline parameter is likely reflecting the true half-life associated with CVD risk reduction following complete cessation. Based on the AIC criteria, the 1-parameter (OPEE) framework is preferable over the TPEE model, in both the binary and packs/day dosings for the EE of smoking (35061.22 vs 35061.96 in binary and 35089.20 vs 35091.20 in packs/day). Meanwhile, the minimum AIC for all of the full sample EE analyses (PLL and algorithmic) selects the binary smoking OPEE model as the preferred metric for CVD hazard by smoking EE.

The second half of table 4.4 shows the OPEE and TPEE algorithm results fol-

Table 4.4: Effective Exposure of Smoking on Risk of Cardiovascular Event. Application of Profile Likelihood Fits and Estimation Algorithms to the Black Women's Health Study Data (1995-2015)

Estimation approach	Exposure Input	HR for Effect of Smoking (95% CI)	Lag Parameter Estimates (95% CI)[2] Incline	Decline	Log-Likelihood	AIC	Total computing cost (in seconds)
Profile Likelihood for One Parameter[1]	Binary	2.50 (2.25-2.77)	7.10 (4.60-10.80)		-17513.61	35061.22	6,984
	Packs/Day	2.63 (2.36-2.93)	5.85 (3.45-9.95)		-17527.60	35089.20	7,397
OPEE	Binary	2.48 (2.21-2.78)	6.75 (3.98-9.52)		-17513.64	35061.29	123
	Packs/Day	2.63 (2.37-2.93)	5.81 (2.84-8.78)		-17527.60	35089.20	205
Profile Likelihood Search Over Two Parameters[1]	Binary	2.78 (2.48-3.12)	9.30 (4.15-20.00)[3]	7.00 (4.15-11.70)	-17512.98	35061.96	2,783,959
	Packs/Day	2.62 (2.35-2.92)	5.75 (2.50-14.45)	5.85 (3.00-12.00)	-17527.6	35091.20	2,956,186
TPEE	Binary	2.76 (2.15-3.55)	9.0 (3.03-14.97)	6.75 (3.91-9.59)	-17513.00	35062.00	467
	Packs/Day	2.64 (2.20-3.16)	6 (1.90-10.10)	6 (2.83-9.17)	-17527.61	35091.21	383

HR: Hazard Ratio; CI: Confidence Interval; OPEE: One Parameter Effective Exposure Algorithm; TPEE: Two Parameter Effective Exposure Algorithm; AIC: Akaike's Information Criterion

[1] Confidence bounds for Profile Likelihood Hazard Ratios based on final model selected, i.e. no adjustments made for multiple testing.

[2] Confidence bounds for half-life lag parameters in OPEE and TPEE algorithms are based on the asymptotically normal approximation of standard errors. Confidence bounds for the profile likelihood methods are based on the likelihood ratio test statistic assuming 2-df and 1-df chi-square distributions for the joint and marginal confidence bounds for the two- and one-parameter profiles, respectively.

[3] Confidence bound stops at 20 years, because profile fits were not performed for half-lives>20.

lowing initialization at a half-life of 1 year. The half-life for the OPEE binary model stopped at 6.75 years vs. 7.10 from the PLL approach. With very similar estimates for the hazard ratio and it's confidence bounds, the algorithmic approach is primarily superior to PLL in the computational burden required ($\sim$2 minutes vs. $\sim$2 hours, OPEE algorithm vs. PLL).[4]. Similarly, the OPEE packs/day model and PLL 1-parameter packs/day model are nearly identical in the results.

The confidence bounds presented for the lag parameters in PLL are wider than those estimated asymptotically following the algorithmic half-life search. For both 2-parameter models (Binary and Packs/Day) and estimation approaches (PLL vs. Algorithm), the decline parameter's half-life confidence bounds are narrower than for the incline parameter. Figures 4.1 and 4.2 show the profile likelihood contours along the incline and decline parameters of the TPEE models using the binary and packs-per-day dose exposures, respectively.

In both figures, the blue solid line represents the values of the decline parameter that maximize the log-likelihood of the full model for a fixed incline parameter. Conversely, the pink dashed line represents the incline value that maximizes the log-likelihood for the full model fit when fixing the decline parameter. The red triangle, where the two lines cross, is the point where the log-likelihood surface is maximized. The contours represent the joint likelihood confidence bounds for the log-likelihood surface [5].

Notably, the contour plot for binary smoking reflects the uneven bounds of the incline and decline parameters, as seen in table 4.4. The width of the incline parameter's 95% joint PLL confidence interval demonstrates the uncertainty sur-

[4]Since the 1-parameter PLL fits were performed within the 2-parameter PLL grid, the computational times presented in Table 4.4 for the PLL modeling approaches are not mutually exclusive

[5]The deviance values that represent the 2-df chi-square distance from the maximum are 2.3, 3.0, and 4.6 for the 90th, 95th, and 99th percentiles, respectively.

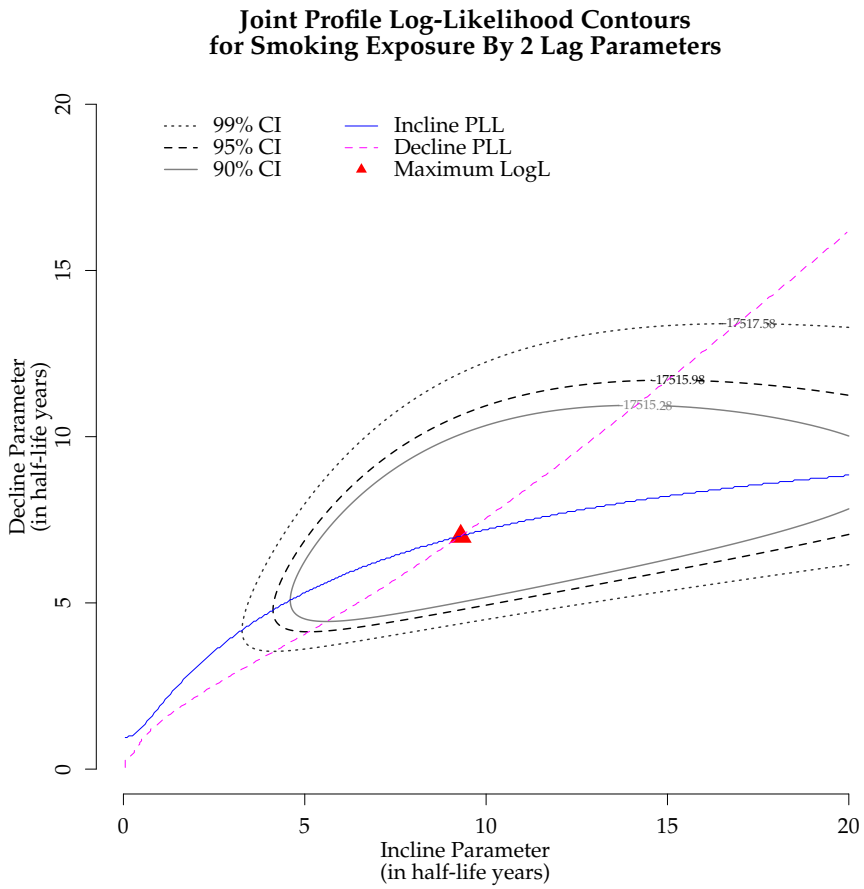


Figure 4.1: Overall Effective Exposure Contour for Log-Likelihoods Across Combinations of Incline and Decline Lag Parameters amongst the full set of BWHS participants.

rounding this parameter that is due to a small number of participants "on the rise". Since smoking is "bad for you", the number of participants that initiate smoking in the study is small, thus it is likely that participants enrolled in a health study are cognizant of the negative effects of smoking, and might therefore be less likely to initiate use during the study. This just implies that the results may not generalize well to new smokers, especially with regards to time-to-plateau of in the CVD hazard after starting to smoke.

In the second contour plot for packs/day (figure 4.2), the individual parame-

ter's profiles are nearly identical for incline and decline half-lives less than 5.75 years. This is consistent with the determination that a single lag parameter may be appropriate in the estimation of the association between smoking and CVD when using a packs/day dosing. As the incline parameter's half-life increases, however, the corresponding decline parameter that maximizes the likelihood stays between 5 and 10 years.

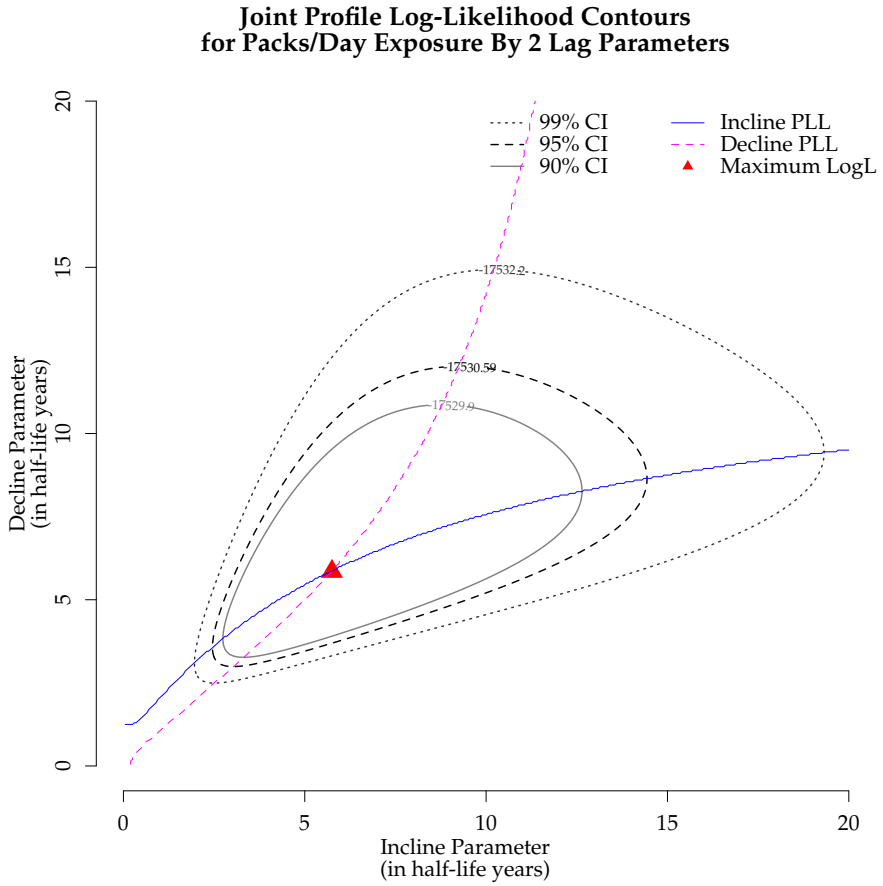


Figure 4.2: Packs Per Day Effective Exposure Contour for Log-Likelihoods Across Combinations of Incline and Decline Lag Parameters amongst the full set of BWHS participants.

The interpretation of the packs/day model is offered in terms of the OPEE framework, due to the points made above. For a consistent 2 packs/day smoker

of 30 years, after accounting for other risk factors of CVD, the CVD hazard associated with a woman's smoking exposure is 6.5 times that of her counterfactual never smoker, while a woman who smoked 1 pack/day for 30 years is at 2.6 times the never smoker's hazard of CVD. A 50% reduction in HR or excess hazard for these same 2- and 1-pack/day smokers, would take 2.9 and 4.1 years following complete and successful cessation. The corresponding hazard ratios would be 3.8 and 1.8, respectively. Alternatively, after 5.85 years of complete and successful quitting of smoking, the 2-pack/day smoker's CVD hazard is expected to reach the 1 pack/day smoker's hazard, i.e., after the half-life number of years, the risk is reduced by 50%. In terms of reduction in hazard ratio, this implies that the hazard ratio of CVD for a 2 pack/day smoker compared to a never smoker reaches the hazard ratio for the 30-year 1 pack/day smoker compared to a never smoker, after 5.85 years of no smoking exposure.

4.3.4 Restricted Sample Results

The baseline characteristics of this restricted sample's participants are shown in Table 4.1. This subset's current and former smokers are older than in the full sample, while the non-smokers are nearly the same age. It could be possible that age and smoking have some interactive effect, but this segmentation is more an artifact of the restriction imposed at baseline to include only those who would already have reached steady state hazard from smoking at baseline (by requiring 10+ years of prior smoking exposure at baseline).

Using the restricted sample of participants, the two models with the largest log-likelihood and smallest AIC, across the conventional measures of smoking, were the same as the "best fit" models from the full sample analyses. The HR estimates

in this subset were consistently larger for the current smokers and smaller for the past and ever smoker measures, likely due to the large number of women that would have contributed to these group being excluded from the varying trajectories group (Table D.1 in Appendix D). The hazard ratios for cumulative years smoked and pack-years did not differ in this analysis, and similarly to the full sample analysis, both of these estimates require interpretation based on a single unit increase in pack-years exposure.

In the restricted sample, the HR estimates are the same for current smokers in terms of packs/day and current smokers in the categorical model. Here, current smokers are at a 2.56-fold CVD hazard compared to never smokers, after adjusting for time-varying factors. This effect size is in the same ballpark as the estimated hazard ratio from the OPEE model in the full sample. It is understandable that the estimated effect size is larger for this analysis, because those classified as current smokers from the "unstable" group are not contributing person-time to the restricted analysis.

As in the full sample, the OPEE PLL and algorithm estimates of lag were nearly the same, with comparable effect sizes (hazard ratios) and lower AICs than each smoking exposure's TPEE model counterpart (28842.84 vs. 28843.36 for binary OPEE vs. TPEE, 28862.56 vs. 28863.84 for packs/day OPEE vs. TPEE). The estimated lag of effect for the binary exposure was 4.5 years with a corresponding 2.52 times the CVD hazard at steady state compared to never smokers. Interestingly, the estimate for the packs/day dosing OPEE half-life was longer (6.9 years) with the CVD hazard associated with a lifetime of smoking 1 pack/day of 2.7 compared to never smokers[6].

[6]after adjustment for covariates, confounders, and age-related baseline hazards

Similar to the full sample results, the restricted set's bounds using the PLL approach were wider than the normally-approximated bounds from the algorithmic results. Also, the decline parameters in both exposure models were closer in magnitude to the OPEE model's single half-life. This is to be expected as this particular set of individuals did not include any initiators or varying risk trajectories.

Contour plots corresponding to the restricted set of BWHS participants can be found in the Appendix D. For this sample, I have chosen to focus on a 3-dimensional visualization of the profile surface, to illustrate the behavior of the log-likelihood. All three plots in figure 4.3 represent the same log-likelihood surface for the BWHS Restricted Subset using the Packs/Day dosing in the TPEE framework (i.e. for coordinates reflecting the fixed incline and decline half-lives). The intersecting plane represents the 95% joint confidence bound for the two lag parameters' profile log-likelihood surface. The maximum likelihood peak occurs at the incline half-life of 8.15 years and decline half-life of 6.4 years.

The likelihood surface is steeper in the direction of the lower bounds for each parameter, indicating that the true half-life of effect is less likely to be close to zero than to be 10-years. This demonstrates that the CVD hazard for women who have quit smoking 1 pack/day does not dissipate immediately, and that it takes at least 3.5 years for the risk to go down by 50%.

4.4 CONCLUSIONS

Compared to conventional smoking exposure measures, the analyses using Effective Exposure showed evidence of a lagged association between smoking and CVD hazard that agrees with previous literature. The nearly three-fold risk of CVD due to smoking, as seen in the restricted sample, has been demonstrated in other

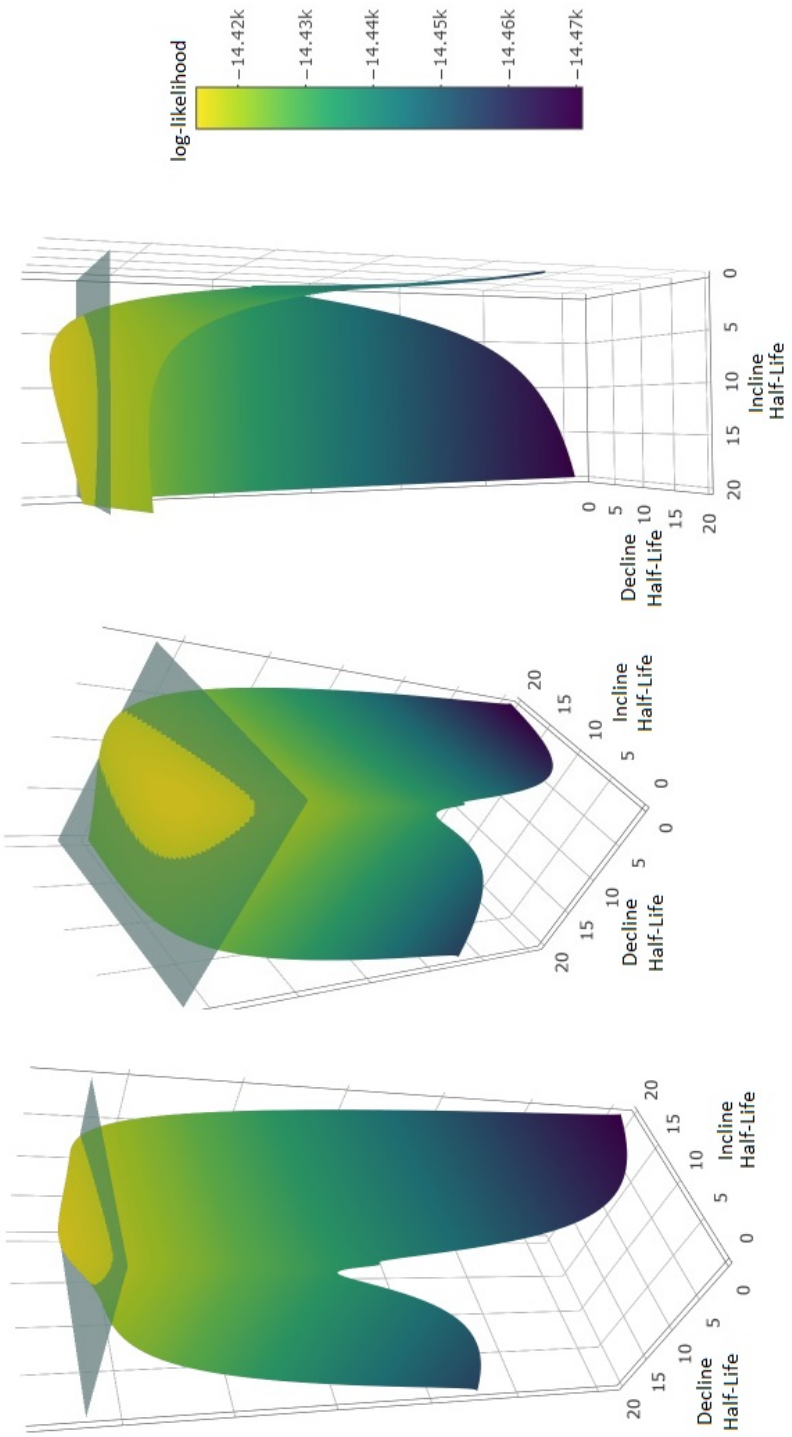


Figure 4.3: 3D Surface of BWHS Profile Log-Likelihood

cohorts and analyses that address the lagged nature of the association between smoking and CVD.(Kawachi et al., 1994; Rachet et al., 2003) However, the current approach allows for researchers to estimate an effect without pre-specifying a lag-time, and is flexible in that the entire sample of participants, regardless of their exposure trajectory, can be included in the analysis.

In the restricted sample's analysis, the half-life estimates for packs/day dosing were longer than those for the binary exposure, which differed from the full sample results.

4.4.1 Limitations

There are many known risk factors for CVD, overall. While I accounted for a majority of these potential confounders, my models did not adjust for oral contraceptives or female hormone use, coffee, education, region of residence, angina, and treatment of type II diabetes. Rosenberg et al. (1990) additionally adjusted for a behavior score, and other literature has noted multiple psychosocial risk factors for acute MI(Tofler, 2017), none of which were considered in this analysis. The lack of adjustment for these variables may confound the results, though it would be interesting to conduct a simulation study to determine the impact of such exclusions on the estimation of the lag. Current literature suggests that there may actually be an interaction between smoking and oral contraceptive use, and potentially other forms of hormone therapy, which could result in different half-life estimates.

My methodology does not account for potential interactions between risk factors, neither between covariates nor with the EE measure. There is a clear need to account for this, as well as, including additional risk factors that were not in these analyses. For example, I would assume that exercise is more difficult for smokers,

and therefore the effect of activity level on CVD hazard may differ by smoking status and time since quitting. This does not even begin to address the possibility of effect modification in the lags of smoking's relationship with CVD due to another risk factor – i.e. typically interaction implies effect modification in the measure of association, though interaction could exist such that time-to-plateau differs by another factor.

One of the major limitations of this applied analysis is the selection of cases. The purpose of the analyses were to demonstrate the methodology for estimating lagged effects in a "real" clinical application. The set of cases selected for the study included various CVD conditions and non-confirmed cases. Currently, the BWHS is performing record abstractions and quality control for CVD cases. Most of the cases that have been self-reported are under review by trained epidemiologists and physicians, meaning that there is a future opportunity to re-analyze the data restricting the outcome to confirmed cases only. This will allow for cleaner information regarding the type of cardiovascular event, as the pathophysiology of the diseases may differ, as well as, the lagged-association between smoking's effective exposure and the hazard.

Smoking cessation is often concurrent with other lifestyle changes, therefore there could be some confounding by indication. My model represents the conditional lagged effect of smoking, based on fixed profiles for the covariates in the model (i.e. adjusting for non-lagged trajectories of other exposures in the model).

The model assumes return to plateau of the never smoker risk of CVD. Therefore, if the underlying biological mechanism is trauma and stiffening of the vascular wall, the model expects this damage to be reversible. However, the CVD hazard could stay elevated for past smokers compared to never smokers, if the assump-

tion of return to pre-smoking risk is incorrect. In this situation, I would expect the estimated lag to appear longer than the true lag, because the lag-time should be infinite with regards to returning to null. On the other hand, any reduction in the hazard, following cessation, should be captured by the two-parameter model. A longer estimated half-life for the decline parameter might account for the decay towards a final exposure level that is greater than null.

The two-parameter model, amongst the entire sample, estimated a slightly longer half-life for the incline lag than the decline lag. The fact that the incline rate appeared slower (i.e. longer half-life) than the decline rate, and with wider confidence intervals, is indicative of the information content available about participants in transition. As mentioned in the results, only 1,160 women began the study as non-smokers and initiated smoking at some point during follow-up. Based on my findings from the simulation study (chapter 3), having few or no individuals on the rise tends to have positively biased estimates for both half-lives, with more egregious overestimation in the incline half-life parameter. This lack of information explains why the confidence interval width is larger for the incline parameter, which intuitively represents the uncertainty in this parameter estimate.

4.4.2 Strengths

Epidemiologic Strengths

Part of the BWHS cohort's strengths are the prospective data collection and nearly complete information on smoking history that includes various trajectories with interruptions in both smoking and quitting. Lack of control for recidivism, has been a weakness in most other approaches that aimed to estimate the time-to-reduction in the CVD hazard due to smoking.(Rachet et al., 2003; Rosenberg et al.,

1990) Where previous literature has investigated smoking cessation and CVD in case-control study designs(Rosenberg et al., 1990), they may be unable to estimate the true CVD hazard associated with smoking and are subject to observation bias.

Prior literature about smoking cessation and risk of CVD/MI may have misclassified menthol smokers as non-smokers, because of the "high" vs. "low" yield cigarette distinctions, as discussed by Rosenberg et al. (1990). Delnevo et al. (2011) found that smoking cessation rates are lower amongst menthol cigarette users compared to non-menthol smokers. It is possible that studies of smoking cessation did not account for menthol cigarette use, which could bias the estimated CVD hazard and the amount of time for the hazard to return to normal. Thus, a strength of my analysis is it's ability to account for menthol smoking, as a source contributing to the effective exposure measure of smoking in relation to CVD hazard.

Methodologic Strengths

My method provides a single estimate of the HR and does not require complicated restrictions for using ex-smoker data.

Most epidemiologic studies prefer the use of pack-years or cigarette-years for quantifying the dose-response relationship of smoking intensity and duration in association with dichotomous outcomes. However, when Leffondré et al. (2002) investigated various smoking metrics in relation to lung cancer, they demonstrated that having separate measures for duration and intensity provided more interpretable and better fitting model estimates than the combined variable. Additionally, they discovered that using the pack-years variable with never smokers set to 0 tended to overestimate the hazard ratio for current- and past-smokers, compared

to models without never smokers.

In the 2010 Surgeon General's report of tobacco smoke and smoking-attributable diseases, it was noted that the risk of coronary heart disease did not appear to be linearly associated with the quantity smoked.(U.S. Department of Health and Human Services, 2010) In particular, Law & Wald (2003) found a plateauing of the risk after the user reached 25 cigarettes per day, though this threshold was not perfectly flat. Therefore, it stands to reason that the packs/day analyses of the effective exposure of smoking may not be the most appropriate, giving more weight to the binary results. This is also consistent with my results that show a smaller AIC in both samples (full and restricted) for the binary smoking EE than the packs/day EE.

CHAPTER 5

Conclusion

In longitudinal cohort studies, one will often assume a fixed amount of time must pass prior to an exposure turning "on". This can be thought of as a delayed or lagged association between an exposure and some risk of a time-to-event outcome. In modeling the lagged exposure-response relationship with adjustment for other time-varying factors, it may be more clinically relevant to consider the underlying action mechanism of the exposure measure associated with the hazard. Thus, the true association between exposure and response can be denoted by a latent quantity (of exposure) that may be unobservable or difficult to measure/obtain in large follow-up studies.

For example, when a lifetime smoker quits, the cardiovascular disease (CVD) hazard may take some time to return to "normal", and most experts agree that the impact of smoking should subside. A simple exposure metric would immediately set this individual to "unexposed" creating misclassification in the population-based estimate of the CVD hazard associated with smoking. The cumulative number of years smoked metric would also create misclassification, because this would not account for the decline of the hazard when the individual quits.

In addition to lag, some exposure-response associations may not be entirely linear, such that prolonged exposures do not appreciably change the hazard level. For example, Law & Wald (2003) showed that smoking's association with ischemic heart disease is not linear and has some plateau of effect following 20 cigarettes per day consumed. In particular, the CVD hazard associated with smoking 2 packs per day may not be that different from the hazard associated with smoking 1 pack per day, but both differ greatly from the CVD hazard associated with smoking 0.5

packs per day.

To account for plateau in effect, delayed action mechanisms, and time-varying exposure profiles, I have defined the "Effective Exposure" (EE) measure that can be used to estimate exposure-response associations in observational time-to-event data. This latent measure is constructed as a lag-parameterized weighted sum of exposure sequences that plateaus at a maximum exposure level and returns to normal based on the lag parameter(s) used. That is, the lag parameter reflects the rate of accumulation and/or decay to/from the maximum EE level.

In addition to estimating the hazard ratio associated with the maximum EE, I have developed a set of algorithms to estimate the lag parameter of the EE measure. Borrowing from pharmacokinetics, I have shown that the lag parameter can also be defined as the half-life of effect, to allow for interpretation in terms of the time-to-plateau or time-to-null of the effective exposure.

5.1 SUMMARY

In Chapter 2 I derived the equations for EE in the context of one- and two-lag parameters (OPEE and TPEE, respectively). The latter represents situations where the time-to-plateau and time-to-null are not equivalent. Additionally, I showed that both parametric forms could be applied to repeated exposures, which accommodates more types of exposure trajectories seen in "real world" data. For example, individuals that start and stop smoking multiple times can still contribute to population-models of CVD hazard, as their total time-varying EE is based on the sum of EE for each interval of smoking history.

To concurrently estimate the hazard ratio and lag parameter(s), I derived algorithms based on profile likelihood methodology. These fit Cox proportional haz-

ards (CPH) or pooled logistic regression (PLR) models of EE versus time-to-event outcome for iterations of fixed value lag(s). Both the OPEE and TPEE algorithms perform a search/grid-search to identify the [combination of] lags that maximize the log-likelihood.

Chapter 3 explored the estimation performance of my two algorithms for a range of simulation scenarios. I demonstrated that the coverage probability of the hazard ratio parameter was consistently close, if not more conservative, than the nominal 95% confidence interval, for the majority of OPEE and TPEE lag-times considered. Deviations from the 95% coverage occurred more often when the half-life of effect exceeded the study follow-up period, and when the TPEE algorithm was applied to an OPEE-generated simulation scenario.

The simulation study primarily showed that estimation of the half-life and hazard depend on the information content in the data. Specifically, information pertaining to number of subjects at plateau and the amount of time spent at this steady state affected estimation bias and coverage of the hazard ratio parameter. This is thought to be the driving force behind the low coverage of the HR in the OPEE half-life=1,000 days scenario, because most individuals will not have reached steady state EE by the end of follow-up.

The information content required for estimating the half-life or lag parameters depends on the number of subjects transitioning in either direction and the amount of time spent in transition. I showed that the estimation thresholds for the single half-life parameter were at 30 and 450 days, which relates back to 1/30th and 1/2 of the follow-up time. For shorter half-lives, the problems with estimation were primarily based on my inability to approximate a standard error for the estimate, which reflects failure in the normality assumption for true values that are located

close to the bound (0 days).

I used real exposure trajectories from participants in the Black Women's Health Study (BWHS) to show that adjustment for covariates and inclusion of multiple time-varying exposure profiles improved half-life and hazard estimation. In particular, I demonstrated that restricting the BWHS set to quitters, smokers, and never-smokers throughout the study imposed additional bias on estimation of the true lag of effect (as compared to the full BWHS set of exposure profiles).

Following the simulation study, I applied my methods to examine the association between smoking and CVD hazard in the BWHS (chapter 4). As recommended by Rothman (1981) and Abrahamowicz et al. (1996), I used Akaike's Information Criteria (Akaike, 1974) to compare conventional measures of smoking exposure to my EE models. The AIC allowed me to account for model complexity in these comparisons, by penalizing the OPEE and TPEE models for the one and two extra half-life parameters estimated in the process. I showed that the difficult-to-interpret metric combination of pack-years plus current smoking fit the best out of all the conventional measures (AIC=35067.8), though both OPEE and TPEE models produced lower/better AIC (35061.3 and 35062, respectively).

Part of the appeal behind my method is that I have estimated the prolonged smoker's hazard of CVD along with the time required to reduce the hazard by 50% for any individual. This hazard ratio estimate was consistent for both the full and restricted samples, such that, after adjusting for confounders and CVD risk factors, the lifetime smoker's hazard of CVD is 2.5 times that of the never smoking counterpart. This hazard ratio may appear to underestimate the excess risk reported by previous literature(Kawachi et al., 1994; Rosenberg et al., 1990), however, the point of these analyses was to demonstrate the methods and, to quote

Dr. Cupples et al. (1988), "[these illustrations] are not to be construed as providing substantive medical conclusions."

5.2 PRIOR LITERATURE

In chapter 1, I discussed some of the previous approaches for dealing with exposures that have a delayed action mechanism.

Similar to my conclusions, Abrahamowicz et al. (2006) showed that a cumulative weighting metric provided better fitting models than traditional measures of time-varying dose and exposure duration in an application to benzodiazepines and risk of injuries from falling. While the authors borrowed weights from the known pharmacokinetic half-life of each drug, the conclusions regarding time-to-reduction in risk varied by the type of drug considered. This could be due to differing biologic mechanisms, as argued by the authors, while I would agree more with the conclusion regarding lack of empirical evidence due to the small number of events by each benzodiazepine drug in the study. I also showed that the conventional models performed better in the small sample size simulations.

The models described by Richardson (2009) assume protracted exposures depend on interval start and end times for multiple exposure events. He defines the cumulative effective exposure as the exposure accrual over a given period, which differs from what I propose in the shape of the function over time. Specifically, Richardson's models assume either a bilinear or log-normal latency function, which do not account for a plateauing of the effect or return to null effect over time.

Langholz et al. (1999) explored several models including one close to my own, with an exponential decay following exposure discontinuation. This "effective dose" structure differs from the EE model as it assumes a linear rise in effect. Simi-

lar to my conclusions, the authors noted the perks of the half-life interpretation for the decline in risk and that precision in the latency parameter estimate depended on having adequate variation in exposure profiles (i.e. information!). Richardson et al. (2011) noted that a decreased variability within-subjects for the follow-up period could result in a flattening of the likelihood with respect to the lag-times considered. This, in turn, leads to biased estimation of the association measures, and potentially narrow confidence intervals.

The more widely-accepted solution to dealing with lagged effects seems to settle on the cubic B–splines approach.(Rachet et al., 2003; Abrahamowicz et al., 2012) While this method is flexible and can be applied to various underlying parametric models, it is sensitive to the number of knots selected and their locations, with instability in the tails. They also require a larger number of degrees of freedom than my method as the polynomial function, alone, takes 4 degrees of freedom.(Abrahamowicz et al., 2012)

Taking the binary OPEE model's estimates of a half-life of 7.1 years and maximum hazard ratio of 2.5: for a 30-year smoker's hazard ratio to decrease by one-third would take roughly 2.9 years following cessation. Meanwhile, Rachet et al. (2003) estimated that the ex-smoker's hazard is 2.7 times the non-smoker with a mean lag of 3.3 years (stdev=0.97).

My analysis of smoking and time-to-first CVD has several strengths over the proposed method and application developed by Rachet et al. For starters, my model included time-varying confounders while Rachet considered only baseline covariates. Given that smoking habits tend to vary over time along with other behavioral factors, such as alcohol use and exercise, my approach likely had better adjustment for the correlation between smoking and other factors of CVD hazard.

Additionally, Rachet included only smoking cessators in the analysis, while my method allows for complex trajectories. This also means that conclusions made using my method are more generalizable to situations where individuals may not quit permanently, or where exposure trajectories are non-monotonic.

5.3 LIMITATIONS

One of the limitations of my approach is the fact that the lag-time may not fall within a study period or window of follow-up that permits estimation. As described in chapter 3 and above, the OPEE and TPEE algorithms depend on adequate information regarding time in transition in order to have unbiased estimation of the half-life of effect. Thus, for EE that decline slowly, it is possible that individuals may die before reaching the level of no effect, which can make estimation difficult. For these types of situations, I may be able to extend my approach to account for competing events, such as death. In general, it would be useful to understand the impact of competing events and other censoring mechanisms, as these are relevant in most time-to-event analyses.

While I was able to account for the accumulated CVD hazard of smoking's exposure in my models, as well as, the time since quitting, I am limited by the fact that I did not explicitly consider a pack-years equivalent in my simulations in chapter 3. It would be useful to understand how this conventional measure compares to the OPEE and TPEE algorithm estimates for the hazard ratio. Particularly, as the OPEE and TPEE model remove the need for classifying individuals by current status, it would be useful to determine whether the maximum hazard plateau matches that of the combined hazard estimate for currently exposed individuals who have been exposed for an extended amount of time.

5.4 FUTURE WORK

5.4.1 Theoretical Next Steps

Floor of Effect

My methods, so far, have all assumed that the hazard will return to "normal" given sufficient passage of time, as determined to be 4-5 half-lives following cessation. In some cases, it may not be appropriate to assume that individuals will return completely to the unexposed level. For example, while I discussed the slow decay in EE of lead following treatment by chelation, a more appropriate model may be one in which the effect of lead is irreversible. This would imply that the level to which the individual can return is not 0 or null, and thus would require an additional estimate for the location or hazard associated with the "floor".

Confounders

I briefly discussed the benefits of adding covariates to the models in chapter 3, in support of my information hypothesis. I also discussed the need to account for potential effect modification and unmeasured confounders in chapter 4. For example, smoking has been shown to increase the risk of stroke for individuals that use oral contraceptives. This type of effect modification could also extend to the use of female hormones for menopause, such that the plateau of smoking's EE varies by hormone use.

Future work is needed to better understand the effect of unmeasured confounding on the estimation of the hazard and lag parameters. This should also be extended to situations where the confounder or covariate may behave as an effect modifier, which could be reflected either in the estimate of the plateau and/or in

the estimated time required to achieve plateau (i.e. lag).

Multiple Lagged Exposures

The methods developed for this dissertation have all focused on estimating the lag and hazard associated with a single exposure. In prospective cohort data, like BWHS or the Framingham Heart Study, the most common approach is to treat each questionnaire cycle as a mini follow-up study. When looking at multiple risk factors for a time-to-event outcome, it is possible to impose a lag on all or some of the variables, by simply using data from prior cycles.

My approach has shown that it is possible to avoid the a priori specification, by estimating the lag associated with the effective exposure and the event of interest. However, my methods do not yet allow for more than one lagged exposure. Thus, a future direction might include development of an [even more complex] algorithm that can iterate the lags of multiple effective exposures simultaneously.

Hypothesis Testing

The majority of this dissertation has focused on the estimation process with no allusion to statistical inference. While my methods are able to estimate HR as low as 1.2, I have not provided a formal approach for testing that the effect is significant. For now, it is possible to use the normally approximated lag-adjusted standard errors to determine whether the hazard is non-zero.

Additionally, the normally approximated standard errors for the lag may provide some insight regarding the magnitude of the delay in effect. For example, if the lag parameter's confidence interval includes 0, then it is possible that using the current metric for exposure may suffice. However, before applying this logic, it

would be important to determine whether the normality assumption is correct, as this tends to be violated near the lower values of the half-life.

Another interesting question that could be answered through hypothesis testing is whether the incline and decline half-life are the same. One way to test this would be to bootstrap the sample and fit both OPEE and TPEE algorithms. The AICs could be compared for each bootstrap pair of results, with an empirical probability assigned to selecting one model over another. The alternative bootstrapping approach would fit just TPEE algorithms to the samples to build a bootstrapped confidence interval for the mean difference in the lag parameters.

5.4.2 Dissemination

Prior to Dissemination

Add a module for the TPEE model to start from the OPEE model's estimate.

Standard errors based on asymptotic theory, though they could also be computed through bootstrapping. The former is computationally efficient, while the latter should work best in the presence of "strange" likelihood surfaces. Future research should compare the estimated confidence intervals for bootstrapped vs. information-derived estimates of both hazard and lag parameters.

Optimization of Functions

One of the primary drawbacks of my current method is the computational efficiency of the iterative algorithmic fitting process. While the "survival" package offers some variants for fitting the time-dependent covariates, the underlying computations in that module utilize C, which is a faster and more efficient programming language. Ideally, I would like to develop the fundamental modules for my

calculations in C, as well.

I could also compare to and consider utilizing functions from the "optim" package in R. This could be particularly useful when I try to approximate the standard errors under failed normality.

R Package Development

I plan to prepare an R package, such that other researchers may use the methods I have developed for my dissertation. Some technical aspects of the programs have been included in the text and appendices, though additional modifications and revisions may be implemented at later points. In particular, I developed modules for data processing that facilitate the use of my estimation algorithms.

Public Health Articles

I plan to summarize and write manuscripts corresponding to the methods and results presented in chapters 3 and 4.

5.4.3 Applications to Public Health Questions

Smoking and CVD in the Million Veteran's Project

One analysis that is in the works, though has not been included in this dissertation, is to look at smoking and CVD using the Veteran's Affairs Million Veteran Project administrative data. The structure of this dataset poses some interesting challenges, and the analytic results would serve as external validation to the analysis described in chapter 4.

In particular, for this analysis, there is no data on individuals who return to smoking, however, there is a fine gradient of information available on covariates

and event timing.

Statins and Risk of Fracture

An interesting clinical question is raised in the conflicting evidence of a protective association between statin use and risk of fracture.(Toh & Hernandez-Diaz, 2007) Researchers have indicated uncertainty in the mechanism of action explaining this association.(Scranton et al., 2005) In this case, my methodology could provide insight regarding the half-life of and whether there is a true association. Specifically, half-life values approaching zero or infinity would indicate that the biologic mechanism behind the association may not be valid and other explanations, such as confounding by indication, could be driving the statistically significant associations observed.

Obesity and Total Knee Replacement

As mentioned in chapter 1, there are many examples of lagged effects that could be interesting to analyze using the proposed approach. For instance, I could look at the trajectories of body mass index (BMI) over time and risk of total knee replacement. There is a known deleterious effect of increased weight on knee osteoarthritis progression, due to a number of hypothesized mechanisms, such as loading and inflammation.(Coggon et al., 2001; Felson et al., 1988) Given recommendations to lose weight, an individual may lower the impact of their weight. However, it us unknown whether the damage over the earlier period of time is reversible, and thus, whether there is an estimate-able lag related to the effective exposure of weight. The TPEE model is particularly interesting to apply to this type of data, as it would allow for the modeling of separate lag-times of obesity's

effect on knee osteoarthritis and later replacement surgery.

5.5 FINAL THOUGHTS

There are several points to consider before using my methodology. The first is that the EE approach should not be the first model considered. There should be substantial rationale for implementing the structure I've described – Does the hazard or risk plateau after sufficient time of exposure? Is there a reason to believe that the effect of exposure is lagged? Can I reasonably identify a clinically-relevant range of lag-times that might exist in my data?

The latter question is meant to focus on the probable biologic mechanisms that are being modeled, i.e. what range of lags would make sense clinically? In addition to this question, the researcher should consider whether the hypothesized lag-time falls within the study length. Particularly, I showed that it is not possible to precisely estimate the lag, when the study is too short. This is especially important when considering use of this method in clinical trials data, as these may be too short to account for the lag in effect. Clinical trials, especially phase I and II, tend to have small numbers of participants, which I showed (small sample size) to be a potential limitation in chapter 3. Because of the small sample sizes and lack of variability in exposure trajectories [point of the design is to control the exposure], it may be best to avoid use of my method in clinical trials.

Secondly, when lagged effects exist, but the sample is small or the hazard ratio is close to null, it is possible to revert back to the conventional measures of exposure to approximate the hazard. These may be slightly biased, however, the confidence intervals produced should have decent coverage of the true measure of association. If the researcher still feels the need to account for lag when the

conventional measure HR is smaller than 1.2, I would recommend starting with several fixed lags. The AIC can then be compared across the fixed-lag models and conventional measures to determine which model is most predictive of the true exposure-response relationship.

Finally, when the conditions for use have been met (variation in exposure trajectories, HR>1.2, lag-time<study length, large enough sample size), the OPEE and TPEE algorithms can be applied to properly formatted/arranged data to estimate the lag and hazard. I recommend that the analyst try more than one initialization value, to ensure that the estimation does not get stuck on a likelihood ridge. If the algorithms continue to produce inconclusive or inconsistent estimates for the lag and HR parameters, for multiple initial guesses, then one could fit a grid of points across all plausible combinations of incline and decline lag. This would be identify the full profile likelihood maximum and PLL confidence bounds. Alternately, should the algorithms converge, but approximation of the standard errors "fail", it is possible to use the PLL confidence bounds to inform the desired uncertainty measure for the parameters estimated.

GLOSSARY OF TERMS

Effective Exposure: An underlying/latent curve reflecting the etiologically-relevant exposure that has a lagged association with a time-to-event outcome. The curve increases like an exponential cumulative density function with a lag parameter λ, and decreases mono-exponentially from the level that corresponds to the change-point amount. In the OPEE context, the lag parameters for incline and decline are assumed to be the same, while for TPEE the incline and decline λ values are different. Can also be understood as the relative proportion of the exposure that is "actively" associated with the event of interest.

Dose: The unit of effective exposure that corresponds to the prolonged exposure's hazard. In the binary context, the Dose takes the value of 1 for exposed and 0 for unexposed. For dosing-based effective exposures, the lag parameters reflect the time-to-plateau for a one unit dose. An example of the latter is the effective exposure of smoking 1 pack/day in relation to risk of cardiovascular disease, where the plateau of a 2 pack/day smoker is at 2-fold the hazard level of the never smoker.

λ ***- Lag Parameter***: The rate of growth/decay for an effect. May be parameterized as the natural log of 2 divided by the half-life of the effect.

Start and Stop Times: The individual initialization and discontinuation of the given dose for an individual. These times may vary across subjects and are assigned relative to the overall study time period.

Steady State: The amount of the drug entering and leaving the body is in equilibrium, such that the concentration in the single compartment does not change over time.

One-Compartment Model: An equation, one compartment pharmacokinetic model, for the concentration of a drug that depends on input and output rates in a single location, typically the plasma/blood. The intravenous (IV) model can be simplified to a time-dependent risk curve that is parameterized by the elimination rate of the specific component.

Event: A binary outcome that follows $Bernoulli(p_{EE})$

p_{EE}: The probability of an event, given the Effective Exposure

Trajectory: The time-varying pattern of exposure for a specific individual. Monotonic trajectories imply that an individual's risk profile transition in a single direction or is static, while a "Multi-Trajectory" implies that the subject's effective exposure has multiple change-points within the follow-up period.

APPENDIX A

Algorithms

OPEE FLOWCHARTS

Figure A.1: Big Picture OPEE Flowchart

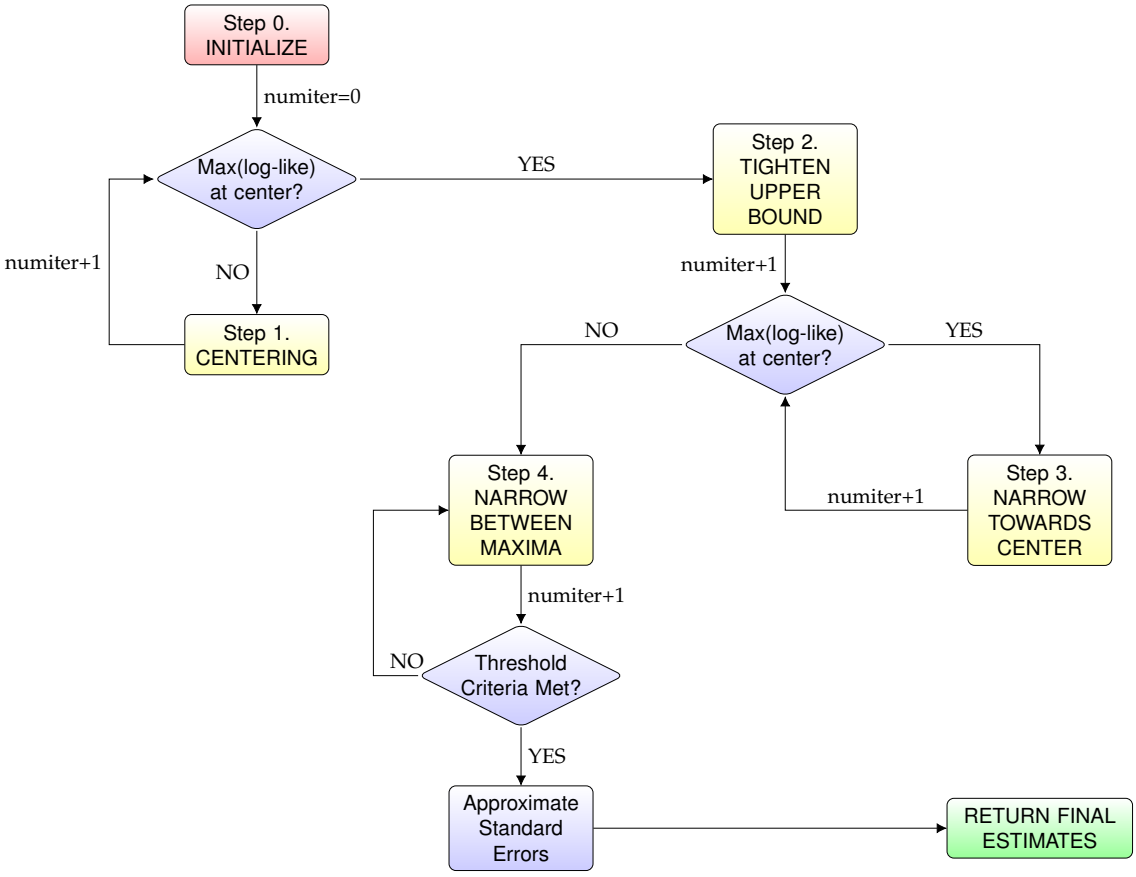


Figure A.2: OPEE Flowchart Steps 0 and 1

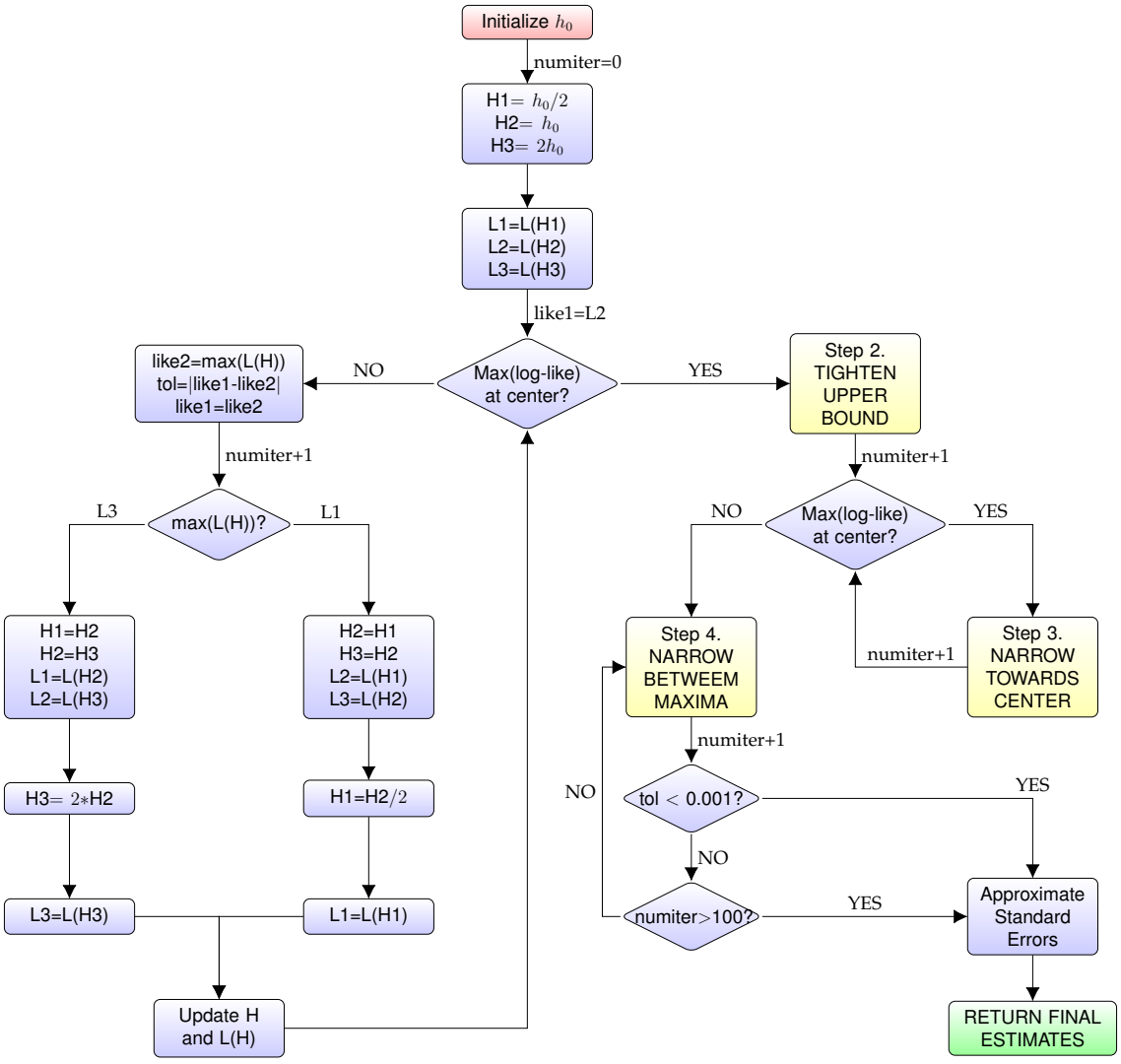


Figure A.3: OPEE Flowchart Steps 2 and 3

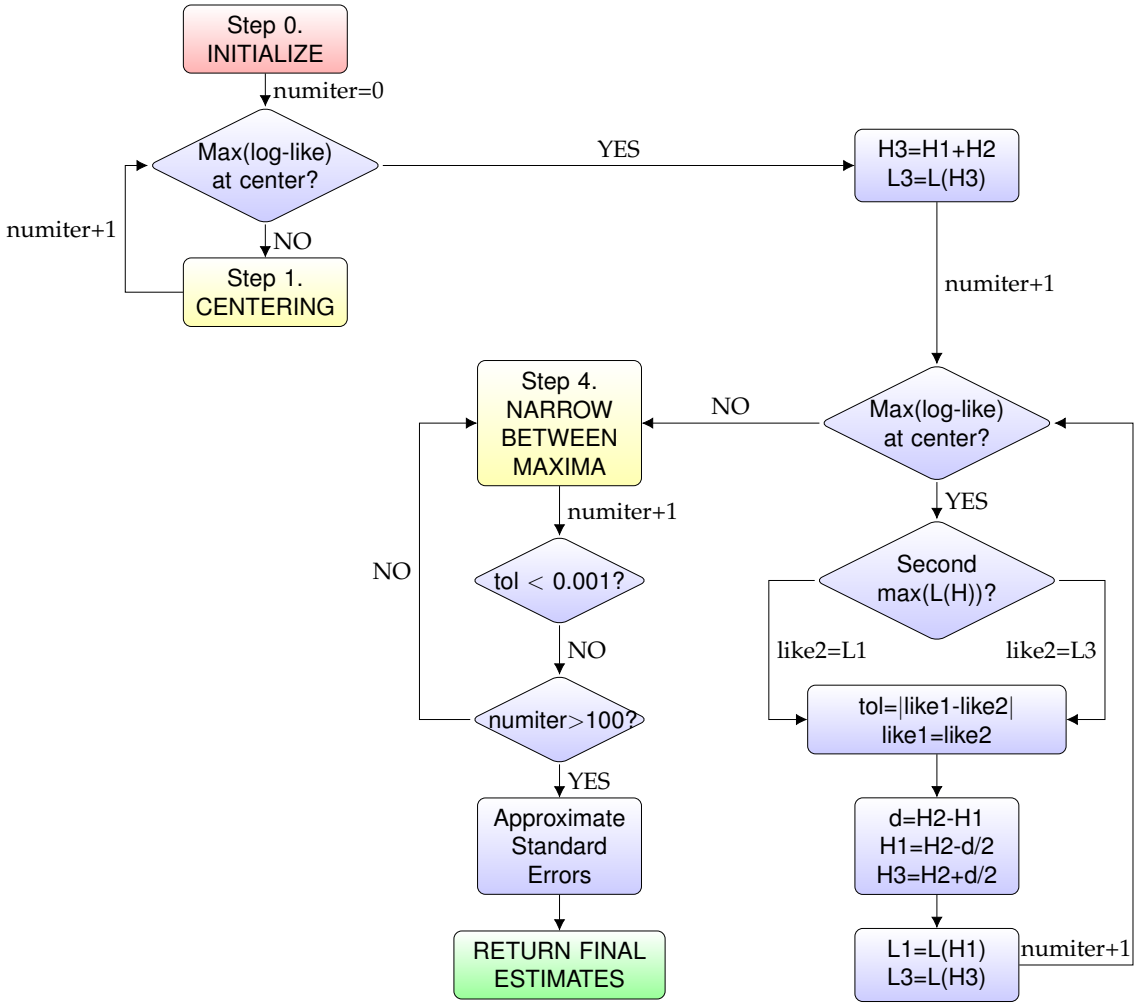
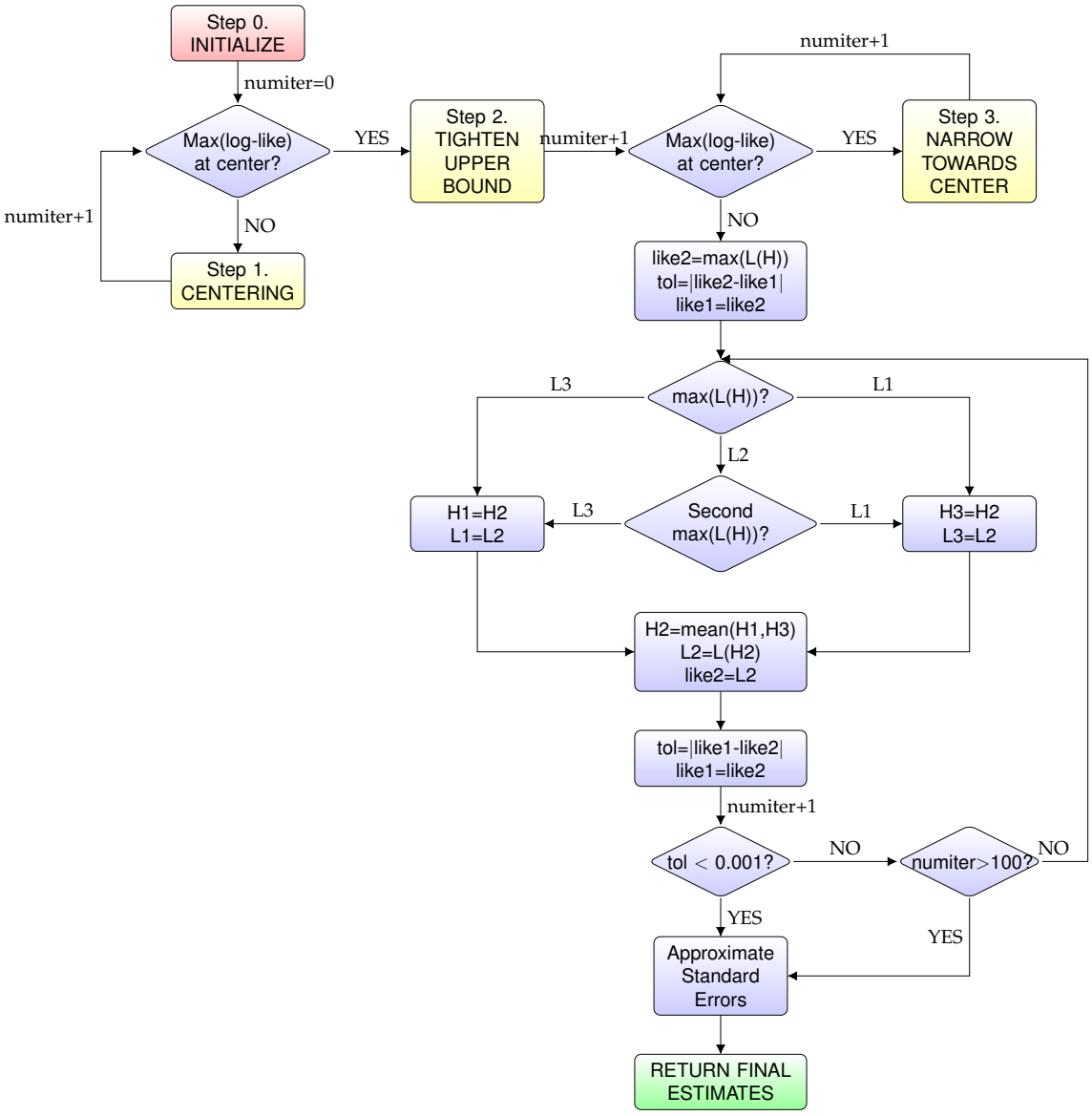


Figure A.4: OPEE Flowchart Step 4



TPEE FLOWCHARTS

Figure A.5: Big-Picture TPEE Flowchart

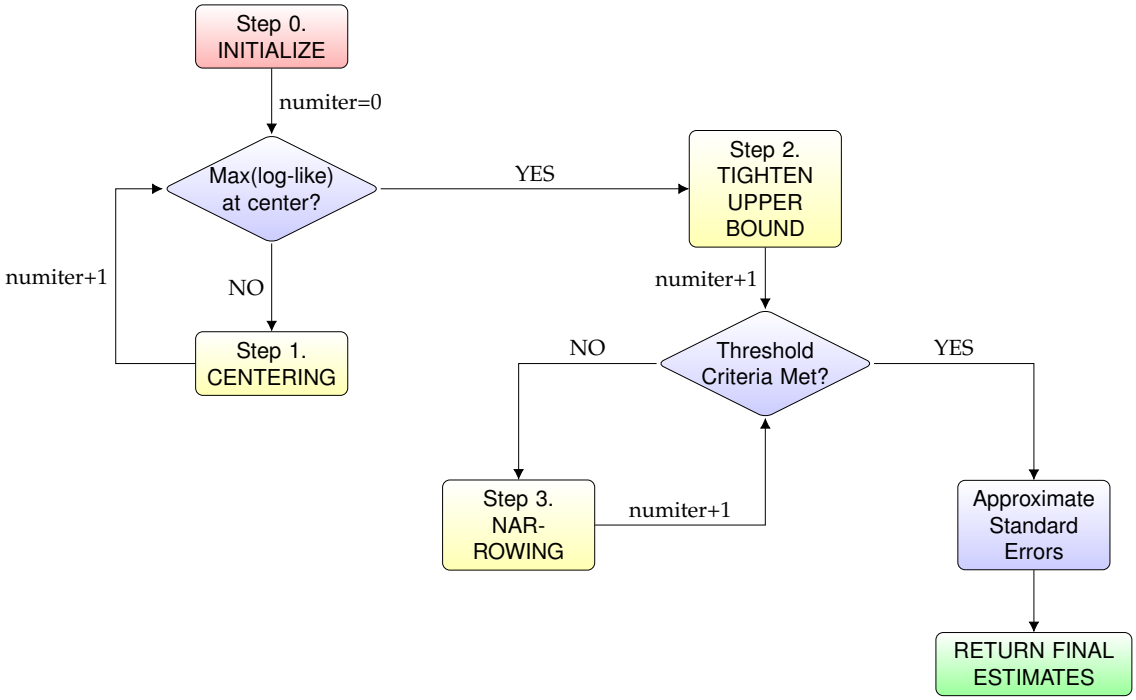
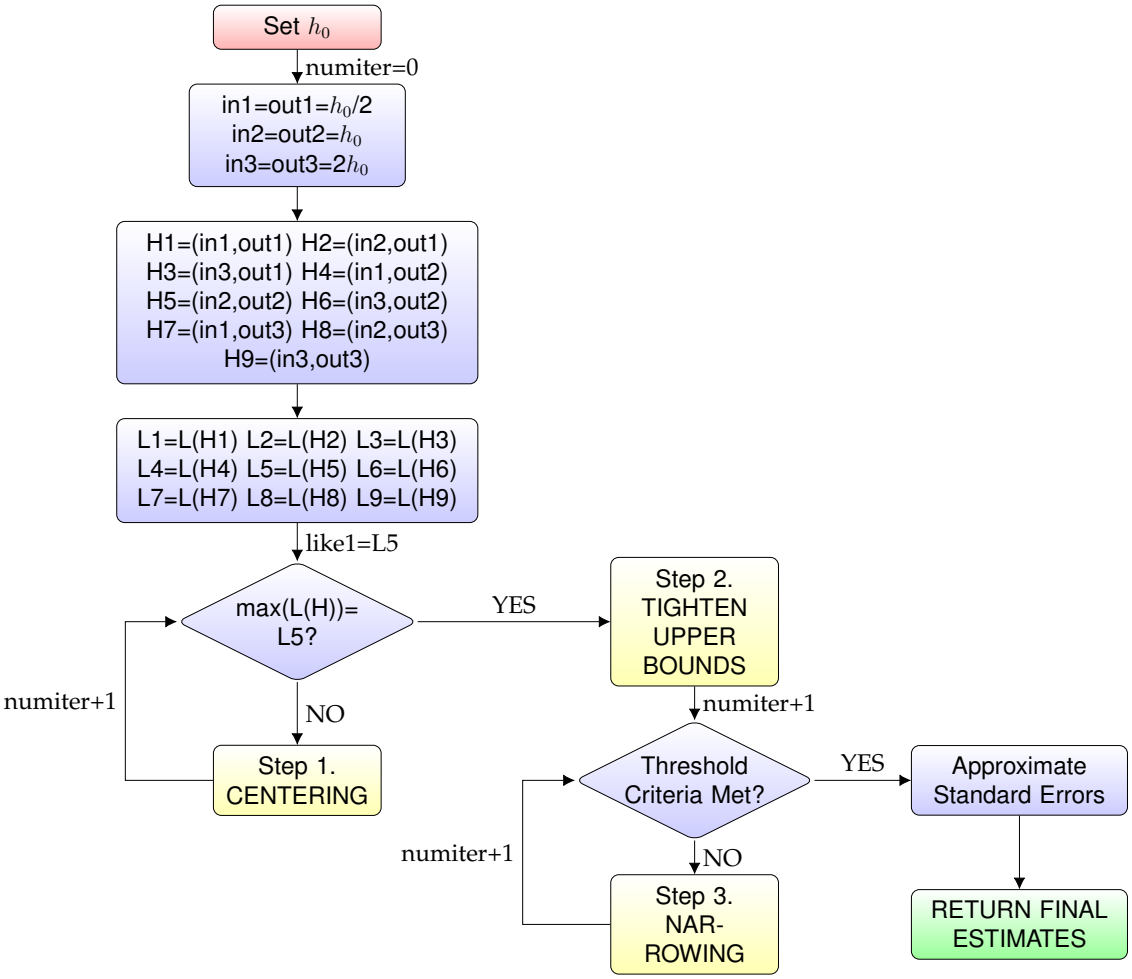


Figure A.6: TPEE Flowchart Step 0



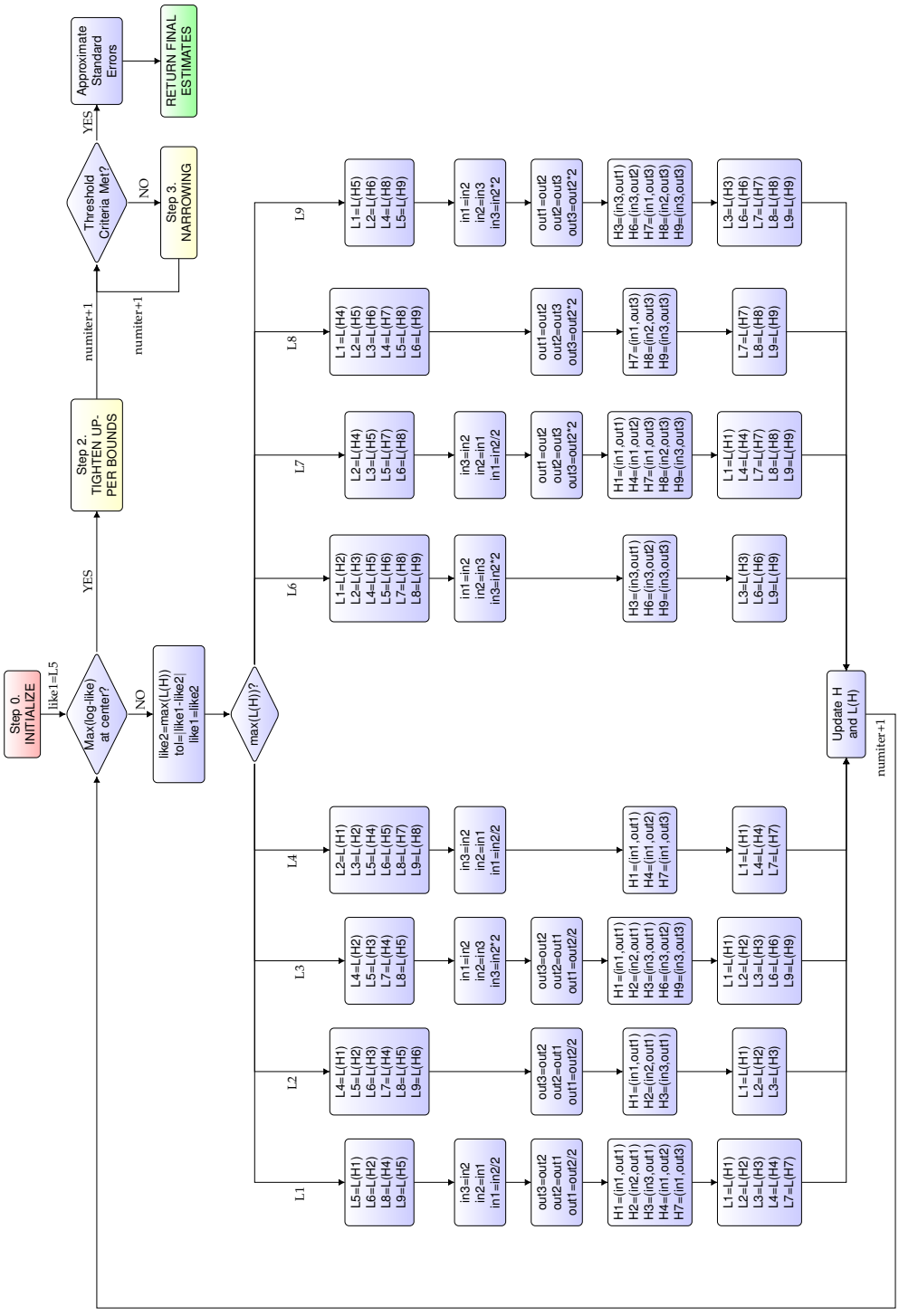


Figure A.7: TPEE Flowchart Step 1

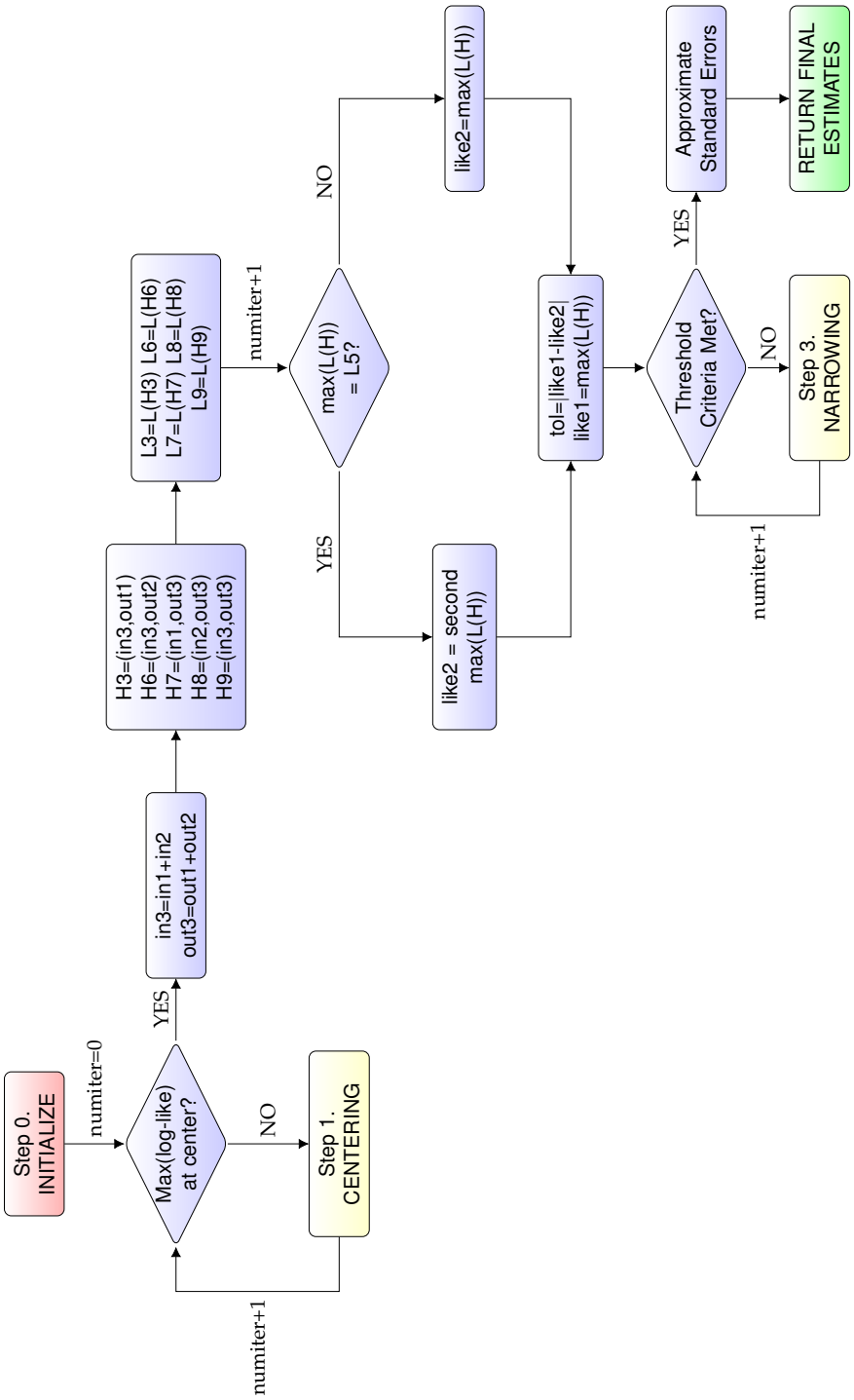


Figure A.8: TPE Flowchart Step 2

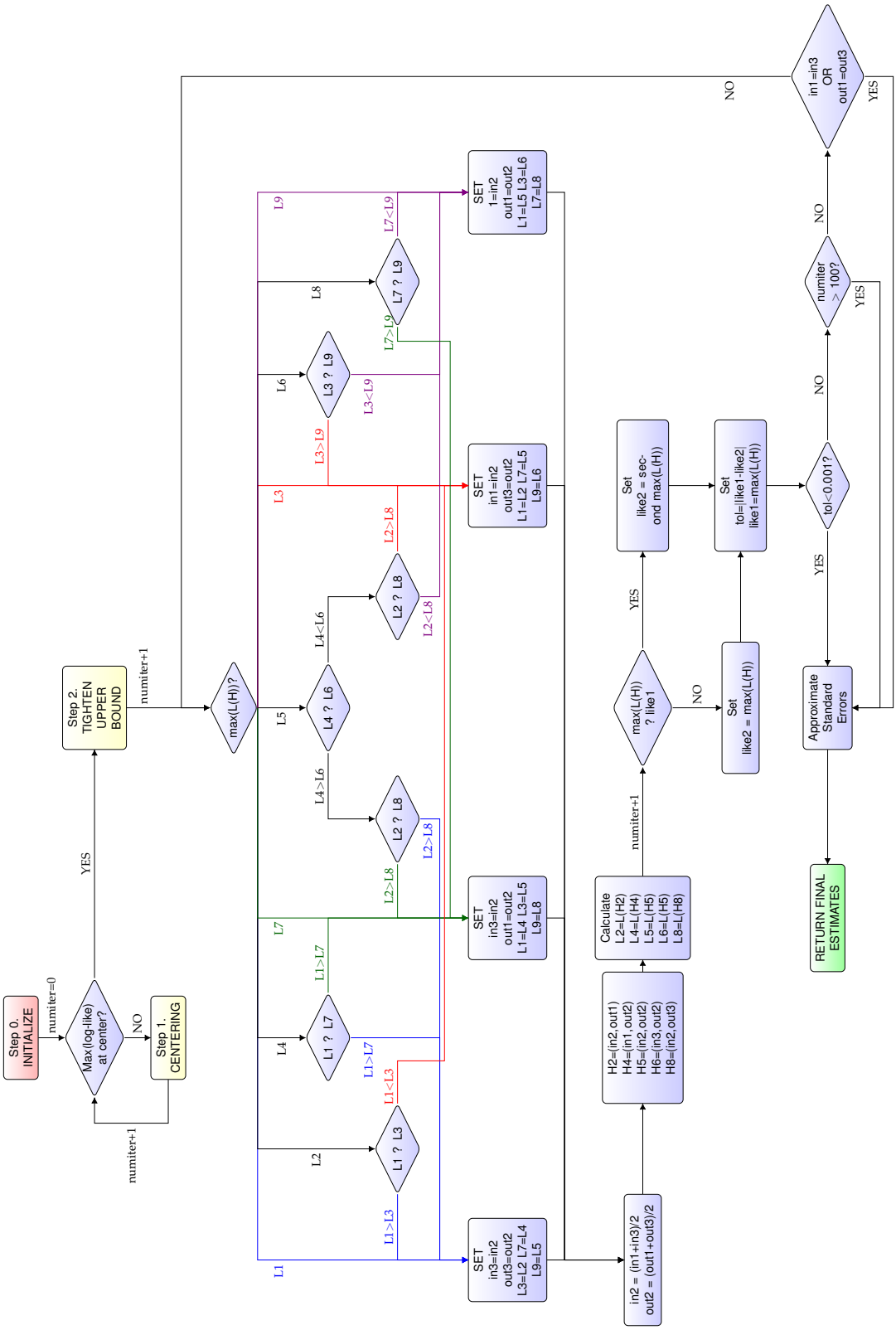


Figure A.9: TPEE Flowchart Step 3

APPENDIX B

Selected Code Documentation

This appendix provides minimal necessary documentation for the functions described in the Dissertation text. Full code and documentation can be found at https://github.com/hgerlovin/Lagest.

---

makeDVecs Split the multiple exposure periods into vectors of dose and time since start and stop

---

Description

Reads in values for treatment regimen doses, start and stop times to create analysis-friendly dataset.

Usage

```
makeDVecs(Cp.vec=c(1), ts.vec=c(0), tf.vec=c(900), intlen=1,
studyt=NULL, struct=0)
```

Arguments

`Cp.vec`	Vector of doses for the regimens. Default assumes the binary exposure plateau and that there is a single regimen of exposure. To have multiple exposures, include the same number of vector components in `Cp.vec`, `ts.vec`, and `tf.vec`.
`ts.vec`	Vector of start times for the regimens. Default assumes the exposure was started at time 0.

`tf.vec` Vector of end times for the regimens. Default assumes the exposure continues through time=900. When `studyt` is not specified, the last specified end-time (last regimen) is used as the total study time.

`intlen` Increment time to use. Default is 1 time unit.

`studyt` Total study follow-up time. Default is `NULL` and will pull the last regimen stop time.

`struct` Structure indicator. If turned on (1), then additional regimen is added for time following discontinuation. Default is off (0), assuming that the total number of regimens is fixed and does not need additional follow-up.

Value

Outputs a `dataframe` with time incremented rows - columns: `time, currD, everD`, and three columns per regimen for 1 to X total exposure events.

`Dose1 – DoseX` Columns indicating the overall doses for each regimen. Repeated throughout for computational ease.

`tStart1 – tStartX` Columns indicating the time since starting the specific regimen – depends on the point in the trajectory. i.e. Takes a value of 0 for times prior to initiation and increments parallel with time following initiation.

`tEnd1 – tEndX` Columns indicating the time since discontinuing the specific regimen – depends on the point in the trajectory. i.e. Takes a value of 0 for times prior to start of regimen and while regimen is "on". Increments parallel to time following discontinuation.

`time`	Column for the study time at the observation.
`currD`	Column with value for the current regimen dose for the subject-time-specific observation.
`everD`	Column indicating whether any exposure has occurred as of (prior to and including) the subject-time-specific observation.

Requires functions

`lapply, unlist, colSums`

C1fn.h Calculate effective exposure for specific subject-time

---

Description

Calculates a relative Effective Exposure component for regimen X for a single time- and subject-specific observation using the one-parameter half-life effective exposure formulation.

Usage

```
C1fn.h(d, s, e, h)
```

Arguments

`d`	dose/concentration value (`DoseX`)
`s`	time since start (`tStartX`)
`e`	time since end (`tEndX`)

`h`	half-life parameter

Value

Returns a single computed value.

C1fun.h Calculating total effective exposure for an entire dataset

---

Description

Combines all individual effective exposure components for a subject at each time to calculate the full effective exposure. Used in simulations and analysis.

Usage

```
C1fun.h(thalf=NULL, dat)
```

Arguments

`thalf`	Single-parameter half-life to use in computation
`dat`	`dataframe` with three columns per exposure event/regimen following the naming conventions for X total regimens: `Dose1 - DoseX, tStart1 - tStartX, tEnd1 - tEndX`.

Value

Outputs a numeric vector of values for concentration that is equal in length to the number of rows/observations in the `dataframe dat`.

Requires functions

`lapply, C1fn.h, Reduce, replace`

Notes

Data should be in long format with one observation per subject per time point - which can be created from the function `makeDVecs`.

---

solve.time Time to reduction in excess hazard

Description

Calculates the time to % reduction in the excess hazard following discontinuation of an exposure.

Usage

```
solve.time(beta, D, half.in, half.out, time.in, reduct)
```

Arguments

`beta`	Effect size coefficient (beta) from the model fitting results.
`D`	Exposure level prior to discontinuation.
`half.in`	Incline half-life parameter if the resulting model selected was the two-parameter effective exposure model. For the one-parameter effective exposure model, use the same value for the `half.in` and `half.out` inputs.

`half.out` Decline half-life parameter if the resulting model selected was the two-parameter effective exposure model. See `half.in` for how to handle OPEE framework.

`time.in` Amount of time exposed prior to discontinuation. Value should exceed 1.

`reduct` Desired excess hazard reduction. For example, to determine the time to 50% reduction in the hazard ratio for a specific individual use `reduct=0.5`.

Value

Returns values for the estimated starting (time=0 at discontinuation) hazard ratio (HR), starting risk or log(HR), ending HR and log(HR), time required to return to the reduced HR, and the relative proportion of reduction from start to end time on both HR and log(HR) scales.

`start.risk` The starting log(HR) after `time.in` units-time of exposure prior to discontinuation.

`start.relrisk` The starting HR after `time.in` units-time of exposure prior to discontinuation.

`time.needed` The calculated time needed to reduce the excess hazard to `reduct`.

`end.risk` The ending log(HR) for the individual following the `time.needed` units-time.

`end.relrisk` The ending HR for the individual following the `time.needed` units-time.

`relrisk.red` Proportion of reduction in excess hazard. Note that this returns the input `reduct` value.

`risk.red`	Proportion of reduction in the log(HR) scale.
`startEE`	Effective Exposure starting value based on the dosing scale with maximum HR at the value of `D` input parameter.
`endEE`	Effective Exposure ending value that corresponds to the reduced excess hazard. Similarly, this is relative to the value of `D` input parameter.

Code for Interpretation Paradigm In Chapter 2

Calculate the time to 50% reduction in the HR for a 2 packs/day smoker of 30-years. Final model being used comes from results in Chapter 4 (OPEE Packs/Day Dosing in Full BWHS Sample).

```
solve.time(beta=log(2.63), D=2, half.in=5.85, half.out=5.85,
time.in=30, reduct=0.5)
```

Calculate the time to 50% reduction in the HR for a 1 pack/day smoker of 30-years. Final model being used comes from results in Chapter 4 (OPEE Packs/Day Dosing in Full BWHS Sample).

```
solve.time(beta=log(2.63), D=1, half.in=5.85, half.out=5.85,
time.in=30, reduct=0.5)
```

APPENDIX C

Simulation Result Tables

BC1 Base case scenario

N1k Base case scenario with 1,000 total subjects

N10k Base case scenario with 10,000 total subjects

N100k Base case scenario with 100,000 total subjects

FourG Variation on the Base case scenario adding 10,000 subjects with "on" binary exposure starting at the same time as the "down" group (900 days prior to study start date), who remain "on" throughout the study follow-up or until censoring at event.

NoUp Variation on the FourG/Base case scenario that removes the "up" group. Remaining sample is 30,000 subjects split evenly between controls, "down" and "on" groups.

DoseMods Variation on the Base case scenario where half of the "down" and "up" groups are assigned a 2-fold risk of event (or dose=2).

HR1 Variation on the base case scenario where there is no change in risk of event due to the exposure. The input half-life is still 90 days, even though this is not clinically relevant.

HR10p Variation on the base case scenario where there is a minimal 10% increase in risk of event due to the exposure.

HR20p Variation on the base case scenario where there is a minimal 20% increase in risk of event due to the exposure.

HR5 Variation on the base case scenario where the steady state hazard due to exposure is 5-fold compared to unexposed.

Half1 Variation on the base case scenario with the half-life set to 1 day.

Half10 Variation on the base case scenario with the half-life set to 10 days.

Half10.4G Variation on Half10 with the additional "on" group.

Half30.4G Variation on the FourG scenario with the half-life set to 30 days.

Half450.4G Variation on the FourG scenario where the half-life of the effect is set to 450 days, or half of the follow-up time.

Half1k Variation on the base case scenario with the half-life set to 1,000 days.

Half1k4G Variation on the FourG scenario where the half-life of the effect is set to 1,000 days. This analysis differs from the "half1k" by the addition of the "on" group of subjects.

MVmono The multivariate monotonic trajectories set of simulations performed on the restricted set of subjects in the Black Women's Health Study data. Individuals included here were set to a 3-fold binary-smoking steady state effect with an assumed 3-year single-half-life effective exposure.

MVmulti Variation on the "MVmono" set of simulations that includes all subjects and all smoking trajectories from the BWHS dataset. Input half-life, risk ratio, and underlying models were the same as in MVmono.

Out10 Two parameter scenario with incline = 90 days and decline = 10 days.

Out10.4G Variation on the "Out10" scenario with additional subjects in the fourth group ("on" group)

Out30.4G Two parameter scenario with incline = 90 days and decline = 30 days, using the four risk profiles.

Out450.4G Two parameter scenario with incline = 90 days and decline = 450 days, using the four risk profiles.

Out1k Two-parameter scenario with incline = 90 days and decline = 1,000 days

Out1k4G Variation on the "Out1k" scenario with additional subjects in the fourth group ("on" group)

Table C.2: Comparing Mean and Median Estimates of the Risk Ratio across scenarios and standard metrics of exposure. Based on 1,000 simulated samples for each scenario.

Simulation	Input	PvsN	CvsN		Current		Ever		TrueEE
Scenario	HR	Mean	Mean	% Bias	Mean	% Bias	Mean	% Bias	Mean
BC1	1.5	1.07	1.42	-5.48	1.37	-8.58	1.24	-17.21	1.50
N1k	1.5	1.10	1.46	-2.48	1.40	-6.94	1.28	-14.79	1.54
N10k	1.5	1.07	1.42	-5.18	1.37	-8.36	1.24	-17.02	1.51
N100k	1.5	1.07	1.42	-5.56	1.37	-8.67	1.24	-17.23	1.50
FourG	1.5	1.07	1.46	-2.75	1.41	-5.94	1.33	-11.53	1.50
NoUp	1.5	1.07	1.50	0.16	1.45	-3.17	1.28	-14.42	1.50
DoseMods	1.5	1.11	1.71	14.28	1.38	-7.99	1.41	-6.08	1.50
HR1	1	1.00	1.00	0.07	1.00	0.05	1.00	0.07	1.00
HR10p	1.1	1.02	1.09	-1.31	1.08	-2.05	1.05	-4.51	1.10
HR20p	1.2	1.03	1.17	-2.52	1.15	-3.91	1.10	-8.39	1.20
HR5	5	1.39	4.10	-18.10	3.43	-31.32	2.68	-46.49	5.00
Half1	1.5	1.00	1.50	0.06	1.50	0.01	1.25	-16.80	1.50
Half10	1.5	1.01	1.49	-0.52	1.49	-0.91	1.25	-16.85	1.50
Half10.4G	1.5	1.01	1.50	-0.26	1.49	-0.65	1.33	-11.28	1.50
Half30.4G	1.5	1.02	1.49	-0.89	1.47	-2.01	1.33	-11.34	1.50
Half450.4G	1.5	1.18	1.31	-12.50	1.20	-19.87	1.27	-15.34	1.50
Half1k	1.5	1.15	1.11	-25.99	1.03	-31.21	1.13	-24.58	1.51
Half1k4G	1.5	1.15	1.19	-20.47	1.11	-26.08	1.18	-21.37	1.50
MVmono	3	1.11	2.96	-1.47	2.86	-4.82	1.51	-49.69	2.98
MVmulti	3	1.18	2.75	-8.18	2.59	-13.52	1.63	-45.82	2.97
Out10	1.5	1.01	1.42	-5.46	1.41	-5.82	1.21	-19.25	1.50
Out10.4G	1.5	1.01	1.46	-2.73	1.45	-3.11	1.31	-12.90	1.50
Out30.4G	1.5	1.02	1.46	-2.74	1.44	-3.84	1.31	-12.55	1.50
Out450.4G	1.5	1.25	1.46	-2.76	1.29	-13.67	1.39	-7.31	1.50
Out1k	1.5	1.36	1.42	-5.48	1.20	-19.71	1.39	-7.47	1.50
Out1k4G	1.5	1.36	1.46	-2.75	1.24	-17.40	1.43	-4.99	1.50

Input HR: The hazard ratio used in data simulation for the effective exposure at steady-state risk; PvsN: Estimate for Past Exposure compared to Never exposed in categorical model; CvsN: Estimate for Current Exposure compared to Never exposed in categorical model; Current: Estimate for Current vs. Not Current exposed risk ratio; Ever: Estimate for Ever vs. Never exposed risk ratio; TrueEE: The simulated exposure's "true" risk ratio if it were measured

Table C.3: Hazard Ratio Estimates across Simulation Scenarios and Effective Exposure Algorithms. Based on 1,000 simulated samples for each scenario.

		OPEE			Non-Failed OPEE		TPEE			Non-Failed TPEE	
	Input	Mean	% Bias	# Failed	Mean	% Bias	Mean	% Bias	# Failed	Mean	% Bias
BC1	1.5	1.51	0.46	1	1.51	0.46	1.62	7.86	74	1.60	7.86
N1k	1.5	2.4e+66	1.6e+68	5	2.4e+66	1.6e+68	Inf	Inf	554	3.4e+282	Inf
N10k	1.5	1.52	1.62	1	1.52	1.62	Inf	Inf	293	4.4e+302	Inf
N100k	1.5	1.50	0.07	0	1.50	0.07	1.51	0.79	1	1.51	0.79
FourG	1.5	1.50	0.20	0	1.50	0.20	1.51	0.59	29	1.51	0.59
NoUp	1.5	1.51	0.56	1	1.51	0.56	Inf	Inf	432	2.5e+130	Inf
DoseMods	1.5	1.50	0.18	0	1.50	0.18	1.52	1.01	1	1.52	1.01
HR1	1	Inf	Inf	49	2.3e+10	Inf	Inf	Inf	824	1.9e+82	Inf
HR10p	1.1	108.96	9805.40	11	110.16	9805.40	Inf	Inf	547	4.6e+132	Inf
HR20p	1.2	1.21	1.15	4	1.21	1.15	Inf	Inf	390	1.8e+161	Inf
HR5	5	5.01	0.22	0	5.01	0.22	5.09	1.75	0	5.09	1.75
Half1	1.5	1.51	0.34	4	1.51	0.34	1.51	0.77	768	1.52	0.77
Half10	1.5	1.51	0.34	2	1.51	0.34	1.51	0.92	561	1.52	0.92
Half10.4G	1.5	1.50	0.20	2	1.50	0.20	1.51	0.50	510	1.51	0.50
Half30.4G	1.5	1.50	0.18	0	1.50	0.18	1.51	0.54	245	1.51	0.54
Half450.4G	1.5	1.53	1.94	0	1.53	1.94	Inf	Inf	2	4.7e+97	Inf
Half1k	1.5	Inf	Inf	60	1.5e+87	Inf	Inf	Inf	511	3.1e+203	Inf
Half1k4G	1.5	Inf	Inf	95	5.4e+20	Inf	Inf	Inf	250	1.4e+30	Inf
MVmono	3	2.98	-0.83	1	2.98	-0.83	2.98	-0.83	60	2.98	-0.83
MVmulti	3	2.97	-0.90	0	2.97	-0.90	2.97	-0.90	20	2.97	-0.90
Out10	1.5	1.46	-2.48	0	1.46	-2.48	1.54	2.62	458	1.55	2.62
Out10.4G	1.5	1.49	-0.48	0	1.49	-0.48	1.51	0.49	460	1.51	0.49
Out30.4G	1.5	1.50	-0.05	0	1.50	-0.05	1.51	0.55	217	1.51	0.55
Out450.4G	1.5	1.48	-1.28	0	1.48	-1.28	1.51	0.39	2	1.51	0.39
Out1k	1.5	1.89	25.73	0	1.89	25.73	1.52	1.12	2	1.52	1.12
Out1k4G	1.5	1.52	1.02	0	1.52	1.02	1.50	0.32	0	1.50	0.32

OPEE: One Parameter Effective Exposure; TPEE: Two Parameter Effective Exposure

Table C.4: Frequency of Minimum AIC Simulations per Scenario by Estimation Method

				#	Amongst Non-Failed		
	OPEE	TPEE	Categorical	Failed	OPEE	TPEE	Categorical
BC1	801	195	4	75	753	170	2
N1k	565	37	398	555	294	7	144
N10k	754	189	57	293	550	122	35
N100k	838	162	0	1	837	162	0
FourG	823	174	3	29	811	158	2
NoUp	846	70	84	432	481	43	44
DoseMods	824	176	0	1	823	176	0
HR1	438	10	552	826	98	2	74
HR10p	622	54	324	550	313	21	116
HR20p	735	120	145	391	464	60	85
HR5	876	124	0	0	876	124	0
Half1	379	52	569	770	106	21	103
Half10	714	102	184	563	334	58	45
Half10.4G	705	118	177	510	373	61	56
Half30.4G	790	167	43	245	628	101	26
Half450.4G	852	148	0	2	852	146	0
Half1k	768	30	202	511	389	19	81
Half1k4G	859	130	11	250	668	74	8
MVmono	997	0	3	61	937	0	2
MVmulti	1000	0	0	20	980	0	0
Out10	526	435	39	458	349	180	13
Out10.4G	341	636	23	460	266	263	11
Out30.4G	569	425	6	217	514	265	4
Out450.4G	58	939	3	2	58	938	2
Out1k	65	917	18	2	65	916	17
Out1k4G	1	990	9	0	1	990	9

Table C.5: Summaries of Estimated Single Half-Life Parameter Using One Parameter Effective Exposure Algorithm Across Scenarios

Simulation Scenario	True Half	Mean	% Bias	Median	# Failed	Amongst Non-Failed Mean	% Bias	Median
BC1	90	92.13	2.4	90	1	92.22	2.5	90
N1k	90	773.8	760	90	5	777.6	764	90
N10k	90	95.5	6.1	90	1	95.6	6.2	90
N100k	90	91.3	1.4	90	0	91.3	1.4	90
FourG[1]	90	91.9	2.1	90	0	91.9	2.1	90
NoUp[2]	90	97.7	8.6	90	1	97.8	8.7	90
DoseMods	90	91.2	1.34	89.1	0	91.2	1.34	89.1
HR1	90	1.9e+9	2.1e+9	75	49	1259	1299	75
HR10p	90	796	785	93.8	11	804	793	93.8
HR20p	90	103.7	15.2	87.5	4	104.1	15.6	87.5
HR5	90	90.4	0.41	90	0	90.4	0.41	90
Half1	1	3.76	275.8	1.41	4	3.77	277	1.41
Half10	10	11.8	18.3	9.38	2	11.85	18.5	9.38
Half10[1]	10	11.81	18.1	9.38	2	11.8	18.3	9.38
Half30[1]	30	31.3	4.3	30	0	31.3	4.3	30
Half450[1]	450	491.9	9.3	450	0	491.9	9.3	450
Half1k	1000	4.1e+9	4.14e+8	900	60	4007	300.7	840
Half1k[1]	1000	9.4e+9	9.45e+8	1020	95	2839	183.4	900
MVmono	3	3.03	1.09	2.97	1	3.04	1.19	2.97
MVmulti	3	3.01	0.26	2.97	0	3.01	0.26	2.97
Out10	(90,10)	40.7	-54.8	37.5	0	40.7	-54.8	37.5
Out10[1]	(90,10)	45	-50	41.3	0	45	-50	41.3
Out30[1]	(90,30)	58.8	-34.7	56.3	0	58.8	-34.7	56.3
Out450[1]	(90,450)	226.9	152	225	0	226.9	152	225
Out1k	(90,1000)	500.5	456	465	0	500.5	456	465
Out1k4G	(90,1000)	387.9	331	360	0	387.9	331	360

[1] Four risk trajectories in simulation scenario
[2] Three risk trajectories included - *down*, *on*, *ctrl*

Table C.6: Summaries of Estimated Incline and Decline Half-Lives Using TPEE Algorithm Across Scenarios

									Amongst Non-Failed			
		Incline			Decline			#	Incline		Decline	
	True Half	Mean	% Bias	Median	Mean	% Bias	Median	Failed	Mean	% Bias	Mean	% Bias
BC1	90	133	47.8	90	94.95	37.1	90	74	123.5	5.5	98.4	9.4
N1k	90	40774	45205	105	13692	45133	71.3	554	40710	15114	878	875.3
N10k	90	15389	16999	90	107.7	11842	80.6	293	10748	19.7	133.8	48.7
N100k	90	97.4	8.2	90	93.8	8.2	90	1	97.4	4.2	93.8	4.2
FourG[1]	90	95.2	5.8	90	96.5	5.6	90	29	95.1	7.2	97.96	8.9
NoUp[2]	90	1e+9	1.2e+9	120	102.5	2699	90	432	2519	13.8	108.9	21
DoseMods	90	100	11.1	100	91.4	11.2	93.8	1	100.1	1.5	91.4	1.5
HR1	90	8.7e+9	9.7e+9	90	6.2e+10	35851	60	824	32356	6.9e+10	1138	11641
HR10p	90	47426	52595	100	9925	51995	75	547	46885	10928	576	540
HR20p	90	25798.6	28565	100	129.5	23359	75	390	21113	43.8	181.3	101.5
HR5	90	94.8	5.4	90	91.1	5.4	90	0	94.8	1.2	91.1	1.2
Half1	1	7.16	616.4	1.41	7.96	555.8	1.4	768	6.6	696.4	24.1	2307
Half10	10	15.79	57.9	8.8	16.6	53.3	9.4	561	15.3	66.1	28.02	180.2
Half10[1]	10	14.58	45.8	9.4	16.2	47	9.8	510	14.7	61.9	25.7	157.4
Half30[1]	30	33.95	13.2	30	35.8	8.7	30	245	32.6	19.4	42.2	40.7
Half450[1]	450	4411	880.2	480	523.4	225.4	480	2	1464.4	16.3	524	16.4
Half1k	1000	1.5e+10	1.5e+9	960	1.8e+11	2527	960	511	26270	1.8e+10	2752.4	175.2
Half1k[1]	1000	1.2e+10	1.2e+9	960	1.8e+11	203.1	960	250	3031.2	1.8e+10	2271.1	127.1
MVmono	3	3.34	11.2	3.12	3.22	13.2	3.12	60	3.40	7.5	3.28	9.4
MVmulti	3	3.22	7.4	3.12	3.18	7.8	3.12	20	3.24	5.97	3.19	6.4
Out10	(90,10)	102.7	14.2	80.6	24.5	11.6	13.1	458	100.5	145.1	37.1	270.9
Out10[1]	(90,10)	92	2.25	90	19	-1.2	11.3	460	88.9	89.6	29	190.2
Out30[1]	(90,30)	92.5	2.8	87.7	37.7	0.43	30	217	90.4	25.5	43.7	45.7
Out450[1]	(90,450)	95.4	6	90	482.7	6.2	480	2	95.6	7.3	482.7	7.3
Out1k	(90,1000)	101.6	12.9	90	1280.5	13	960	2	101.71	28.1	1235	23.5
Out1k[1]	(90,1000)	95	5.6	90	1221.4	5.6	990	0	95	22.1	1221.4	22.1

TPEE: Two Parameter Effective Exposure

[1] Four risk trajectories

[2] Three risk trajectories included - *down, on, ctrl*

Table C.7: OPEE Coverage Probabilities of the Incline, Decline, and Hazard Ratio by Scenario. Amongst full set of 1,000 simulations and Non-failed Algorithm Estimation Procedures

	Incline	Decline	Hazard Ratio	# Failed	NF Incline	NF Decline	NF Hazard Ratio
BC1	92.8		95.8	1	92.9		95.9
N1k	85		95.3	5	85.4		95.3
N10k	90.5		95.1	1	90.6		95.2
N100k	94.9		94	0	94.9		94
FourG	92.3		95.7	0	92.3		95.7
NoUp	91.5		95.8	1	91.6		95.9
DoseMods	94.2		95.4	0	94.2		95.4
HR1	70.5		89	49	74.1		92.2
HR10p	80		96.4	11	80.9		96.7
HR20p	85.6		95.9	4	85.9		95.9
HR5	94.2		96.5	0	94.2		96.5
Half1	98		95.2	4	98.4		95.2
Half10	85.9		94.7	2	86.1		94.7
Half10.4G	85.3		94.8	2	85.5		94.8
Half30.4G	90.7		95	0	90.7		95
Half450.4G	92.2		95.4	0	92.2		95.4
Half1k	75.9		81.6	60	80.4		86.8
Half1k4G	77.9		79.8	95	86.1		88.2
MVmono	93.7		94	1	93.8		94
MVmulti	94.9		92.8	0	94.9		92.8
Out10	35.5	67.3	89.3	0	35.5	67.3	89.3
Out10.4G	39.7	59.7	93.5	0	39.7	59.7	93.5
Out30.4G	57.9	77.5	95	0	57.9	77.5	95
Out450.4G	24.2	10.7	93.5	0	24.2	10.7	93.5
Out1k	1.2	14.3	42.4	0	1.2	14.3	42.4
Out1k4G	7.8	7.2	95.8	0	7.8	7.2	95.8

OPEE: One Parameter Effective Exposure
NF: Non-Failed Set of Simulations

Table C.8: TPEE Coverage Probabilities of the Incline, Decline, and Hazard Ratio by Scenario. Amongst full set of 1,000 simulations and Non-failed Algorithm Estimation Procedures

	Incline	Decline	Hazard Ratio	# Failed	NF Incline	NF Decline	NF Hazard Ratio
BC1	86.9	82.5	95.3	74	90.5	88.9	97.4
N1k	81	45.1	90.6	554	85.4	99.1	96.9
N10k	78.8	67.4	93	293	82.6	94.3	97.3
N100k	94.5	91.3	96.5	1	94.6	91.4	96.6
FourG	91.3	89.3	95.5	29	91.7	91.9	96
NoUp	61.4	57.1	65	432	100	93	99.1
DoseMods	95	89.8	97.2	1	95.1	89.8	97.3
HR1	52	17.7	65.5	824	76.1	100	85.2
HR10p	78.3	44.4	93.9	547	81.9	97.8	97.4
HR20p	79.8	59.2	94.2	390	82.6	95.6	98.2
HR5	86.1	61.4	94.7	0	86.1	61.4	94.7
Half1	86	21.6	82.4	768	100	92.7	92.2
Half10	79.4	41.5	89.9	561	85.4	94.5	94.1
Half10.4G	81.6	47.2	91.7	510	86.5	96.1	94.5
Half30.4G	84.1	70.6	93.4	245	85.7	93.5	94.8
Half450.4G	92.8	94.4	95.6	2	92.8	94.5	95.7
Half1k	29.8	45.9	30.1	511	58.7	93.5	60.3
Half1k4G	62.8	65	64.3	250	82.3	86	84.7
MVmono	86.2	77.6	93.7	60	91.6	82.6	94.4
MVmulti	90.7	85	92.8	20	92.6	86.7	93.6
Out10	79.9	51.1	93.5	458	81.5	94.3	95.8
Out10.4G	88.4	51.2	93.5	460	89.3	94.8	93.9
Out30.4G	89.7	73.3	94.6	217	90.4	93.6	95.3
Out450.4G	92.2	93.8	95.3	2	92.4	94	95.5
Out1k	91.6	89.8	96.6	2	91.7	90	96.7
Out1k4G	92.3	91.3	95.2	0	92.3	91.3	95.2

TPEE: Two Parameter Effective Exposure
NF: Non-Failed Set of Simulations

Table C.9: Interval-Based TPEE Results

									Amongst Non-Failed					
		Incline		Decline		Hazard Ratio		#	Incline		Decline		Hazard Ratio	
intlen	modtype	CP	Mean	CP	Mean	CP	Mean	Failed	CP	Mean	CP	Mean	CP	Mean
	cph1	88.3	454.6	85.5	98.5	94.7	1.1e+114	76	92.6	133.5	91.99	101.9	97.7	1.70
10	cph2	71.4	459.2	56.6	94.2	42.4	4e+114	425	88.4	69.1	93.7	119.7	66.96	1.49
Days	plr1	91.4	123.65	83.8	94.84	93.4	1.6	16	91.46	122.8	85.16	96.21	93.5	1.56
	plr2	90.80	118.88	84.30	90.68	93.20	1.6	18	91.34	118.69	85.74	91.67	93.48	1.56
	cph1	92.90	785.71	88.10	112.04	94.90	4e+211	64	97.76	160.77	93.91	114.85	98.08	1.70
50	cph2	80.20	70.46	56.10	124.68	36.70	1.5	394	98.02	25.00	86.96	142.32	44.72	1.51
Days	plr1	89.80	3.1e+4	81.80	112.90	90.40	Inf	38	92.83	3.2e+4	84.93	113.25	93.14	Inf
	plr2	94.00	2.2e+4	87.10	110.35	86.40	Inf	93	97.79	2.4e+4	95.59	111.31	90.52	Inf
	cph1	96.5	1770.6	92.4	129.6	95.3	2.7e+218	43	100.00	1583.7	96.5	131.10	98.1	2.9e+218
100	cph2	85.9	9.4	67.1	173.3	30.7	1.6	320	98.09	2.13	97.65	187.3	31.03	1.6
Days	plr1	93.7	7.4e+4	81.4	134.11	90.1	Inf	38	97.3	7.5e+4	84.5	133.25	93.14	Inf
	plr2	87.9	9174.1	64.6	150.7	57.9	Inf	258	100	886.9	86.9	156.8	66.71	2.9e+259
	cph1	95.60	3.3e+4	83.60	194.30	95.10	Inf	47	100.00	2e+4	85.52	195.72	99.27	Inf
300	cph2	80.20	0.03	0.10	216.10	71.80	1.8	997	100.00	0.03	33.33	333.33	100.00	1.7
Days	plr1	96.4	411.7	55.7	202	93.4	1.3e+27	43	100	395	58.2	203.97	97.39	1.3e+27
	plr2	84.5	0.03	3.4	211.3	0.30	1.8	965	100	0.03	97.14	244.73	5.71	1.8
	cph1	0.10	150.04	0	374.08	0.10	1.42	999	100.00	150.00	0	800.00	100.00	1.58
900	cph2	0.00	0.03	0	400.00	0.00	2.30	1000						
Days	plr1	0.00	115.69	130	369.21	0.00	1.48	1000						
	plr2	67.60	0.03	0	400.00	0.00	2.41	1000						

HR: Hazard Ratio; CP: Coverage Probability; μ: Mean across simulations; CPH: Cox Proportional Hazards Regression; PLR: Pooled Logistic Regression
Structure 1 (cph1, plr1): Cases and Non-Cases assigned exposure at end-point of interval times.
Structure 2 (cph2, plr2): Cases assigned exposure at event time and Non-Cases designated by exposure at end-point of interval times.

Table C.10: Interval-Based Standard Metrics Results

intlen	modtype	CP CalcC	Mean CalcC	CP PvsN	Mean PvsN	CP CvsN	Mean CvsN	CP Curr	Mean Curr	CP Ever	Mean Ever	CP TrueEE	Mean TrueEE
10 Days	cph1	94.20	1.49	0	1.07	68.00	1.41	21.40	1.36	0.20	1.24	94.20	1.49
	cph2			0	1.07	68.50	1.41	22.30	1.36	0.20	1.24	94.30	1.49
	plr1	95.10	1.50	0	1.07	73.30	1.42	30.30	1.37	0.20	1.24	95.00	1.50
50 Days	cph1	92.80	1.48	0	1.07	66.40	1.41	23.30	1.36	0.20	1.24	93.20	1.48
	cph2			0	1.07	66.70	1.41	23.90	1.36	0.10	1.24	92.80	1.48
	plr1	94.90	1.50	0	1.07	73.80	1.42	31.30	1.37	0.20	1.24	94.70	1.50
100 Days	cph1	88.00	1.47	0	1.07	65.20	1.40	24.50	1.36	0.10	1.23	89.30	1.47
	cph2			0	1.07	66.40	1.41	26.10	1.36	0.20	1.24	89.80	1.47
	plr1	93.70	1.48	0	1.11	74.90	1.42	32.80	1.37	0.20	1.24	94.40	1.49
300 Days	cph1	61.60	1.41	0	1.07	67.20	1.40	22.90	1.36	0.10	1.23	65.60	1.41
	cph2			0	1.07	71.00	1.41	29.20	1.37	0.20	1.24	70.60	1.42
	plr1	69.90	1.42	0	1.07	78.90	1.43	39.40	1.38	0.40	1.25	80.90	1.44
900 Days	cph1	34.20	1.38	0	1.06	59.60	1.39	14.40	1.35	0.10	1.23	19.80	1.36
	cph2			0	1.07	73.20	1.42	29.90	1.37	0.20	1.24	30.20	1.38
	plr1	59.10	1.41	0	1.07	88.50	1.46	58.30	1.40	0.80	1.26	71.40	1.42

CP: Coverage Probability, in percentage points; μ: Mean across simulations; CPH: Cox Proportional Hazards Regression; PLR: Pooled Logistic Regression
Structure 1 (cph1, plr1): Data split into time intervals at end of interval. Cases and Non-Cases assigned exposure at interval times.
Structure 2 (cph2): Data split into time intervals at end of interval. Cases assigned exposure associated with event timing and Non-Cases designated by exposure at interval times.
PvsN: Past versus Never Estimated Ratio in Categorical Model; CvsN: Current versus Never Estimated Ratio in Categorical Model; Curr: Current vs. Not Current Estimated Ratio; Ever: Ever vs. Never Estimated Ratio; CalcC: Calculated True Effective Exposure Ratio Estimate at Interval-Based Times (not relevant for cph2, because cases have true EE assigned); TrueEE: True Effective Exposure Ratio Estimate with value from Interval for Non-Cases and Event Time for Cases

APPENDIX D

BWHS Results

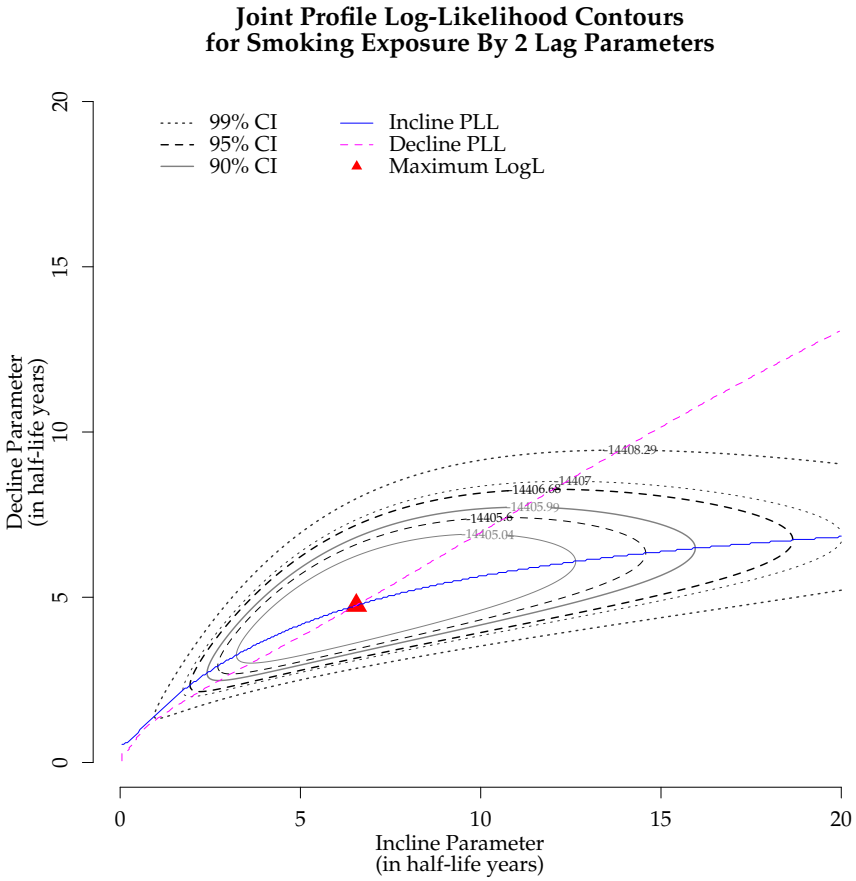


Figure D.1: Overall Effective Exposure Contour for Log-Likelihoods Across Combinations of Incline and Decline Lag Parameters amongst the simple subset of BWHS participants.

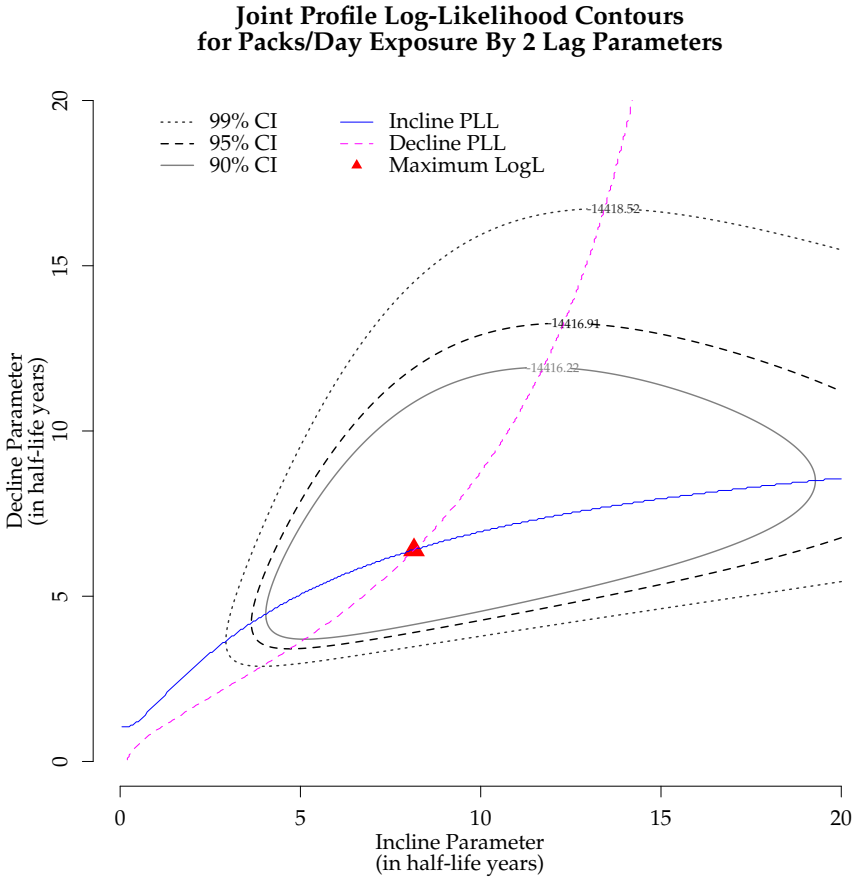


Figure D.2: Packs Per Day Effective Exposure Contour for Log-Likelihoods Across Combinations of Incline and Decline Lag Parameters amongst the simple subset of BWHS participants.

Table D.1: Smoking and risk of CVD: standard analyses amongst the **Restricted Subset** of BWHS participants (1995-2015)

Smoking Exposure Model		Hazard Ratio	95% Confidence Interval Lower	Upper	Log-Likelihood	AIC
Smoking Categories	Current	2.55	2.26	2.87	-14407.77	28849.53
	Past	1.11	1.01	1.23		
	Ever vs. Never	1.43	1.31	1.55	-14479.75	28991.49
	Current vs. Not	2.46	2.19	2.75	-14409.98	28851.97
	Cumulative Years Smoked	1.02	1.02	1.02	-14442.71	28917.42
	Current Smokers: Packs/Day	2.55	2.23	2.90	-14437.65	28907.31
Combination Model 1[1]	Current Smoker	2.16	1.90	2.47	-14403.18	28840.36
	Pack-Years	1.01	1.00	1.01		
Combination Model 2[2]	Ever Smoker	1.17	1.05	1.30	-14460.42	28954.83
	Pack-Years	1.01	1.01	1.02		

AIC: Akaike's Information Criterion

[1,2] Combination models 1 and 2 assume that never smokers have zero pack-years.

Table D.2: Effective Exposure of Smoking on Risk of Cardiovascular Event. Application of Profile Likelihood Fits and Estimation Algorithms to **Restricted Subset** of Participants from BWHS (1995-2015)

Estimation Approach	Exposure Input	Smoking HR (95% CI)	Lag Parameter(s) (95% CI)[2] Incline	Decline	Log-Likelihood	AIC
Profile Likelihood for One Parameter[1]	Binary	2.52 (2.25-2.83)	4.50 (2.35-7.30)		-14404.42	28842.84
	Packs/Day	2.74 (2.42-3.10)	6.90 (4.15-11.85)		-14414.28	28862.56
OPEE Algorithm	Binary	2.52 (2.25-2.83)	4.50 (2.10-6.91)		-14404.42	28842.85
	Packs/Day	2.73 (2.41-3.10)	6.88 (3.40-10.36)		-14414.28	28862.55
Profile Likelihood Search Over Two Parameters[1]	Binary	2.77 (2.44-3.15)	6.55 (1.95-18.65)	4.75 (2.15-8.25)	-14403.68	28843.36
	Packs/Day	2.96 (2.59-3.38)	8.15 (3.65-20.00)[3]	6.40 (3.45-13.25)	-14413.92	28863.84
TPEE Algorithm	Binary	2.73 (2.20-3.39)	6.00 (0-12.01)	4.5 (1.16-7.84)	-14403.71	28843.43
	Packs/Day	2.99 (2.35-3.80)	8.00 (2.16-13.84)	6.00 (2.39-9.61)	-14413.96	28863.91

HR: Hazard Ratio; CI: Confidence Interval; OPEE: One Parameter Effective Exposure; TPEE: Two Parameter Effective Exposure; AIC: Akaike's Information Criterion

[1] Confidence bounds for Profile Likelihood Hazard Ratios based on final model selected, i.e. no adjustments made for multiple testing.

[2] Confidence bounds for half-life lag parameters in OPEE and TPEE algorithms are based on the asymptotically normal approximation of standard errors, while CI for the profile likelihood methods are based on the likelihood ratio test statistic assuming χ^2_2 and χ^2_1 distributions for the joint and marginal confidence bounds for the two- and one-parameter profiles, respectively.

[3] Bound stops at 20 years, because profile fits were not performed for half-lives>20.

REFERENCE EQUATIONS

INTRODUCTION

Likelihood and Log-Likelihood of Cox Proportional Hazards Regression across k distinct event times. The m_k and $j \epsilon R(t_k)$ represent the total number of events and the index of subjects at risk at event time t_k.

$$\begin{aligned} \mathcal{L}(X, \beta, Y) &= \prod_{k=1}^{K} \frac{\sum\limits_{j \epsilon R(t_k, Y_j=1)} exp(X_j \beta)}{\left[\sum\limits_{j \epsilon R(t_k)} exp(X_j \beta) \right]^{m_k}} \\ \ell(X, \beta, Y) &= \sum_{k=1}^{K} \left[\sum_{j \epsilon R(t_k, Y_j=1)} \left(\sum_{q} x_{qj} \beta_q \right) - m_k \ln \left(\sum_{j \epsilon R(t_k)} e^{\sum\limits_{q} x_{qj} \beta_q} \right) \right] \end{aligned} \tag{D.1}$$

Cox Proportional Hazards.

$$h(t|\boldsymbol{X}(t)) = h_0(t) exp[x_{1j}\beta_1 + ... + x_{qj}\beta_q] \tag{D.2}$$

Risk Sum - Multivariate

$$\hat{r_i} = \sum_{q} x_{qi} \hat{\beta}_q = x_{1i} \hat{\beta}_1 + ... + x_{qi} \hat{\beta}_q \tag{D.3}$$

Likelihood ($\mathcal{L}$) and Log-Likelihood (ℓ) of the Logistic Regression Models

$$\begin{aligned}\mathcal{L}(X,\beta,Y) &= \prod_{i=1}^{n}\prod_{t=1}^{T} p_{it}^{Y_{it}}(1-p_{it})^{(1-Y_{it})} \\ \ell(X,\beta,Y) &= \sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it} ln p_{it} + (1-Y_{it}) ln(1-p_{it})\end{aligned} \tag{D.4}$$

Probability of Event, p_{it}, for subject i at time t with q time-varying risk factors in the logistic multivariate model.

$$p_{it} = \frac{exp(\beta_0 + x_{1it}\beta_1 + ... + x_{qit}\beta_q)}{1 + exp(\beta_0 + x_{1it}\beta_1 + ... + x_{qit}\beta_q)} \tag{D.5}$$

One-Compartment Model for Infusion

$$C_p^t = \begin{cases} \frac{k_0}{k_e V}\left[1 - e^{-k_e t}\right] & \text{if } t \leq D \\ \frac{k_0}{k_e V}\left[1 - e^{-k_e D}\right] e^{-k_e (t-D)} & \text{if } t > D \end{cases} \tag{D.6}$$

Where D is the end–time for the infusion, $t - D$ is the time elapsed since ending the infusion, and where the first condition is equivalent to the second by replacing D with t, when t is less than or equal to D. k_0 and k_e are the infusion and elimination rates, and V represents the volume of the administered infusion. At steady state, the total concentration is equivalent to the first part of the equation:

$$C_p^{ss} = \frac{k_0}{k_e V} \tag{D.7}$$

Thus, the relative concentration at time t vs. the steady state (ss) level can be described by the current concentration, C_p^t relative to the steady state concentration,

C_p^{ss}:

$$\frac{C_p^t}{C_p^{ss}} = \frac{k_0}{k_e V}\left[1 - e^{-k_e D}\right] * e^{-k_e(t-D)} \times \frac{k_e V}{k_0} = \left[1 - e^{-k_e D}\right] * e^{-k_e(t-D)} \tag{D.8}$$

Turning OCM into a generalized form of the relative ratio

$$C_{ratio} = \frac{C_p^t}{C_p^{ss}} = \left[1 - e^{-k_e D}\right] * e^{-k_e(t-D)}$$

Condition of assumed curves

$$C_{ratio} = \begin{cases} 0 & \text{if } t \leq b \\ 1 - e^{-k_e(t-b)} & \text{if } b < t \leq f \\ \left[1 - e^{-k_e(f-b)}\right] * e^{-k_e(t-f))} & \text{if } t > f \end{cases} \tag{D.9}$$

DERIVATION

Basic formulation of the Effective Exposure over time given a single lag parameter, known exposure level, and known start and stop times. D can be assumed to take the value of 1 for exposed and 0 for unexposed, which will be referred to as the "Binary Dosing Scheme".

$$\begin{aligned} E_{it}(\lambda, b, f) &= D\left(1 - e^{-\lambda(t-b)}\right) * I\left(t\epsilon\left[b, f\right]\right) + D\left(1 - e^{-\lambda(f-b)}\right) e^{-\lambda(t-f)} * I\left(t > f\right) \\ &= D * \left[e^{-\lambda * \max(0,t-f)} - e^{-\lambda * \max(0,t-b)}\right] = D\left[e^{-\lambda z_2} - e^{-\lambda z_1}\right] \end{aligned} \tag{D.10}$$

where

$$\begin{aligned} z_1 = \max(0, t-b) &= \begin{cases} t-b & \text{if } t > b \\ 0 & \text{otherwise} \end{cases} \\ z_2 = \max(0, t-f) &= \begin{cases} t-f & \text{if } t > f \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{D.11}$$

and where b and f represent the times the exposure starts and stops or is discontinued, respectively, for a given subject. D, b, and f are fixed values, i.e. not time-varying.

Log-Likelihood of the Cox Proportional Hazards Model - Univariate for the Effective Exposure Measure.

$$\boldsymbol{\ell(X, \beta, \lambda, Y)} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \left[\sum_{\boldsymbol{j \epsilon R(t_k, Y_j = 1)}} \boldsymbol{\beta E_j(\lambda)} - \boldsymbol{m_k} \ln \left(\sum_{\boldsymbol{j \epsilon R(t_k)}} \boldsymbol{e^{\beta E_j(\lambda)}} \right) \right] \tag{D.12}$$

Predicted probability of event for subject i at time t based on the logistic regression parameters. (Relevant for simulation development)

$$\hat{p_{it}} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 E_{it}(\hat{\lambda})}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 E_{it}(\hat{\lambda})}} \tag{D.13}$$

where $\hat{\beta}_0 + \hat{\beta}_1 E_{it}(\hat{\lambda}) \rightarrow \hat{\beta}_0 + \hat{\beta}_1 E_{it}(\hat{\lambda}) + \hat{\gamma_1} x_{1it} + ... + \hat{\gamma_q} x_{qit}$ for the multivariate setting. Fisher's Information Matrix for the One Parameter Effective Exposure Models. Variations in λ and h simply require the appropriate substitution of the second derivatives for either model's (logistic or Cox PH) log-likelihood.

$$I(\lambda, \beta) = - \begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta^2} & \frac{\partial^2 \ell}{\partial \beta \partial \lambda} \\ \frac{\partial^2 \ell}{\partial \beta \partial \lambda} & \frac{\partial^2 \ell}{\partial \lambda^2} \end{bmatrix} \tag{D.14}$$

Fisher's Information Matrix for the Two Parameter Effective Exposure Models. Variations in (λ_1, λ_2) and (h_1, h_2) simply require the appropriate substitution of the second derivatives for either model's (logistic or Cox PH) log-likelihood.

$$I(\lambda_1, \lambda_2, \beta) = - \begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta^2} & \frac{\partial^2 \ell}{\partial \beta \partial \lambda_1} & \frac{\partial^2 \ell}{\partial \beta \partial \lambda_2} \\ \frac{\partial^2 \ell}{\partial \beta \partial \lambda_1} & \frac{\partial^2 \ell}{\partial \lambda_1^2} & \frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_2} \\ \frac{\partial^2 \ell}{\partial \beta \partial \lambda_2} & \frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_2} & \frac{\partial^2 \ell}{\partial \lambda_2^2} \end{bmatrix} \tag{D.15}$$

Effective Exposure Specification

Single Dosing, Single Lag Parameter equation (Main text equation (2.3)) and Single Dosing, Two Lag Parameters equation (Main text equation (2.8)) where z_1 and z_2 retain the same specifications as mentioned in condition (2.2).

$$\begin{aligned} E_{it}(\lambda) &= D\left(e^{-\lambda z_2} - e^{-\lambda z_1}\right) \\ E_{it}(h) &= D\left(e^{-z_2 \log 2/h} - e^{-z_1 \log 2/h}\right) \end{aligned} \tag{D.16}$$

$$\begin{aligned} E_{it}(\lambda_1, \lambda_2) &= D\left(1 - e^{-\lambda_1 (z_1 - z_2)}\right) e^{-\lambda_2 z_2} \\ E_{it}(h_1, h_2) &= D\left[1 - e^{-(z_1 - z_2)\log 2/h_1}\right] e^{-z_2 \log 2/h_2} \end{aligned} \tag{D.17}$$

First Derivative with respect to lag or half-life parameter for equation (2.3)

$$\begin{aligned} \frac{\partial E_{it}(\lambda)}{\partial \lambda} &= D\left[z_1 e^{-\lambda z_1} - z_2 e^{-\lambda z_2}\right] \\ \frac{\partial E_{it}(h)}{\partial h} &= \frac{D \log 2}{h^2}\left[z_2 e^{-z_2 \log 2/h} - z_1 e^{-z_1 \log 2/h}\right] \end{aligned} \tag{D.18}$$

Second Derivative with respect to lag or half-life parameter for equation (2.3)

$$\begin{aligned} \frac{\partial^2 E_{it}(\lambda)}{\partial \lambda^2} &= D\left[z_2^2 e^{-\lambda z_2} - z_1^2 e^{-\lambda z_1}\right] \\ \frac{\partial^2 E_{it}(h)}{\partial h^2} &= \frac{D\left(\log 2\right)^2}{h^4}\left(z_2^2 e^{-z_2 \log 2/h} - z_1^2 e^{-z_1 \log 2/h}\right) \\ &\quad - \frac{2D \log 2}{h^3}\left(z_2 e^{-z_2 \log 2/h} - z_1 e^{-z_1 \log 2/h}\right) \end{aligned} \tag{D.19}$$

First Derivative with respect to incline parameter for equation (2.8)

$$\begin{aligned} \frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1} &= D\left(z_1 - z_2\right) e^{-\lambda_1 (z_1 - z_2) - \lambda_2 z_2} \\ \frac{\partial E_{it}(h_1, h_2)}{\partial h_1} &= \frac{-D\left(z_1 - z_2\right)\log 2}{h_1^2} e^{-\log 2[(z_1 - z_2)/h_1 + z_2/h_2]} \end{aligned} \tag{D.20}$$

First Derivative with respect to decline parameter for equation (2.8)

$$\begin{aligned}\frac{\partial E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_2} &= -Dz_2\left(1-e^{-\lambda_1(z_1-z_2)}\right)e^{-\lambda_2 z_2}\\ \frac{\partial E_{it}(h_1,h_2)}{\partial h_2} &= \frac{Dz_2\log 2}{h_2^2}\left(1-e^{-(z_1-z_2)\log 2/h_1}\right)e^{-z_2\log 2/h_2}\end{aligned} \tag{D.21}$$

Second Derivative with respect to incline parameter for equation (2.8)

$$\begin{aligned}\frac{\partial^2 E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_1^2} &= -D\left(z_1-z_2\right)^2 e^{-\lambda_1(z_1-z_2)-\lambda_2 z_2}\\ \frac{\partial^2 E_{it}(h_1,h_2)}{\partial {h_1}^2} &= \frac{D\left(z_1-z_2\right)\log 2}{h_1^2}\left(\frac{2}{h_1}-\frac{(z_1-z_2)\log 2}{h_1^2}\right)e^{-\log 2[(z_1-z_2)/h_1+z_2/h_2]}\end{aligned} \tag{D.22}$$

Second Derivative with respect to decline parameter for equation (2.8)

$$\begin{aligned}\frac{\partial^2 E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_2^2} &= Dz_2^2\left(1-e^{-\lambda_1(z_1-z_2)}\right)e^{-\lambda_2 z_2}\\ \frac{\partial^2 E_{it}(h_1,h_2)}{\partial {h_2}^2} &= \frac{Dz_2\log 2}{h_2^2}\left(\frac{2}{h_2}-\frac{z_2\log 2}{h_1^2}\right)\left(1-e^{-(z_1-z_2)\log 2/h_1}\right)e^{-z_2\log 2/h_2}\end{aligned} \tag{D.23}$$

Second Derivative with respect to both incline and decline parameters for equation (2.8)

$$\begin{aligned}\frac{\partial^2 E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_1\partial \lambda_2} &= -Dz_2\left(z_1-z_2\right)e^{-\lambda_1(z_1-z_2)-\lambda_2 z_2}\\ \frac{\partial^2 E_{it}(h_1,h_2)}{\partial h_1\partial h_2} &= \frac{-Dz_2\left(z_1-z_2\right)(\log 2)^2}{h_1^2h_2^2}e^{-\log 2[(z_1-z_2)/h_1+z_2/h_2]}\end{aligned} \tag{D.24}$$

Multiple Dosing

When considering multiple dosings, or trajectories that may increase or decrease more than once, the formulation of $E_{it}(\Lambda)$ becomes a sum function of the individual dosing exposures. To illustrate, assume a subject is exposed from years 0 to 20

and 60 to 80 at a dose level of 2, while being exposed at a dose level of 1 from time 20 to 60. The current effective exposure at time t can be calculated as:

$$\begin{aligned} E_{it}^{(tot)}(\lambda) = D_1 \left(1 - e^{-(z_1 - z_2)\lambda}\right) e^{-z_2\lambda} \\ + D_2 \left(1 - e^{-(z_3 - z_4)\lambda}\right) e^{-z_4\lambda} \\ + D_3 \left(1 - e^{-(z_5 - z_6)\lambda}\right) e^{-z_6\lambda} \end{aligned} \tag{D.25}$$

where

$$z_1 = \begin{cases} t & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases} \qquad z_4 = z_5 = \begin{cases} t - 60 & \text{if } t > 60 \\ 0 & \text{otherwise} \end{cases}$$

$$z_2 = z_3 = \begin{cases} t - 20 & \text{if } t > 20 \\ 0 & \text{otherwise} \end{cases} \qquad z_6 = \begin{cases} t - 80 & \text{if } t > 80 \\ 0 & \text{otherwise} \end{cases}$$

This could also be written in a piecewise fashion, which implies that the first and second derivatives can be readily calculated for each individual exposure event. Let $E_{it}^{(1)}(\lambda)$, $E_{it}^{(2)}(\lambda)$, and $E_{it}^{(3)}(\lambda)$, be the effective exposure components for each of the exposure events.

$$\begin{aligned} E_{it}^{(1)}(\lambda) &= D_1 \left(1 - e^{-(z_1 - z_2)\lambda}\right) e^{-z_2\lambda} \\ E_{it}^{(2)}(\lambda) &= D_2 \left(1 - e^{-(z_3 - z_4)\lambda}\right) e^{-z_4\lambda} \\ E_{it}^{(3)}(\lambda) &= D_3 \left(1 - e^{-(z_5 - z_6)\lambda}\right) e^{-z_6\lambda} \\ E_{it}^{(tot)}(\lambda) &= E_{it}^{(1)}(\lambda) + E_{it}^{(2)}(\lambda) + E_{it}^{(3)}(\lambda) \end{aligned} \tag{D.26}$$

As mentioned previously, the λ can be interchanged with $\log 2/h$, and $E_{it}^{(tot)}(h)$ represents the longitudinal function of EE based on this parameterization. To transition from OPEE to TPEE, I can substitute the single parameter with (λ_1, λ_2) or

(h_1, h_2) in equation (2.11), updating the piecewise component exposures that feed into $E_{it}^{(tot)}$.

Cox Proportional Hazards Models

OPEE Log-Likelihood

$$\ell(\beta, \lambda) = \sum_{k=1}^{K} \left[A_1(t_k) - m_k \log \left(C_1(t_k) \right) \right] \tag{D.27}$$

OPEE Score Function

$$\begin{aligned} \frac{\partial \ell}{\partial \beta} &= \sum_{k=1}^{K} \left[A_{2\beta}(t_k) - m_k \frac{B_1(t_k)}{C_1(t_k)} \right] \\ \frac{\partial \ell}{\partial \lambda} &= \sum_{k=1}^{K} \left[A_{2\lambda}(t_k) - m_k \frac{C_{2\lambda}(t_k)}{C_1(t_k)} \right] \end{aligned} \tag{D.28}$$

OPEE Fisher's Information

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \beta^2} &= \sum_{k=1}^{K} m_k \frac{\left[B_1(t_k) \right]^2 - B_{2\beta}(t_k) C_1(t_k)}{\left[C_1(t_k) \right]^2} \\ \frac{\partial^2 \ell}{\partial \lambda^2} &= \sum_{k=1}^{K} \left[A_3(t_k) - m_k \frac{C_3(t_k) C_1(t_k) - \left(C_{2\lambda}(t_k) \right)^2}{\left(C_1(t_k) \right)^2} \right] \\ \frac{\partial^2 \ell}{\partial \beta \partial \lambda} &= \sum_{k=1}^{K} \left[\frac{A_{2\lambda}(t_k)}{\beta} - m_k \frac{B_{2\lambda}(t_k) C_1(t_k) - C_{2\lambda}(t_k) B_1(t_k)}{\left(C_1(t_k) \right)^2} \right] \end{aligned} \tag{D.29}$$

TPEE Score Function

$$\begin{aligned} \frac{\partial \ell}{\partial \beta} &= \sum_{k=1}^{K} \left[A_{2\beta}(t_k) - m_k \frac{B_1(t_k)}{C_1(t_k)} \right] \\ \frac{\partial \ell}{\partial \lambda_1} &= \sum_{k=1}^{K} \left[A_4(t_k) - m_k \frac{C_4(t_k)}{C_1(t_k)} \right] \\ \frac{\partial \ell}{\partial \lambda_2} &= \sum_{k=1}^{K} \left[A_5(t_k) - m_k \frac{C_5(t_k)}{C_1(t_k)} \right] \end{aligned} \tag{D.30}$$

TPEE Fisher's Information

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial \beta^2} &= \sum_{k=1}^{K} m_k \frac{\left[B_1(t_k)\right]^2 - B_{2\beta}(t_k) C_1(t_k)}{\left[C_1(t_k)\right]^2} \\
\frac{\partial^2 \ell}{\partial \lambda_1^2} &= \sum_{k=1}^{K} \left[A_6(t_k) - m_k \frac{C_6(t_k) C_1(t_k) - \left(C_4(t_k)\right)^2}{\left(C_1(t_k)\right)^2} \right] \\
\frac{\partial^2 \ell}{\partial \lambda_2^2} &= \sum_{k=1}^{K} \left[A_7(t_k) - m_k \frac{C_7(t_k) C_1(t_k) - \left(C_5(t_k)\right)^2}{\left(C_1(t_k)\right)^2} \right] \\
\frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_2} &= \sum_{k=1}^{K} \left[A_8(t_k) - m_k \frac{C_8(t_k) C_1(t_k) - C_4(t_k) C_5(t_k)}{\left(C_1(t_k)\right)^2} \right] \\
\frac{\partial^2 \ell}{\partial \beta \partial \lambda_1} &= \sum_{k=1}^{K} \left[\frac{A_4(t_k)}{\beta} - m_k \frac{B_{2\lambda_1}(t_k) C_1(t_k) - C_4(t_k) B_1(t_k)}{\left(C_1(t_k)\right)^2} \right] \\
\frac{\partial^2 \ell}{\partial \beta \partial \lambda_2} &= \sum_{k=1}^{K} \left[\frac{A_5(t_k)}{\beta} - m_k \frac{B_{2\lambda_2}(t_k) C_1(t_k) - C_5(t_k) B_1(t_k)}{\left(C_1(t_k)\right)^2} \right]
\end{aligned}
\tag{D.31}
$$

Where we define the following compute-able quantities[1]:

$$A_1(t_k) = \sum_{j \epsilon \mathcal{R}(t_k, Y_j = 1)} \beta E_j(\lambda) = \sum_{j \epsilon \mathcal{R}(t_k, Y_j = 1)} \beta E_j(\lambda_1, \lambda_2)$$

$$A_{2\beta}(t_k) = \frac{\partial A_1(t_k)}{\partial \beta} = \sum_{j \epsilon \mathcal{R}(t_k, Y_j = 1)} E_j(\lambda) = \sum_{j \epsilon \mathcal{R}(t_k, Y_j = 1)} E_j(\lambda_1, \lambda_2)$$

$$A_{2\lambda}(t_k) = \frac{\partial A_1(t_k)}{\partial \lambda} = \sum_{j \epsilon \mathcal{R}(t_k, Y_j = 1)} \beta \frac{\partial E_j(\lambda)}{\partial \lambda}$$

$$A_3(t_k) = \frac{\partial^2 A_1(t_k)}{(\partial \lambda)^2} = \sum_{j \epsilon \mathcal{R}(t_k, Y_j = 1)} \beta \frac{\partial^2 E_j(\lambda)}{\partial \lambda^2}$$

$$A_4(t_k) = \frac{\partial A_1(t_k)}{\partial \lambda_1} = \sum_{j \epsilon \mathcal{R}(t_k, Y_j = 1)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1}$$

$$A_5(t_k) = \frac{\partial A_1(t_k)}{\partial \lambda_2} = \sum_{j \epsilon \mathcal{R}(t_k, Y_j = 1)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2}$$

[1] As these functions utilize the first and second derivatives of the effective exposure, one can switch between the λ and h by adjusting which internal formulas to use, maintaining consistency once estimation has started.

$$A_6(t_k) = \frac{\partial^2 A_1(t_k)}{\partial \lambda_1^2} = \sum_{j \epsilon \mathcal{R}(t_k, Y_j=1)} \beta \frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_1^2}$$

$$A_7(t_k) = \frac{\partial^2 A_1(t_k)}{\partial \lambda_2^2} = \sum_{j \epsilon \mathcal{R}(t_k, Y_j=1)} \beta \frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_2^2}$$

$$A_8(t_k) = \frac{\partial^2 A_1(t_k)}{\partial \lambda_1 \partial \lambda_2} = \sum_{j \epsilon \mathcal{R}(t_k, Y_j=1)} \beta \frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_1 \partial \lambda_2}$$

$$B_1(t_k) = \sum_{j \epsilon \mathcal{R}(t_k)} E_j(\lambda) e^{\beta E_j(\lambda)} = \sum_{j \epsilon \mathcal{R}(t_k)} E_j(\lambda_1, \lambda_2) e^{\beta E_j(\lambda_1, \lambda_2)}$$

$$B_{2\beta}(t_k) = \frac{\partial B_1(t_k)}{\partial \beta} = \sum_{j \epsilon \mathcal{R}(t_k)} (E_j(\lambda))^2 e^{\beta E_j(\lambda)} = \sum_{j \epsilon \mathcal{R}(t_k)} (E_j(\lambda_1, \lambda_2))^2 e^{\beta E_j(\lambda_1, \lambda_2)}$$

$$B_{2\lambda}(t_k) = \frac{\partial B_1(t_k)}{\partial \lambda} = \sum_{j \epsilon \mathcal{R}(t_k)} \frac{\partial E_j(\lambda)}{\partial \lambda} e^{\beta E_j(\lambda)} \left[1 + \beta E_j(\lambda)\right]$$

$$B_{2\lambda_1}(t_k) = \frac{\partial B_1(t_k)}{\partial \lambda_1} = \sum_{j \epsilon \mathcal{R}(t_k)} \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} e^{\beta E_j(\lambda_1, \lambda_2)} \left[1 + \beta E_j(\lambda_1, \lambda_2)\right]$$

$$B_{2\lambda_2}(t_k) = \frac{\partial B_1(t_k)}{\partial \lambda_2} = \sum_{j \epsilon \mathcal{R}(t_k)} \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} e^{\beta E_j(\lambda_1, \lambda_2)} \left[1 + \beta E_j(\lambda_1, \lambda_2)\right]$$

$$C_1(t_k) = \sum_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)} = \sum_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda_1, \lambda_2)}$$

$$C_{2\lambda}(t_k) = \frac{\partial C_1(t_k)}{\partial \lambda} = \sum_{j \epsilon \mathcal{R}(t_k)} \beta \frac{\partial E_j(\lambda)}{\partial \lambda} e^{\beta E_j(\lambda)}$$

$$C_3(t_k) = \frac{\partial^2 C_1(t_k)}{(\partial \lambda)^2} = \sum_{j \epsilon \mathcal{R}(t_k)} \beta e^{\beta E_j(\lambda)} \left[\frac{\partial^2 E_j(\lambda)}{\partial \lambda^2} + \beta \left(\frac{\partial E_j(\lambda)}{\partial \lambda} \right)^2 \right]$$

$$C_4(t_k) = \frac{\partial C_1(t_k)}{\partial \lambda_1} = \sum_{j \epsilon \mathcal{R}(t_k)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} e^{\beta E_j(\lambda_1, \lambda_2)}$$

$$C_5(t_k) = \frac{\partial C_1(t_k)}{\partial \lambda_2} = \sum_{j \epsilon \mathcal{R}(t_k)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} e^{\beta E_j(\lambda_1, \lambda_2)}$$

$$C_6(t_k) = \frac{\partial^2 C_1(t_k)}{\partial \lambda_1^2} = \sum_{j \epsilon \mathcal{R}(t_k)} \beta e^{\beta E_j(\lambda_1, \lambda_2)} \left[\frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_1^2} + \beta \left(\frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} \right)^2 \right]$$

$$C_7(t_k) = \frac{\partial^2 C_1(t_k)}{\partial \lambda_2^2} = \sum_{j \epsilon \mathcal{R}(t_k)} \beta e^{\beta E_j(\lambda_1, \lambda_2)} \left[\frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_2^2} + \beta \left(\frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} \right)^2 \right]$$

$$C_8(t_k) = \frac{\partial^2 C_1(t_k)}{\partial \lambda_1 \partial \lambda_2} = \sum_{j \epsilon \mathcal{R}(t_k)} \beta e^{\beta E_j(\lambda_1, \lambda_2)} \left[\frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_1 \partial \lambda_2} + \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} \right]$$

Pooled Logistic Regression Models

Pooled Log-Likelihood

$$\boldsymbol{\ell(Y, p) = \sum_{i=1}^{n} \sum_{t=1}^{T} Y_{it} \ln p_{it} + (1 - Y_{it}) \ln(1 - p_{it})} \tag{D.32}$$

where Y_{it} takes the value of 1 for events and 0 otherwise, for subject i at time t. The probability of event, p_{it}, under the logistic model is computed by (D.33):

$$p_{it} = \begin{cases} \dfrac{exp(\beta_0 + \beta_1 E_{it}(\lambda))}{1 + exp(\beta_0 + \beta_1 E_{it}(\lambda))} & \text{for OPEE} \\ \\ \dfrac{exp(\beta_0 + \beta_1 E_{it}(\lambda_1, \lambda_2))}{1 + exp(\beta_0 + \beta_1 E_{it}(\lambda_1, \lambda_2))} & \text{for TPEE} \end{cases} \tag{D.33}$$

OPEE Score Function

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_0} &= \sum_{i=1}^{n} \sum_{t=1}^{T} Y_{it} - p_{it} \\ \frac{\partial \ell}{\partial \beta_1} &= \sum_{i=1}^{n} \sum_{t=1}^{T} E_{it}(\lambda) \left(Y_{it} - p_{it} \right) \\ \frac{\partial \ell}{\partial \lambda} &= \sum_{i=1}^{n} \sum_{t=1}^{T} \beta_1 \left(Y_{it} - p_{it} \right) \frac{\partial E_{it}(\lambda)}{\partial \lambda} \end{aligned} \tag{D.34}$$

OPEE Fisher's Information

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial \beta_0^2} &= -\sum_{i=1}^{n}\sum_{t=1}^{T} \frac{\partial p_{it}}{\partial \beta_0} \\
\frac{\partial^2 \ell}{\partial \beta_1^2} &= -\sum_{i=1}^{n}\sum_{t=1}^{T} E_{it}(\lambda)\frac{\partial p_{it}}{\partial \beta_1} \\
\frac{\partial^2 \ell}{\partial \lambda^2} &= \sum_{i=1}^{n}\sum_{t=1}^{T} \beta_1 \left(Y_{it} - p_{it}\right)\left(\frac{\partial^2 E_{it}(\lambda)}{\partial \lambda^2}\right) - \beta_1 \left(\frac{\partial p_{it}}{\partial \lambda}\right)\left(\frac{\partial E_{it}(\lambda)}{\partial \lambda}\right) \\
\frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} &= -\sum_{i=1}^{n}\sum_{t=1}^{T} \frac{\partial p_{it}}{\partial \beta_1} \\
\frac{\partial^2 \ell}{\partial \beta_0 \partial \lambda} &= -\sum_{i=1}^{n}\sum_{t=1}^{T} \frac{\partial p_{it}}{\partial \lambda} \\
\frac{\partial^2 \ell}{\partial \beta_1 \partial \lambda} &= \sum_{i=1}^{n}\sum_{t=1}^{T} \left(Y_{it} - p_{it}\right)\frac{\partial E_{it}(\lambda)}{\partial \lambda} - E_{it}(\lambda)\frac{\partial p_{it}}{\partial \lambda}
\end{aligned}
\tag{D.35}
$$

TPEE Score Function

$$
\begin{aligned}
\frac{\partial \ell}{\partial \beta_0} &= \sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it} - p_{it} \\
\frac{\partial \ell}{\partial \beta_1} &= \sum_{i=1}^{n}\sum_{t=1}^{T} E_{it}(\lambda_1, \lambda_2)\left(Y_{it} - p_{it}\right) \\
\frac{\partial \ell}{\partial \lambda_1} &= \sum_{i=1}^{n}\sum_{t=1}^{T} \beta_1 \left(Y_{it} - p_{it}\right)\frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1} \\
\frac{\partial \ell}{\partial \lambda_2} &= \sum_{i=1}^{n}\sum_{t=1}^{T} \beta_1 \left(Y_{it} - p_{it}\right)\frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_2}
\end{aligned}
\tag{D.36}
$$

TPEE Fisher's Information

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial \beta_0^2} &= -\sum_{i=1}^{n}\sum_{t=1}^{T} \frac{\partial p_{it}}{\partial \beta_0} \\
\frac{\partial^2 \ell}{\partial \beta_1^2} &= -\sum_{i=1}^{n}\sum_{t=1}^{T} \left(E_{it}(\lambda_1,\lambda_2)\right)^2 p_{it}\left(1-p_{it}\right) \\
\frac{\partial^2 \ell}{\partial \lambda_1^2} &= \sum_{i=1}^{n}\sum_{t=1}^{T} \beta_1\left(Y_{it}-p_{it}\right)\left(\frac{\partial^2 E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_1^2}\right) - \beta_1\left(\frac{\partial E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_1}\right)\left(\frac{\partial p_{it}}{\partial \lambda_1}\right) \\
\frac{\partial^2 \ell}{\partial \lambda_2^2} &= \sum_{i=1}^{n}\sum_{t=1}^{T} \beta_1\left(Y_{it}-p_{it}\right)\left(\frac{\partial^2 E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_2^2}\right) - \beta_1\left(\frac{\partial E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_2}\right)\left(\frac{\partial p_{it}}{\partial \lambda_2}\right) \\
\frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} &= -\sum_{i=1}^{n}\sum_{t=1}^{T} \frac{\partial p_{it}}{\partial \beta_1} \\
\frac{\partial^2 \ell}{\partial \beta_0 \partial \lambda_1} &= -\sum_{i=1}^{n}\sum_{t=1}^{T} \frac{\partial p_{it}}{\partial \lambda_1} \\
\frac{\partial^2 \ell}{\partial \beta_0 \partial \lambda_2} &= -\sum_{i=1}^{n}\sum_{t=1}^{T} \frac{\partial p_{it}}{\partial \lambda_2} \\
\frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_2} &= \sum_{i=1}^{n}\sum_{t=1}^{T} \beta_1\left(Y_{it}-p_{it}\right)\left(\frac{\partial^2 E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_1 \partial \lambda_2}\right) - \beta_1\left(\frac{\partial E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_1}\right)\left(\frac{\partial p_{it}}{\partial \lambda_2}\right) \\
\frac{\partial^2 \ell}{\partial \beta_1 \partial \lambda_1} &= \sum_{i=1}^{n}\sum_{t=1}^{T} \left(Y_{it}-p_{it}\right)\left(\frac{\partial E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_1}\right) - E_{it}(\lambda_1,\lambda_2)\frac{\partial p_{it}}{\partial \lambda_1} \\
\frac{\partial^2 \ell}{\partial \beta_1 \partial \lambda_2} &= \sum_{i=1}^{n}\sum_{t=1}^{T} \left(Y_{it}-p_{it}\right)\left(\frac{\partial E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_2}\right) - E_{it}(\lambda_1,\lambda_2)\frac{\partial p_{it}}{\partial \lambda_2}
\end{aligned}
\tag{D.37}
$$

Derivatives of Probability Function

$$
\begin{aligned}
\frac{\partial p_{it}}{\partial \beta_0} &= p_{it}(1-p_{it}) \\
\frac{\partial p_{it}}{\partial \beta_1} &= \begin{cases} E_{it}(\lambda)p_{it}(1-p_{it}) & \text{for OPEE} \\ E_{it}(\lambda_1,\lambda_2)p_{it}(1-p_{it}) & \text{for TPEE} \end{cases} \\
\frac{\partial p_{it}}{\partial \lambda} &= \beta_1 p_{it}(1-p_{it})\frac{\partial E_{it}(\lambda)}{\partial \lambda} \\
\frac{\partial p_{it}}{\partial \lambda_1} &= \beta_1 p_{it}(1-p_{it})\frac{\partial E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_1} \\
\frac{\partial p_{it}}{\partial \lambda_2} &= \beta_1 p_{it}(1-p_{it})\frac{\partial E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_2}
\end{aligned}
\tag{D.38}
$$

Note: Moving from the OPEE to TPEE framework, $E_{it}(\lambda)$ can be substituted as $E_{it}(\lambda_1, \lambda_2)$, with corresponding substitutions made in $\frac{\partial p_{it}}{\partial \lambda_1}$ and $\frac{\partial p_{it}}{\partial \lambda_2}$ to account for the derivatives with respect to both lag parameters. Additionally, h and (h_1, h_2) should replace λ and (λ_1, λ_2) in the denominator of the partial derivative functions to obtain appropriate predicted probability estimates when calculating the Score and Information values for a model fit using the half–life parameterization.

DERIVATIONS: SINGLE DOSING, SINGLE LAG PARAMETER

Let D represent a specific steady state infusion dose that starts at $t = b$ and ends at $t = f$ for the set of for subject i. We can represent subject i's effective exposure, $E_{it}(\lambda)$ at time t by the following:

$$\begin{aligned} E_{it}(\lambda) &= D\left(1 - e^{-\lambda(t-b)}\right) I\left(t\epsilon\left[b, f\right]\right) + D\left(1 - e^{-\lambda(f-b)}\right) e^{-\lambda(t-f)} I\left(t > f\right) \\ &= D\left(e^{-\lambda \max(0,t-f)} - e^{-\lambda \max(0,t-b)}\right) = D\left(e^{-\lambda z_2} - e^{-\lambda z_1}\right) \end{aligned}$$

where

$$z_1 = \max(0, t-b) = \begin{cases} t-b & \text{if } t > b \\ 0 & \text{otherwise} \end{cases}$$

$$z_2 = \max(0, t-f) = \begin{cases} t-f & \text{if } t > f \\ 0 & \text{otherwise} \end{cases}$$

Let us recall that $E_{it}(\lambda)$ depends on the lag parameter, λ:

$$\begin{aligned} E_{it}(\lambda) &= D\left[e^{-\lambda z_2} - e^{-\lambda z_1}\right] \\ \frac{\partial E_{it}(\lambda)}{\partial \lambda} &= D\left[z_1 e^{-\lambda z_1} - z_2 e^{-\lambda z_2}\right] \\ \frac{\partial^2 E_{it}(\lambda)}{\partial \lambda^2} &= D\left[z_2^2 e^{-\lambda z_2} - z_1^2 e^{-\lambda z_1}\right] \end{aligned}$$

Choosing to parameterize using h instead of λ requires slightly more complicated equations for the first and second derivatives. The first component to identify is

the relationship between h and λ: $\lambda = \dfrac{\log 2}{h} \longrightarrow \dfrac{\partial \lambda}{\partial h} = \dfrac{-\log 2}{h^2}$

$$\begin{aligned}
\frac{\partial E_{it}(\lambda)}{\partial h} &= -z_2 \frac{\partial \lambda}{\partial h} D \exp\left(\frac{-z_2 \log 2}{h}\right) + z_1 \frac{\partial \lambda}{\partial h} D \exp\left(\frac{-z_1 \log 2}{h}\right) \\
&= \frac{z_2 \log 2}{h^2} D \exp\left(\frac{-z_2 \log 2}{h}\right) - \frac{z_1 \log 2}{h^2} D \exp\left(\frac{-z_1 \log 2}{h}\right) \\
&= \frac{\log 2}{h^2} D \left[z_2 e^{-z_2 \log 2/h} - z_1 e^{-z_1 \log 2/h}\right]
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 E_{it}(\lambda)}{\partial h^2} &= \frac{\partial \left(\dfrac{\log 2}{h^2} D \left[z_2 e^{-z_2 \log 2/h} - z_1 e^{-z_1 \log 2/h}\right]\right)}{\partial h} \\
&= \frac{-2 z_2 \log 2}{h^3} D e^{-z_2 \log 2/h} + \left[\frac{z_2 \log 2}{h^2}\right]^2 D e^{-z_2 \log 2/h} \\
&\quad - \left(\frac{-2 z_1 \log 2}{h^3} D e^{-z_1 \log 2/h} + \left[\frac{z_1 \log 2}{h^2}\right]^2 D e^{-z_1 \log 2/h}\right) \\
&= \frac{D \log 2}{h^3} \left(\left[\frac{z_2^2 \log 2}{h} - 2 z_2\right] e^{-z_2 \log 2/h} - \left[\frac{z_1^2 \log 2}{h} - 2 z_1\right] e^{-z_1 \log 2/h}\right)
\end{aligned}$$

POOLED LOGISTIC REGRESSION

Define the odds of an event for subject i at time t:

$$\exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right] = \exp\left[\beta_0 + \beta_1 D\left(e^{-\lambda z_2} - e^{-\lambda z_1}\right)\right]$$

Define the probability of an event for subject i at time t:

$$p_{it} = \frac{\exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]}{1 + \exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]} = \frac{\exp\left[\beta_0 + \beta_1 D\left(e^{-\lambda z_2} - e^{-\lambda z_1}\right)\right]}{1 + \exp\left[\beta_0 + \beta_1 D\left(e^{-\lambda z_2} - e^{-\lambda z_1}\right)\right]}$$

Likelihood function of logistic regression:

$$\mathcal{L}(\beta_0, \beta_1, \lambda, Y) = \prod_{i=1}^{n} \prod_{t=1}^{T} p_{it}^{Y_{ij}} (1 - p_{it})^{1 - Y_{it}}$$

Log-likelihood function for the logistic model:

$$\begin{aligned} \ell(\beta_0, \beta_1, \lambda, Y) &= \sum_{i=1}^{n} \sum_{t=1}^{T} Y_{it} \ln p_{it} + (1 - Y_{it}) \ln (1 - p_{it}) \\ &= \sum_{i=1}^{n} \sum_{t=1}^{T} Y_{it}\beta_0 + Y_{it}\beta_1 E_{it}(\lambda) - \ln\left(1 + \exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]\right) \end{aligned}$$

Score and Information Matrix Derivations

The following quantities can be used to solve for the asymptotic distribution of the lag parameter, in that the Score and Fisher's Information can be used to numerically estimate the lag, or for an approximate Hessian from which to pull an estimate of standard error. In all locations where $\frac{\partial E_{it}(\lambda)}{\partial \lambda}$ is used, one may substitute $\frac{\partial E_{it}(\lambda)}{\partial h}$ appropriately to converge towards estimates in the half-life lag-parameterization.

$$\begin{aligned}\frac{\partial p_{it}}{\partial \beta_0} &= \frac{\exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]}{\left(1 + \exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]\right)} - \left(\frac{\exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]}{\left(1 + \exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]\right)}\right)^2 \\ &= p_{it}(1 - p_{it})\end{aligned}$$

$$\begin{aligned}\frac{\partial p_{it}}{\partial \beta_1} &= E_{it}(\lambda)\left[\frac{\exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]}{\left(1 + \exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]\right)}\left(\frac{\exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]}{1 + \exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]}\right)^2\right] \\ &= D\left(e^{-\lambda z_2} - e^{-\lambda z_1}\right) p_{it}(1 - p_{it})\end{aligned}$$

$$\begin{aligned}\frac{\partial p_{it}}{\partial \lambda} &= \beta_1 \frac{\partial E_{it}(\lambda)}{\partial \lambda}\left[\frac{\exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]}{\left(1 + \exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]\right)} - \left(\frac{\exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]}{\left(1 + \exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]\right)}\right)^2\right] \\ &= \beta_1 \frac{\partial E_{it}(\lambda)}{\partial \lambda} p_{it}(1 - p_{it}) \\ &= \beta_1 D\left(z_1 e^{-\lambda z_1} - z_2 e^{-\lambda z_2}\right) p_{it}(1 - p_{it})\end{aligned}$$

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_0} &= \sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it} - \frac{\exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]}{1+\exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]} \\ &= \sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it} - p_{it}\end{aligned}$$

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_1} &= \sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it}E_{it}(\lambda) - \frac{E_{it}(\lambda)\exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]}{1+\exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]} \\ &= \sum_{i=1}^{n}\sum_{t=1}^{T} E_{it}(\lambda)\left(Y_{it} - p_{it}\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial \ell}{\partial \lambda} &= \sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it}\beta_1\frac{\partial E_{it}(\lambda)}{\partial \lambda} - \beta_1\frac{\partial E_{it}(\lambda)}{\partial \lambda}\frac{\exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]}{1+\exp\left[\beta_0 + \beta_1 E_{it}(\lambda)\right]} \\ &= \sum_{i=1}^{n}\sum_{t=1}^{T} \beta_1\frac{\partial E_{it}(\lambda)}{\partial \lambda}\left(Y_{it} - p_{it}\right) \\ &= \sum_{i=1}^{n}\sum_{t=1}^{T} \beta_1\left(Y_{it} - p_{it}\right)D\left(z_1 e^{-\lambda z_1} - z_2 e^{-\lambda z_2}\right)\end{aligned}$$

$$\frac{\partial^2 \ell}{\partial \beta_0^2} = \frac{\partial\left(\frac{\partial \ell}{\partial \beta_0}\right)}{\partial \beta_0} = \frac{\partial\left(\sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it} - p_{it}\right)}{\partial \beta_0} = -\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\partial p_{it}}{\partial \beta_0}$$
$$= -\sum_{i=1}^{n}\sum_{t=1}^{T} p_{it}\left(1 - p_{it}\right)$$

$$\frac{\partial^2 \ell}{\partial \beta_1^2} = \frac{\partial\left(\frac{\partial \ell}{\partial \beta_1}\right)}{\partial \beta_1} = \frac{\partial\left(\sum_{i=1}^{n}\sum_{t=1}^{T} E_{it}(\lambda)\left(Y_{it} - p_{it}\right)\right)}{\partial \beta_1} = -\sum_{i=1}^{n}\sum_{t=1}^{T} E_{it}(\lambda)\frac{\partial p_{it}}{\partial \beta_1}$$
$$= -\sum_{i=1}^{n}\sum_{t=1}^{T} E_{it}(\lambda)^2 p_{it}\left(1 - p_{it}\right)$$

$$\frac{\partial^2 \ell}{\partial \lambda^2} = \frac{\partial\left[\frac{\partial \ell}{\partial \lambda}\right]}{\partial \lambda} = \frac{\partial\left[\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\frac{\partial E_{it}(\lambda)}{\partial \lambda}\right)\beta_1\left(Y_{it} - p_{it}\right)\right]}{\partial \lambda}$$
$$= \sum_{i=1}^{n}\sum_{t=1}^{T}\left(\frac{\partial^2 E_{it}(\lambda)}{\partial \lambda^2}\right)\beta_1\left(Y_{it} - p_{it}\right) - \left(\frac{\partial E_{it}(\lambda)}{\partial \lambda}\right)\beta_1\left(\frac{\partial p_{it}}{\partial \lambda}\right)$$
$$= \sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\partial^2 E_{it}(\lambda)}{\partial \lambda^2}\beta_1\left(Y_{it} - p_{it}\right) - \sum_{i=1}^{n}\sum_{t=1}^{T}\left(\frac{\partial E_{it}(\lambda)}{\partial \lambda}\right)^2 \beta_1^2 p_{it}\left(1 - p_{it}\right)$$
$$= \sum_{i=1}^{n}\sum_{t=1}^{T}\beta_1\left(Y_{it} - p_{it}\right) D\left([z_2]^2 e^{-\lambda z_2} - [z_1]^2 e^{-\lambda z_1}\right)$$
$$- \sum_{i=1}^{n}\sum_{t=1}^{T}\beta_1^2 p_{it}\left(1 - p_{it}\right) D^2\left[z_1 e^{-\lambda z_1} - z_2 e^{-\lambda z_2}\right]^2$$

$$\frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} = \frac{\partial\left(\frac{\partial \ell}{\partial \beta_0}\right)}{\partial \beta_1} = \frac{\partial\left(\sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it} - p_{it}\right)}{\partial \beta_1} = -\sum_{i=1}^{n}\sum_{t=1}^{T} \frac{\partial p_{it}}{\partial \beta_1}$$
$$= -\sum_{i=1}^{n}\sum_{t=1}^{T} E_{it}(\lambda) p_{it}\left(1 - p_{it}\right)$$

$$\frac{\partial^2 \ell}{\partial \beta_0 \partial \lambda} = \frac{\partial\left(\frac{\partial \ell}{\partial \beta_0}\right)}{\partial \lambda} = \frac{\partial\left(\sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it} - p_{it}\right)}{\partial \lambda} = -\sum_{i=1}^{n}\sum_{t=1}^{T} \frac{\partial p_{it}}{\partial \lambda}$$
$$= -\sum_{i=1}^{n}\sum_{t=1}^{T} \beta_1 \frac{\partial E_{it}(\lambda)}{\partial \lambda} p_{it}\left(1 - p_{it}\right)$$
$$= -\sum_{i=1}^{n}\sum_{t=1}^{T} \beta_1 p_{it}\left(1 - p_{it}\right) D\left(z_1 e^{-\lambda z_1} - z_2 e^{-\lambda z_2}\right)$$

$$\frac{\partial^2 \ell}{\partial \beta_1 \partial \lambda} = \frac{\partial\left(\frac{\partial \ell}{\partial \beta_1}\right)}{\partial \lambda} = \frac{\partial\left(\sum_{i=1}^{n}\sum_{t=1}^{T} E_{it}(\lambda) Y_{it} - E_{it}(\lambda) p_{it}\right)}{\partial \lambda}$$
$$= \sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it} \frac{\partial E_{it}(\lambda)}{\partial \lambda} - p_{it} \frac{\partial E_{it}(\lambda)}{\partial \lambda} - E_{it}(\lambda) \frac{\partial p_{it}}{\partial \lambda}$$
$$= \sum_{i=1}^{n}\sum_{t=1}^{T} \frac{\partial E_{it}(\lambda)}{\partial \lambda} \left[Y_{it} - p_{it} - \beta_1 E_{it}(\lambda) p_{it}\left(1 - p_{it}\right)\right]$$
$$= \sum_{i=1}^{n}\sum_{t=1}^{T} \left[Y_{it} - p_{it} - \beta_1 p_{it}\left(1 - p_{it}\right) D\left(e^{-\lambda z_2} - e^{-\lambda z_1}\right)\right]$$
$$\times D\left(z_1 e^{-\lambda z_1} - z_2 e^{-\lambda z_2}\right)$$

COX PROPORTIONAL HAZARDS

Likelihood Function with Breslow's handling of tied event times:

$$\mathcal{L}(\beta, \lambda, Y) = \prod_{k=1}^{K} \frac{\prod_{j \epsilon \mathcal{R}(t_k, Y_j=1)} e^{\beta E_j(\lambda)}}{\left(\sum_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)} \right)^{m_k}}$$

where m_k is the total number of events at time k, $\mathcal{R}(t_k)$ is the set of subjects at risk at time t_k, and where the numerator for each unique event time, t_k, is the product of the exponential risk of event, $e^{\beta E_j(\lambda)}$, for all subjects with events at time t_k as noted by $\mathcal{R}(t_k, Y_{t_k} = 1)$.

Thus, the log-likelihood function takes the form:

$$\begin{aligned} \ell(\beta, \lambda, Y) &= \sum_{k=1}^{K} \left[\sum_{j \epsilon \mathcal{R}(t_k, Y_j=1)} \beta E_j(\lambda) - m_k \log \left(\sum_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)} \right) \right] \\ &= \sum_{k=1}^{K} \left[A_1(t_k) - m_k \log \left(C_1(t_k) \right) \right] \end{aligned}$$

Two important notes to remember about the Cox likelihood function:

1. The sum from $k = 1$ to K imply a risk set be defined by unique stop-time and strata, since the assumption is being made that the baseline hazard is different across strata.

2. The form $\beta E_j(\lambda)$ is a stand-in for the risk score of an individual, and would, more correctly, be written as $\beta E_j(\lambda) + \sum \Gamma X_{jq}$, in a multivariate model with $q = 1, ..., Q$ covariates.

Score and Information Matrix Derivations

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta} &= \sum_{k=1}^{K} \left[\sum_{j \epsilon \mathcal{R}(t_k, Y_j = 1)} E_j(\lambda) - m_k \frac{\sum\limits_{j \epsilon \mathcal{R}(t_k)} E_j(\lambda) e^{\beta E_j(\lambda)}}{\sum\limits_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)}} \right] \\
&= \sum_{k=1}^{K} \left[A_{2\beta}(t_k) - m_k \frac{B_1(t_k)}{C_1(t_k)} \right] \\
\frac{\partial \ell}{\partial \lambda} &= \sum_{k=1}^{K} \left[\beta \sum_{j \epsilon \mathcal{R}(t_k, Y_j = 1)} \frac{\partial E_j(\lambda)}{\partial \lambda} - m_k \frac{\sum\limits_{j \epsilon \mathcal{R}(t_k)} \beta \frac{\partial E_j(\lambda)}{\partial \lambda} e^{\beta E_j(\lambda)}}{\sum\limits_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)}} \right] \\
&= \sum_{k=1}^{K} \left[A_{2\lambda}(t_k) - m_k \frac{C_2(t_k)}{C_1(t_k)} \right] \\
\frac{\partial^2 \ell}{\partial \beta^2} &= \frac{\partial \left(\frac{\partial \ell}{\partial \beta} \right)}{\partial \beta} = \frac{\partial \left(\sum\limits_{k=1}^{K} \left[\sum\limits_{j \epsilon \mathcal{R}(t_k, Y_j = 1)} E_j(\lambda) - m_k \frac{\sum\limits_{j \epsilon \mathcal{R}(t_k)} E_j(\lambda) e^{\beta E_j(\lambda)}}{\sum\limits_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)}} \right] \right)}{\partial \beta} \\
&= -\sum_{k=1}^{K} m_k \frac{\left(\sum\limits_{j \epsilon \mathcal{R}(t_k)} E_j(\lambda)^2 e^{\beta E_j(\lambda)} \right) \left(\sum\limits_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)} \right)}{\left(\sum\limits_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)} \right)^2} \\
&\quad + \sum_{k=1}^{K} m_k \frac{\left(\sum\limits_{j \epsilon \mathcal{R}(t_k)} E_j(\lambda) e^{\beta E_j(\lambda)} \right) \left(\sum\limits_{j \epsilon \mathcal{R}(t_k)} E_j(\lambda) e^{\beta E_j(\lambda)} \right)}{\left(\sum\limits_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)} \right)^2} \\
&= \sum_{k=1}^{K} m_k \frac{\left[B_1(t_k) \right]^2 - B_{2\beta}(t_k) C_1(t_k)}{\left[C_1(t_k) \right]^2}
\end{aligned}$$

$$\frac{\partial^2 \ell}{\partial \lambda^2} = \frac{\partial \left(\sum_{k=1}^{K} \left[\beta \sum_{j \epsilon \mathcal{R}(t_k, Y_j=1)} \frac{\partial E E_k}{\partial \lambda} - m_k \frac{\sum_{j \epsilon \mathcal{R}(t_k)} \beta \frac{\partial E_j(\lambda)}{\partial \lambda} e^{\beta E_j(\lambda)}}{\sum_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)}} \right] \right)}{\partial \lambda}$$

$$= \beta \sum_{k=1}^{K} \sum_{j \epsilon \mathcal{R}(t_k, Y_j=1)} \frac{\partial^2 E_j(\lambda)}{\partial \lambda^2} + \sum_{k=1}^{K} m_k \frac{\left(\sum_{j \epsilon \mathcal{R}(t_k)} \beta \frac{\partial E_j(\lambda)}{\partial \lambda} e^{\beta E_j(\lambda)} \right)^2}{\left(\sum_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)} \right)^2}$$

$$- \sum_{k=1}^{k} m_k \frac{\left(\sum_{j \epsilon \mathcal{R}(t_k)} \beta e^{\beta E_j(\lambda)} \left[\frac{\partial^2 E_j(\lambda)}{\partial \lambda^2} + \beta \left(\frac{\partial E_j(\lambda)}{\partial \lambda} \right)^2 \right] \right) \left(\sum_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)} \right)}{\left(\sum_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)} \right)^2}$$

$$= \sum_{k=1}^{K} \left[A_3(t_k) - m_k \frac{C_3(t_k) C_1(t_k) - (C_2(t_k))^2}{(C_1(t_k))^2} \right]$$

$$\frac{\partial^2 \ell}{\partial \beta \partial \lambda} = \frac{\partial \left(\sum_{k=1}^{K} \left[\sum_{j \epsilon \mathcal{R}(t_k, Y_j=1)} E_j(\lambda) - m_k \frac{\sum_{j \epsilon \mathcal{R}(t_k)} E_j(\lambda) e^{\beta E_j(\lambda)}}{\sum_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)}} \right] \right)}{\partial \lambda}$$

$$= \sum_{k=1}^{K} \sum_{j \epsilon \mathcal{R}(t_k, Y_j=1)} \frac{\partial E_j(\lambda)}{\partial \lambda}$$

$$- \sum_{k=1}^{K} m_k \frac{\left(\sum_{j \epsilon \mathcal{R}(t_k)} \frac{\partial E_j(\lambda)}{\partial \lambda} e^{\beta E_j(\lambda)} \left[1 + \beta E_j(\lambda) \right] \right) \left(\sum_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)} \right)}{\left(\sum_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)} \right)^2}$$

$$+ \sum_{k=1}^{K} m_k \frac{\left(\sum_{j \epsilon \mathcal{R}(t_k)} \beta e^{\beta E_j(\lambda)} \frac{\partial E_j(\lambda)}{\partial \lambda} \right) \left(\sum_{j \epsilon \mathcal{R}(t_k)} E_j(\lambda) e^{\beta E_j(\lambda)} \right)}{\left(\sum_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)} \right)^2}$$

$$= \sum_{k=1}^{K} \left[\frac{A_{2\lambda}(t_k)}{\beta} - m_k \frac{B_{2\lambda}(t_k) C_1(t_k) - C_2(t_k) B_1(t_k)}{(C_1(t_k))^2} \right]$$

Where we define the following compute-able quantities[2]:

- $A_1(t_k) = \sum_{j \epsilon \mathcal{R}(t_k, Y_j=1)} \beta E_j(\lambda)$
- $A_{2\beta}(t_k) = \frac{\partial A_1(t_k)}{\partial \beta} = \sum_{j \epsilon \mathcal{R}(t_k, Y_j=1)} E_j(\lambda)$
- $A_{2\lambda}(t_k) = \frac{\partial A_1(t_k)}{\partial \lambda} = \sum_{j \epsilon \mathcal{R}(t_k, Y_j=1)} \beta \frac{\partial E_j(\lambda)}{\partial \lambda}$
- $A_3(t_k) = \frac{\partial^2 A_1(t_k)}{(\partial \lambda)^2} = \sum_{j \epsilon \mathcal{R}(t_k, Y_j=1)} \beta \frac{\partial^2 E_j(\lambda)}{\partial \lambda^2}$
- $B_1(t_k) = \sum_{j \epsilon \mathcal{R}(t_k)} E_j(\lambda) e^{\beta E_j(\lambda)}$
- $B_{2\beta}(t_k) = \frac{\partial B_1(t_k)}{\partial \beta} = \sum_{j \epsilon \mathcal{R}(t_k)} E_j(\lambda)^2 e^{\beta E_j(\lambda)}$
- $B_{2\lambda}(t_k) = \frac{\partial B_1(t_k)}{\partial \lambda} = \sum_{j \epsilon \mathcal{R}(t_k)} \frac{\partial E_j(\lambda)}{\partial \lambda} e^{\beta E_j(\lambda)} \left[1 + \beta E_j(\lambda)\right]$
- $C_1(t_k) = \sum_{j \epsilon \mathcal{R}(t_k)} e^{\beta E_j(\lambda)}$
- $C_2(t_k) = \frac{\partial C_1(t_k)}{\partial \lambda} = \sum_{j \epsilon \mathcal{R}(t_k)} \beta \frac{\partial E_j(\lambda)}{\partial \lambda} e^{\beta E_j(\lambda)}$
- $C_3(t_k) = \frac{\partial^2 C_1(t_k)}{(\partial \lambda)^2} = \sum_{j \epsilon \mathcal{R}(t_k)} \beta e^{\beta E_j(\lambda)} \left[\frac{\partial^2 E_j(\lambda)}{\partial \lambda^2} + \beta \left(\frac{\partial E_j(\lambda)}{\partial \lambda} \right)^2 \right]$

[2] As these functions utilize the first and second derivatives of the effective exposure, one can switch between the λ and h by adjusting which internal formulas to use, maintaining consistency once estimation has started.

DERIVATIONS: SINGLE DOSING, TWO LAG PARAMETERS

Let D represent a specific steady state infusion dose that starts at $t = b$ and ends at $t = f$ for the set of for subject i. We can represent subject i's effective exposure, $E_{it}(\lambda_1, \lambda_2)$ at time t by the following:

$$\begin{aligned} E_{it}(\lambda_1, \lambda_2) &= D\left(1 - e^{-\lambda_1(t-b)}\right) I\left(t\epsilon\left[b, f\right]\right) + D\left(1 - e^{-\lambda_1(f-b)}\right) e^{-\lambda_2(t-f)} I\left(t > f\right) \\ &= D\left(1 - e^{-\lambda_1(max(0,t-b)-max(0,t-f))}\right) e^{-\lambda_2(max(0,t-f))} \\ &= D\left(1 - e^{-\lambda_1(z_1-z_2)}\right) e^{-\lambda_2 z_2} \end{aligned}$$

where

$$z_1 = \max(0, t-b) = \begin{cases} t-b & \text{if } t > b \\ 0 & \text{otherwise} \end{cases}$$

$$z_2 = \max(0, t-f) = \begin{cases} t-f & \text{if } t > f \\ 0 & \text{otherwise} \end{cases}$$

Let us recall that $E_{it}(\lambda_1, \lambda_2)$ depends on the lag parameters, $\lambda_1$and λ_2:

$$\frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1} = \left(z_1 - z_2\right) D e^{-\lambda_1(z_1-z_2)} e^{-\lambda_2 z_2}$$

$$\frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_2} = -z_2 D\left(1 - e^{-\lambda_1(z_1-z_2)}\right) e^{-\lambda_2 z_2}$$

$$\frac{\partial^2 E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1^2} = -\left(z_1 - z_2\right)^2 D e^{-\lambda_1(z_1-z_2)} e^{-\lambda_2 z_2}$$

$$\frac{\partial^2 E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_2^2} = z_2^2 D\left(1-e^{-\lambda_1(z_1-z_2)}\right)e^{-\lambda_2 z_2}$$

$$\frac{\partial^2 E_{it}(\lambda_1,\lambda_2)}{\partial \lambda_1 \partial \lambda_2} = -\left(z_1-z_2\right)z_2 D e^{-\lambda_1(z_1-z_2)}e^{-\lambda_2 z_2}$$

$$\frac{\partial E_{it}(\lambda_1,\lambda_2)}{\partial h_1} = \left(z_1-z_2\right)\frac{-\log 2}{{h_1}^2} D e^{-(z_1-z_2)\log 2/h_1}e^{-z_2\log 2/h_2}$$

$$\frac{\partial E_{it}(\lambda_1,\lambda_2)}{\partial h_2} = z_2\frac{\log 2}{{h_2}^2} D\left(1-e^{-(z_1-z_2)\log 2/h_1}\right)e^{-z_2\log 2/h_2}$$

$$\frac{\partial^2 E_{it}(\lambda_1,\lambda_2)}{\partial {h_1}^2} = \frac{\left(z_1-z_2\right)D\log 2}{{h_1}^2}e^{-(z_1-z_2)\log 2/h_1}e^{-z_2\log 2/h_2}\left(\frac{2}{h_1}-\frac{\left(z_1-z_2\right)\log 2}{{h_1}^2}\right)$$

$$\frac{\partial^2 E_{it}(\lambda_1,\lambda_2)}{\partial {h_2}^2} = \frac{z_2 D\log 2}{{h_2}^2}\left(1-e^{-(z_1-z_2)\log 2/h_1}\right)e^{-z_2\log 2/h_2}\left(\frac{2}{h_2}-\frac{z_2\log 2}{{h_1}^2}\right)$$

$$\frac{\partial^2 E_{it}(\lambda_1,\lambda_2)}{\partial h_1 \partial h_2} = -\left(z_1-z_2\right)z_2\frac{\left(\log 2\right)^2}{{h_1}^2{h_2}^2} D e^{-(z_1-z_2)\log 2/h_1}e^{-z_2\log 2/h_2}$$

POOLED LOGISTIC REGRESSION

Similar to the One-Parameter, we now substitute $E_{it}(\lambda)$ with $E_{it}(\lambda_1, \lambda_2)$ to define the odds of an event for subject i at time t:

$$\exp\left[\beta_0 + \beta_1 E_{it}(\lambda_1, \lambda_2)\right] = \exp\left[\beta_0 + \beta_1 D\left(1 - e^{-\lambda_1(z_1 - z_2)}\right) e^{-\lambda_2 z_2}\right]$$

Where now the probability of an event for subject i at time t:

$$\begin{aligned} p_{it} &= \frac{\exp\left[\beta_0 + \beta_1 E_{it}(\lambda_1, \lambda_2)\right]}{1 + \exp\left[\beta_0 + \beta_1 E_{it}(\lambda_1, \lambda_2)\right]} \\ &= \frac{\exp\left[\beta_0 + \beta_1 D\left(1 - e^{-\lambda_1(z_1 - z_2)}\right) e^{-\lambda_2 z_2}\right]}{1 + \exp\left[\beta_0 + \beta_1 D\left(1 - e^{-\lambda_1(z_1 - z_2)}\right) e^{-\lambda_2 z_2}\right]} \end{aligned}$$

Likelihood function of logistic regression:

$$\mathcal{L}(\beta, \lambda_1, \lambda_2, Y) = \prod_{i=1}^{n} \prod_{t=1}^{T} p_{it}^{Y_{ij}} (1 - p_{it})^{1 - Y_{it}}$$

Log-likelihood function for the logistic model:

$$\begin{aligned} \ell(\beta, \lambda_1, \lambda_2, Y) &= \sum_{i=1}^{n} \sum_{t=1}^{T} Y_{it} \ln p_{it} + (1 - Y_{it}) \ln\left(1 - p_{it}\right) \\ &= \sum_{i=1}^{n} \sum_{t=1}^{T} Y_{it}\beta_0 + Y_{it}\beta_1 E_{it}(\lambda_1, \lambda_2) - \ln\left(1 + \exp\left[\beta_0 + \beta_1 E_{it}(\lambda_1, \lambda_2)\right]\right) \end{aligned}$$

$$\frac{\partial p_{it}}{\partial \beta_1} = E_{it}(\lambda_1, \lambda_2) p_{it}(1 - p_{it})$$

$$\begin{aligned}
\frac{\partial p_{it}}{\partial \lambda_1} &= \beta_1 \frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1} p_{it}(1 - p_{it}) \\
&= (z_1 - z_2)\, \beta_1 D e^{-\lambda_1 (z_1 - z_2)} e^{-\lambda_2 z_2} p_{it}(1 - p_{it}) \\
&= (z_1 - z_2)\, \beta_1 D e^{-\lambda_1 (z_1 - z_2) - \lambda_2 z_2} p_{it}(1 - p_{it})
\end{aligned}$$

$$\begin{aligned}
\frac{\partial p_{it}}{\partial \lambda_2} &= \beta_1 \frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_2} p_{it}(1 - p_{it}) \\
&= -z_2 \beta_1 D \left(1 - e^{-\lambda_1 (z_1 - z_2)}\right) e^{-\lambda_2 z_2} p_{it}(1 - p_{it}) \\
&= -z_2 \beta_1 D \left(e^{-\lambda_2 z_2} - e^{-\lambda_1 (z_1 - z_2) - \lambda_2 z_2}\right) p_{it}(1 - p_{it})
\end{aligned}$$

Score and Information Matrix Derivations

$$\frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^{n} \sum_{t=1}^{T} E_{it}(\lambda_1, \lambda_2) \left(Y_{it} - p_{it}\right)$$

$$\begin{aligned}\frac{\partial \ell}{\partial \lambda_1} &= \sum_{i=1}^{n} \sum_{t=1}^{T} \beta_1 \frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1} \left(Y_{it} - p_{it}\right) \\ &= \sum_{i=1}^{n} \sum_{t=1}^{T} \left(z_1 - z_2\right) \beta_1 D e^{-\lambda_1 (z_1 - z_2) - \lambda_2 z_2} \left(Y_{it} - p_{it}\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial \ell}{\partial \lambda_2} &= \sum_{i=1}^{n} \sum_{t=1}^{T} \beta_1 \frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_2} \left(Y_{it} - p_{it}\right) \\ &= \sum_{i=1}^{n} \sum_{t=1}^{T} -z_2 \beta_1 D \left(e^{-\lambda_2 z_2} - e^{-\lambda_1 (z_1 - z_2) - \lambda_2 z_2}\right) \left(Y_{it} - p_{it}\right)\end{aligned}$$

$$\frac{\partial^2 \ell}{\partial \beta_1^2} = -\sum_{i=1}^{n} \sum_{t=1}^{T} E_{it}(\lambda_1, \lambda_2)^2 p_{it} \left(1 - p_{it}\right)$$

$$\frac{\partial^2 \ell}{\partial \lambda_1^2} = \frac{\partial \left[\frac{\partial \ell}{\partial \lambda_1} \right]}{\partial \lambda_1} = \frac{\partial \left[\sum_{i=1}^{n} \sum_{t=1}^{T} \left(\frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1} \right) \beta_1 \left(Y_{it} - p_{it} \right) \right]}{\partial \lambda_1}$$

$$= \sum_{i=1}^{n} \sum_{t=1}^{T} \left(\frac{\partial^2 E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1^2} \right) \beta_1 \left(Y_{it} - p_{it} \right) - \left(\frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1} \right) \beta_1 \left(\frac{\partial p_{it}}{\partial \lambda_1} \right)$$

$$= \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{\partial^2 E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1^2} \beta_1 \left(Y_{it} - p_{it} \right) - \sum_{i=1}^{n} \sum_{t=1}^{T} \left(\frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1} \right)^2 \beta_1^2 p_{it} \left(1 - p_{it} \right)$$

$$= \sum_{i=1}^{n} \sum_{t=1}^{T} - \left(z_1 - z_2 \right)^2 \beta_1 D e^{-\lambda_1 (z_1 - z_2) - \lambda_2 z_2} \left(Y_{it} - p_{it} \right)$$
$$- \sum_{i=1}^{n} \sum_{t=1}^{T} \left[\left(z_1 - z_2 \right) D e^{-\lambda_1 (z_1 - z_2) - \lambda_2 z_2} \right]^2 \beta_1^2 p_{it} \left(1 - p_{it} \right)$$

$$\frac{\partial^2 \ell}{\partial \lambda_2^2} = \frac{\partial \left[\frac{\partial \ell}{\partial \lambda_2} \right]}{\partial \lambda_2} = \frac{\partial \left[\sum_{i=1}^{n} \sum_{t=1}^{T} \left(\frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_2} \right) \beta_1 \left(Y_{it} - p_{it} \right) \right]}{\partial \lambda_2}$$

$$= \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{\partial^2 E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_2^2} \beta_1 \left(Y_{it} - p_{it} \right) - \sum_{i=1}^{n} \sum_{t=1}^{T} \left(\frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_2} \right)^2 \beta_1^2 p_{it} \left(1 - p_{it} \right)$$

$$= \sum_{i=1}^{n} \sum_{t=1}^{T} z_2^2 \beta_1 D \left(e^{-\lambda_2 z_2} - e^{-\lambda_1 (z_1 - z_2) - \lambda_2 z_2} \right) \left(Y_{it} - p_{it} \right)$$
$$- \sum_{i=1}^{n} \sum_{t=1}^{T} \left[-z_2 D \left(e^{-\lambda_2 z_2} - e^{-\lambda_1 (z_1 - z_2) - \lambda_2 z_2} \right) \right]^2 \beta_1^2 p_{it} \left(1 - p_{it} \right)$$

$$\frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_2} = \frac{\partial \left[\frac{\partial \ell}{\partial \lambda_1} \right]}{\partial \lambda_2} = \frac{\partial \left[\sum_{i=1}^{n} \sum_{t=1}^{T} \left(\frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1} \right) \beta_1 \left(Y_{it} - p_{it} \right) \right]}{\partial \lambda_2}$$

$$= \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{\partial^2 E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1 \partial \lambda_2} \beta_1 \left(Y_{it} - p_{it} \right) - \left(\frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1} \right) \beta_1 \left(\frac{\partial p_{it}}{\partial \lambda_2} \right)$$

$$= \sum_{i=1}^{n} \sum_{t=1}^{T} -z_2 \left(z_1 - z_2 \right) D e^{-\lambda_1 (z_1 - z_2) - \lambda_2 z_2} \beta_1 \left(Y_{it} - p_{it} \right)$$

$$+ \sum_{i=1}^{n} \sum_{t=1}^{T} z_2 \left(z_1 - z_2 \right) \beta_1^2 D \left(e^{-\lambda_2 z_2} - e^{-\lambda_1 (z_1 - z_2) - \lambda_2 z_2} \right) p_{it} \left(1 - p_{it} \right)$$

$$\frac{\partial^2 \ell}{\partial \beta_1 \partial \lambda_1} = \frac{\partial \left(\frac{\partial \ell}{\partial \beta_1} \right)}{\partial \lambda_1} = \frac{\partial \left(\sum_{i=1}^{n} \sum_{t=1}^{T} E_{it}(\lambda_1, \lambda_2) Y_{it} - E_{it}(\lambda_1, \lambda_2) p_{it} \right)}{\partial \lambda_1}$$

$$= \sum_{i=1}^{n} \sum_{t=1}^{T} Y_{it} \frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1} - p_{it} \frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1} - E_{it}(\lambda_1, \lambda_2) \frac{\partial p_{it}}{\partial \lambda_1}$$

$$= \sum_{i=1}^{n} \sum_{t=1}^{T} \left(z_1 - z_2 \right) D e^{-\lambda_1 (z_1 - z_2) - \lambda_2 z_2} \left(Y_{it} - p_{it} \right)$$

$$- \sum_{i=1}^{n} \sum_{t=1}^{T} \left(z_1 - z_2 \right) \left[D e^{-\lambda_1 (z_1 - z_2) - \lambda_2 z_2} \right]^2 \beta_1 p_{it} \left(1 - p_{it} \right)$$

$$\frac{\partial^2 \ell}{\partial \beta_1 \partial \lambda_2} = \frac{\partial \left(\frac{\partial \ell}{\partial \beta_1} \right)}{\partial \lambda_2} = \frac{\partial \left(\sum_{i=1}^{n} \sum_{t=1}^{T} E_{it}(\lambda_1, \lambda_2) Y_{it} - E_{it}(\lambda_1, \lambda_2) p_{it} \right)}{\partial \lambda_2}$$

$$= \sum_{i=1}^{n} \sum_{t=1}^{T} Y_{it} \frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_2} - p_{it} \frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_2} - E_{it}(\lambda_1, \lambda_2) \frac{\partial p_{it}}{\partial \lambda_2}$$

$$= \sum_{i=1}^{n} \sum_{t=1}^{T} -z_2 D \left(e^{-\lambda_2 z_2} - e^{-\lambda_1 (z_1 - z_2) - \lambda_2 z_2} \right) \left(Y_{it} - p_{it} \right)$$

$$+ \sum_{i=1}^{n} \sum_{t=1}^{T} z_2 \left[D \left(e^{-\lambda_2 z_2} - e^{-\lambda_1 (z_1 - z_2) - \lambda_2 z_2} \right) \right]^2 \beta_1 p_{it} \left(1 - p_{it} \right)$$

COX PROPORTIONAL HAZARDS

For the Cox Proportional Hazards model, the likelihood and log-likelihood functions take the same general form as the one-parameter lag equations. Also, the partial first and second derivatives of the log-likelihood with respect to β, alone, remain unchanged. The equations that follow complete the forms needed to derive the Score and Hessian matrices for the two λ parameters. We add definitions for the following compute-able quantities:

$$A_4(t_i) = \frac{\partial A_1(t_i)}{\partial \lambda_1} = \sum_{j \epsilon \mathcal{R}(t_i, Y_j = 1)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1}$$

$$A_5(t_i) = \frac{\partial A_1(t_i)}{\partial \lambda_2} = \sum_{j \epsilon \mathcal{R}(t_i, Y_j = 1)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2}$$

$$A_6(t_i) = \frac{\partial^2 A_1(t_i)}{\partial \lambda_1^2} = \sum_{j \epsilon \mathcal{R}(t_i, Y_j = 1)} \beta \frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_1^2}$$

$$A_7(t_i) = \frac{\partial^2 A_1(t_i)}{\partial \lambda_2^2} = \sum_{j \epsilon \mathcal{R}(t_i, Y_j = 1)} \beta \frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_2^2}$$

$$A_8(t_i) = \frac{\partial^2 A_1(t_i)}{\partial \lambda_1 \partial \lambda_2} = \sum_{j \epsilon \mathcal{R}(t_i, Y_j = 1)} \beta \frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_1 \partial \lambda_2}$$

$$B_{2\lambda_1}(t_i) = \frac{\partial B_1(t_i)}{\partial \lambda_1} = \sum_{j \epsilon \mathcal{R}(t_i)} \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} e^{\beta E_j(\lambda_1, \lambda_2)} \left[1 + \beta E_j(\lambda_1, \lambda_2)\right]$$

$$B_{2\lambda_2}(t_i) = \frac{\partial B_1(t_i)}{\partial \lambda_2} = \sum_{j \epsilon \mathcal{R}(t_i)} \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} e^{\beta E_j(\lambda_1, \lambda_2)} \left[1 + \beta E_j(\lambda_1, \lambda_2)\right]$$

$$C_4(t_i) = \frac{\partial C_1(t_i)}{\partial \lambda_1} = \sum_{j \epsilon \mathcal{R}(t_i)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} e^{\beta E_j(\lambda_1, \lambda_2)}$$

$$C_5(t_i) = \frac{\partial C_1(t_i)}{\partial \lambda_2} = \sum_{j \epsilon \mathcal{R}(t_i)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} e^{\beta E_j(\lambda_1, \lambda_2)}$$

$$C_6(t_i) = \frac{\partial^2 C_1(t_i)}{\partial \lambda_1^2} = \sum_{j \epsilon \mathcal{R}(t_i)} \beta e^{\beta E_j(\lambda_1, \lambda_2)} \left[\frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_1^2} + \beta \left(\frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} \right)^2 \right]$$

$$C_7(t_i) = \frac{\partial^2 C_1(t_i)}{\partial \lambda_2^2} = \sum_{j \epsilon \mathcal{R}(t_i)} \beta e^{\beta E_j(\lambda_1, \lambda_2)} \left[\frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_2^2} + \beta \left(\frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} \right)^2 \right]$$

$$C_8(t_i) = \frac{\partial^2 C_1(t_i)}{\partial \lambda_1 \partial \lambda_2} = \sum_{j \epsilon \mathcal{R}(t_i)} \beta e^{\beta E_j(\lambda_1, \lambda_2)} \left[\frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_1 \partial \lambda_2} + \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} \right]$$

$$\begin{aligned} \frac{\partial \ell}{\partial \lambda_1} &= \sum_{i=1}^{I} \left[\beta \sum_{j \epsilon \mathcal{R}(t_i, Y_j = 1)} \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} - m_i \frac{\sum_{j \epsilon \mathcal{R}(t_i)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} e^{\beta E_j(\lambda_1, \lambda_2)}}{\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)}} \right] \\ &= \sum_{i=1}^{I} \left[A_4(t_i) - m_i \frac{C_4(t_i)}{C_1(t_i)} \right] \end{aligned}$$

$$\begin{aligned} \frac{\partial \ell}{\partial \lambda_2} &= \sum_{i=1}^{I} \left[\beta \sum_{j \epsilon \mathcal{R}(t_i, Y_j = 1)} \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} - m_i \frac{\sum_{j \epsilon \mathcal{R}(t_i)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} e^{\beta E_j(\lambda_1, \lambda_2)}}{\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)}} \right] \\ &= \sum_{i=1}^{I} \left[A_5(t_i) - m_i \frac{C_5(t_i)}{C_1(t_i)} \right] \end{aligned}$$

$$\frac{\partial^2 \ell}{\partial \lambda_1^2} = \frac{\partial \left(\sum_{i=1}^{I} \left[\beta \sum_{j \epsilon \mathcal{R}(t_i, Y_j=1)} \frac{\partial EE_i}{\partial \lambda_1} - m_i \frac{\sum_{j \epsilon \mathcal{R}(t_i)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} e^{\beta E_j(\lambda_1, \lambda_2)}}{\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)}} \right] \right)}{\partial \lambda_1}$$

$$= \beta \sum_{i=1}^{I} \sum_{j \epsilon \mathcal{R}(t_i, Y_j=1)} \frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_1^2} + \sum_{i=1}^{I} m_i \frac{\left(\sum_{j \epsilon \mathcal{R}(t_i)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} e^{\beta E_j(\lambda_1, \lambda_2)} \right)^2}{\left(\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)} \right)^2}$$

$$- \sum_{i=1}^{I} m_i \frac{\left(\sum_{j \epsilon \mathcal{R}(t_i)} \beta e^{\beta E_j(\lambda_1, \lambda_2)} \left[\frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_1^2} + \beta \left(\frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} \right)^2 \right] \right) \left(\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)} \right)}{\left(\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)} \right)^2}$$

$$= \sum_{i=1}^{I} \left[A_6(t_i) - m_i \frac{C_6(t_i) C_1(t_i) - (C_4(t_i))^2}{(C_1(t_i))^2} \right]$$

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial \lambda_2^2} &= \frac{\partial \left(\sum_{i=1}^{I} \left[\beta \sum_{j \epsilon \mathcal{R}(t_i, Y_j = 1)} \frac{\partial EE_i}{\partial \lambda_2} - m_i \frac{\sum_{j \epsilon \mathcal{R}(t_i)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} e^{\beta E_j(\lambda_1, \lambda_2)}}{\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)}} \right] \right)}{\partial \lambda_2} \\
&= \beta \sum_{i=1}^{I} \sum_{j \epsilon \mathcal{R}(t_i, Y_j = 1)} \frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_2^2} \\
&+ \sum_{i=1}^{I} m_i \frac{\left(\sum_{j \epsilon \mathcal{R}(t_i)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} e^{\beta E_j(\lambda_1, \lambda_2)} \right)^2}{\left(\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)} \right)^2} \\
&- \sum_{i=1}^{I} m_i \frac{\left(\sum_{j \epsilon \mathcal{R}(t_i)} \beta e^{\beta E_j(\lambda_1, \lambda_2)} \left[\frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_2^2} + \beta \left(\frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} \right)^2 \right] \right) \left(\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)} \right)}{\left(\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)} \right)^2} \\
&= \sum_{i=1}^{I} \left[A_7(t_i) - m_i \frac{C_7(t_i) C_1(t_i) - (C_5(t_i))^2}{(C_1(t_i))^2} \right]
\end{aligned}
$$

$$\frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_2} = \frac{\partial \left(\sum_{i=1}^{I} \left[\beta \sum_{j \epsilon \mathcal{R}(t_i, Y_j=1)} \frac{\partial EE_i}{\partial \lambda_1} - m_i \frac{\sum_{j \epsilon \mathcal{R}(t_i)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} e^{\beta E_j(\lambda_1, \lambda_2)}}{\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)}} \right] \right)}{\partial \lambda_2}$$

$$= \beta \sum_{i=1}^{I} \sum_{j \epsilon \mathcal{R}(t_i, Y_j=1)} \frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_1 \partial \lambda_2}$$

$$- \sum_{i=1}^{I} m_i \frac{\left(\sum_{j \epsilon \mathcal{R}(t_i)} \beta e^{\beta E_j(\lambda_1, \lambda_2)} \left[\frac{\partial^2 E_j(\lambda_1, \lambda_2)}{\partial \lambda_1 \partial \lambda_2} + \beta \left(\frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} \right) \left(\frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} \right) \right] \right) \left(\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)} \right)}{\left(\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)} \right)^2}$$

$$+ \sum_{i=1}^{I} m_i \frac{\left(\sum_{j \epsilon \mathcal{R}(t_i)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} e^{\beta E_j(\lambda_1, \lambda_2)} \right) \left(\sum_{j \epsilon \mathcal{R}(t_i)} \beta \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} e^{\beta E_j(\lambda_1, \lambda_2)} \right)}{\left(\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)} \right)^2}$$

$$= \sum_{i=1}^{I} \left[A_8(t_i) - m_i \frac{C_8(t_i) C_1(t_i) - C_4(t_i) C_5(t_i)}{(C_1(t_i))^2} \right]$$

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial \beta \partial \lambda_1} &= \frac{\partial \left(\sum_{i=1}^{I} \left[\sum_{j \epsilon \mathcal{R}(t_i, Y_j=1)} E_j(\lambda_1, \lambda_2) - m_i \frac{\sum_{j \epsilon \mathcal{R}(t_i)} E_j(\lambda_1, \lambda_2) e^{\beta E_j(\lambda_1, \lambda_2)}}{\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)}} \right] \right)}{\partial \lambda_1} \\
&= \sum_{i=1}^{I} \sum_{j \epsilon \mathcal{R}(t_i, Y_j=1)} \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} \\
&\quad - \sum_{i=1}^{I} m_i \frac{\left(\sum_{j \epsilon \mathcal{R}(t_i)} \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} e^{\beta E_j(\lambda_1, \lambda_2)} \left[1 + \beta E_j(\lambda_1, \lambda_2) \right] \right) \left(\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)} \right)}{\left(\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)} \right)^2} \\
&\quad + \sum_{i=1}^{I} m_i \frac{\left(\sum_{j \epsilon \mathcal{R}(t_i)} \beta e^{\beta E_j(\lambda_1, \lambda_2)} \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_1} \right) \left(\sum_{j \epsilon \mathcal{R}(t_i)} E_j(\lambda_1, \lambda_2) e^{\beta E_j(\lambda_1, \lambda_2)} \right)}{\left(\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)} \right)^2} \\
&= \sum_{i=1}^{I} \left[\frac{A_4(t_i)}{\beta} - m_i \frac{B_{2\lambda_1}(t_i) C_1(t_i) - C_4(t_i) B_1(t_i)}{\left(C_1(t_i) \right)^2} \right]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial \beta \partial \lambda_2} &= \frac{\partial \left(\sum_{i=1}^{I} \left[\sum_{j \epsilon \mathcal{R}(t_i, Y_j = 1)} E_j(\lambda_1, \lambda_2) - m_i \frac{\sum_{j \epsilon \mathcal{R}(t_i)} E_j(\lambda_1, \lambda_2) e^{\beta E_j(\lambda_1, \lambda_2)}}{\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)}} \right] \right)}{\partial \lambda_2} \\
&= \sum_{i=1}^{I} \sum_{j \epsilon \mathcal{R}(t_i, Y_j = 1)} \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} \\
&\quad - \sum_{i=1}^{I} m_i \frac{\left(\sum_{j \epsilon \mathcal{R}(t_i)} \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} e^{\beta E_j(\lambda_1, \lambda_2)} \left[1 + \beta E_j(\lambda_1, \lambda_2) \right] \right) \left(\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)} \right)}{\left(\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)} \right)^2} \\
&\quad + \sum_{i=1}^{I} m_i \frac{\left(\sum_{j \epsilon \mathcal{R}(t_i)} \beta e^{\beta E_j(\lambda_1, \lambda_2)} \frac{\partial E_j(\lambda_1, \lambda_2)}{\partial \lambda_2} \right) \left(\sum_{j \epsilon \mathcal{R}(t_i)} E_j(\lambda_1, \lambda_2) e^{\beta E_j(\lambda_1, \lambda_2)} \right)}{\left(\sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta E_j(\lambda_1, \lambda_2)} \right)^2} \\
&= \sum_{i=1}^{I} \left[\frac{A_5(t_i)}{\beta} - m_i \frac{B_{2\lambda_2}(t_i) C_1(t_i) - C_5(t_i) B_1(t_i)}{\left(C_1(t_i) \right)^2} \right]
\end{aligned}
$$

DERIVATIONS: MULTIPLE DOSING

TWO DOSES, SINGLE LAG

Let D_1 represent the first (in terms of when it occurred) exposure specific steady state infusion dose that starts at $t = b_1$ and ends at $t = f_1$ for the subject i. This same individual was later exposed at another steady state dosing level, D_2 starting at $t = b_2$ and ending at $t = f_2$. The full effective exposure, $E_{it}(\lambda)$, for this individual, i, at time, t, can be written as follows:

$$\begin{aligned}
E_{it}(\lambda) &= D_1\left(1 - e^{-\lambda(t-b_1)}\right) I\left(t\epsilon\left[b_1, f_1\right]\right) + D_1\left(1 - e^{-\lambda(f_1-b_1)}\right) e^{-\lambda(t-f_1)} I\left(t > f_1\right) \\
&\quad + D_2\left(1 - e^{-\lambda(t-b_2)}\right) I\left(t\epsilon\left[b_2, f_2\right]\right) + D_2\left(1 - e^{-\lambda(f_2-b_2)}\right) e^{-\lambda(t-f_2)} I\left(t > f_2\right) \\
&= D_1\left(e^{-\lambda \max(0,t-f_1)} - e^{-\lambda \max(0,t-b_1)}\right) + D_2\left(e^{-\lambda \max(0,t-f_2)} - e^{-\lambda \max(0,t-b_2)}\right) \\
&= D_1\left(e^{-\lambda z_2} - e^{-\lambda z_1}\right) + D_2\left(e^{-\lambda z_4} - e^{-\lambda z_3}\right)
\end{aligned}$$

where

$$z_1 = \max(0, t - b_1) = \begin{cases} t - b_1 & \text{if } t > b_1 \\ 0 & \text{otherwise} \end{cases}$$

$$z_2 = \max(0, t - f_1) = \begin{cases} t - f_1 & \text{if } t > f_1 \\ 0 & \text{otherwise} \end{cases}$$

$$z_3 = \max(0, t - b_2) = \begin{cases} t - b_2 & \text{if } t > b_2 \\ 0 & \text{otherwise} \end{cases}$$

$$z_4 = \max(0, t - f_2) = \begin{cases} t - f_2 & \text{if } t > f_2 \\ 0 & \text{otherwise} \end{cases}$$

Let us recall that $E_{it}(\lambda)$ depends on the lag parameter, λ:

$$\begin{aligned} E_{it}(\lambda) &= D_1 \left(e^{-\lambda z_2} - e^{-\lambda z_1}\right) + D_2 \left(e^{-\lambda z_4} - e^{-\lambda z_3}\right) \\ \frac{\partial E_{it}(\lambda)}{\partial \lambda} &= D_1 \left(z_1 e^{-\lambda z_1} - z_2 e^{-\lambda z_2}\right) + D_2 \left(z_3 e^{-\lambda z_3} - z_4 e^{-\lambda z_4}\right) \\ \frac{\partial^2 E_{it}(\lambda)}{\partial \lambda^2} &= D_1 \left(z_2^2 e^{-\lambda z_2} - z_1^2 e^{-\lambda z_1}\right) + D_2 \left(z_4^2 e^{-\lambda z_4} - z_3^2 e^{-\lambda z_3}\right) \end{aligned}$$

Choosing to parameterize the lag using h instead of λ:

$$\frac{\partial E_{it}(\lambda)}{\partial h} = \frac{\log 2}{h^2} \left[D_1 \left(z_2 e^{-z_2 \log 2/h} - z_1 e^{-z_1 \log 2/h}\right) + D_2 \left(z_4 e^{-z_4 \log 2/h} - z_3 e^{-z_3 \log 2/h}\right)\right]$$

$$\begin{aligned} \frac{\partial^2 E_{it}(\lambda)}{\partial h^2} = &\frac{D_1 \log 2}{h^3} \left(\frac{z_2^2 \log 2}{h} e^{\frac{-z_2 \log 2}{h}} - 2 z_2 e^{\frac{-z_2 \log 2}{h}} - \frac{z_1^2 \log 2}{h} e^{\frac{-z_1 \log 2}{h}} + 2 z_1 e^{\frac{-z_1 \log 2}{h}}\right) \\ &+ \frac{D_2 \log 2}{h^3} \left(\frac{z_4^2 \log 2}{h} e^{\frac{-z_4 \log 2}{h}} - 2 z_4 e^{\frac{-z_4 \log 2}{h}} - \frac{z_3^2 \log 2}{h} e^{\frac{-z_3 \log 2}{h}} + 2 z_3 e^{\frac{-z_3 \log 2}{h}}\right) \end{aligned}$$

TWO DOSES, TWO LAG PARAMETERS

For the same individual as described in the previous section, the effective exposure using two lag parameters, $E_{it}(\lambda_1, \lambda_2)$, can be written as:

$$\begin{aligned}
E_{it}(\lambda_1, \lambda_2) &= D_1 \left(1 - e^{-\lambda_1 (t - b_1)}\right) I\left(t \epsilon [b_1, f_1]\right) \\
&\quad + D_1 \left(1 - e^{-\lambda_1 (f_1 - b_1)}\right) e^{-\lambda_2 (t - f_1)} I\left(t > f_1\right) \\
&\quad + D_2 \left(1 - e^{-\lambda_1 (t - b_2)}\right) I\left(t \epsilon [b_2, f_2]\right) \\
&\quad + D_2 \left(1 - e^{-\lambda_1 (f_2 - b_2)}\right) e^{-\lambda_2 (t - f_2)} I\left(t > f_2\right) \\
&= D_1 \left(1 - e^{-\lambda_1 (\max(0, t - b_1) - \max(0, t - f_1))}\right) e^{-\lambda_2 \max(0, t - f_1)} \\
&\quad + D_2 \left(1 - e^{-\lambda_1 (\max(0, t - b_2) - \max(0, t - f_2))}\right) e^{-\lambda_2 \max(0, t - f_2)} \\
&= D_1 \left(1 - e^{-\lambda_1 (z_1 - z_2)}\right) e^{-\lambda_2 z_2} + D_2 \left(1 - e^{-\lambda_1 (z_3 - z_4)}\right) e^{-\lambda_2 z_4}
\end{aligned}$$

Thus we derive the first and second derivatives of $E_{it}(\lambda_1, \lambda_2)$ with respect to both $\lambda_1$and λ_2:

$$\begin{aligned}
\frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1} &= (z_1 - z_2) D_1 e^{-\lambda_1 (z_1 - z_2)} e^{-\lambda_2 z_2} + (z_3 - z_4) D_2 e^{-\lambda_1 (z_3 - z_4)} e^{-\lambda_2 z_4} \\
\frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_2} &= -z_2 D_1 \left(1 - e^{-\lambda_1 (z_1 - z_2)}\right) e^{-\lambda_2 z_2} - z_4 D_2 \left(1 - e^{-\lambda_1 (z_3 - z_4)}\right) e^{-\lambda_2 z_4} \\
\frac{\partial^2 E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1^2} &= -\left(z_1 - z_2\right)^2 D_1 e^{-\lambda_1 (z_1 - z_2)} e^{-\lambda_2 z_2} - \left(z_3 - z_4\right)^2 D_2 e^{-\lambda_1 (z_3 - z_4)} e^{-\lambda_2 z_4} \\
\frac{\partial^2 E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_2^2} &= z_2^2 D_1 \left(1 - e^{-\lambda_1 (z_1 - z_2)}\right) e^{-\lambda_2 z_2} + z_4^2 D_2 \left(1 - e^{-\lambda_1 (z_3 - z_4)}\right) e^{-\lambda_2 z_4} \\
\frac{\partial^2 E_{it}(\lambda_1, \lambda_2)}{\partial \lambda_1 \partial \lambda_2} &= -\left(z_1 - z_2\right) z_2 D_1 e^{-\lambda_1 (z_1 - z_2)} e^{-\lambda_2 z_2} - \left(z_3 - z_4\right) z_4 D_2 e^{-\lambda_1 (z_3 - z_4)} e^{-\lambda_2 z_4}
\end{aligned}$$

Choosing to parameterize the lag using (h_1, h_2) instead of (λ_1, λ_2):

$$\frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial h_1} = \frac{-(z_1 - z_2) \log 2}{h_1^2} D_1 e^{-(z_1 - z_2) \log 2 / h_1} e^{-z_2 \log 2 / h_2}$$

$$- \frac{(z_3 - z_4)\log 2}{h_1^2} D_2 e^{-(z_3-z_4)\log 2/h_1} e^{-z_4 \log 2/h_2}$$

$$\begin{aligned}\frac{\partial E_{it}(\lambda_1, \lambda_2)}{\partial h_2} &= \frac{z_2 \log 2 D_1}{h_2^2} \left(1 - e^{-(z_1-z_2)\log 2/h_1}\right) e^{-z_2 \log 2/h_2} \\ &+ \frac{z_4 \log 2 D_2}{h_2^2} \left(1 - e^{-(z_3-z_4)\log 2/h_1}\right) e^{-z_4 \log 2/h_2}\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 E_{it}(\lambda_1, \lambda_2)}{\partial {h_1}^2} &= \frac{(z_1 - z_2)\, D_1 \log 2}{{h_1}^2} e^{-(z_1-z_2)\log 2/h_1} e^{-z_2 \log 2/h_2} \left(\frac{2}{h_1} - \frac{(z_1 - z_2)\log 2}{{h_1}^2}\right) \\ &+ \frac{(z_3 - z_4)\, D_2 \log 2}{{h_1}^2} e^{-(z_3-z_4)\log 2/h_1} e^{-z_4 \log 2/h_2} \left(\frac{2}{h_1} - \frac{(z_3 - z_4)\log 2}{{h_1}^2}\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 E_{it}(\lambda_1, \lambda_2)}{\partial {h_2}^2} &= \frac{z_2 D_1 \log 2}{{h_2}^2} \left(1 - e^{-(z_1-z_2)\log 2/h_1}\right) e^{-z_2 \log 2/h_2} \left(\frac{2}{h_2} - \frac{z_2 \log 2}{{h_1}^2}\right) \\ &+ \frac{z_4 D_2 \log 2}{{h_2}^2} \left(1 - e^{-(z_3-z_4)\log 2/h_1}\right) e^{-z_4 \log 2/h_2} \left(\frac{2}{h_2} - \frac{z_4 \log 2}{{h_1}^2}\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 E_{it}(\lambda_1, \lambda_2)}{\partial h_1 \partial h_2} &= -z_2 \left(z_1 - z_2\right) \frac{(\log 2)^2}{{h_1}^2 {h_2}^2} D_1 e^{-(z_1-z_2)\log 2/h_1} e^{-z_2 \log 2/h_2} \\ &- z_4 \left(z_3 - z_4\right) \frac{(\log 2)^2}{{h_1}^2 {h_2}^2} D_2 e^{-(z_3-z_4)\log 2/h_1} e^{-z_4 \log 2/h_2}\end{aligned}$$

LIST OF JOURNAL ABBREVIATIONS

Am J Epidemiol	American Journal of Epidemiology
Am J Hum Biol	American Journal of Human Biology
Am J Ind Med	The American Journal of Industrial Medicine
Am J Prev Med	American Journal of Preventative Medicine
Am J Public Health ...	American Journal of Public Health
Am Stat	The American Statistician
Ann Intern Med	Annals of Internal Medicine
Ann Stat	The Annals of Statistics
Arch Intern Med	Archives of Internal Medicine
BMC	BioMed Central
BMC Med Res Methodol .	BioMed Central Medical Research Methodology
BMJ	The British Medical Journal
BMJ Open	BMJ Open Access
Breast Cancer Res	Breast Cancer Research
Calcif Tissue Int	Calcified Tissue International Journal
Can J Stat	The Canadian Journal of Statistics
IEEE Trans Automat Contr	Institute of Electrical and Electronics Engineers Transactions on Automatic Control
Int J Cancer	International Journal of Cancer
Int J Exerc Sci	International Journal of Exercise Science
Int J Obes	International Journal of Obesity

JAMA	Journal of the American Medical Association
JASA	Journal of the American Statistical Association
J Am Med Womens Assoc	The Journal of the American Medical Women's Association
J Bone Miner Res	Journal of Bone and Mineral Research
J Clin Epidemiol	Journal of Clinical Epidemiology
J Chronic Dis	Journal of Chronic Diseases
J Lab Clin Med	The Journal of Laboratory and Clinical Medicine
J Pharmacokinet Biopharm	Journal of Pharmacokinetics and Biopharmaceutics
J Psychiatry Neurosci ..	Journal of Psychiatry and Neuroscience
J R Stat Soc Ser B	Journal of the Royal Statistical Society. Series B (Statistical Methodology)
J R Stat Soc Ser C	Journal of the Royal Statistical Society. Series C (Applied Statistics)
J Stat Softw	The Journal of Statistical Software
N Engl J Med	New England Journal of Medicine
Prog Cardiovasc Dis ...	Progress in Cardiovascular Diseases
PM R	Journal of the American Academy of Physical Medicine and Rehabilitation
Rev Esp Cardiol	Revista Espanola de Cardiologia
Sci Rep	Scientific Reports
Stat Med	Statistics in Medicine

BIBLIOGRAPHY

Abrahamowicz, M., Bartlett, G., Tamblyn, R., & Du Berger, R. (2006). Modeling cumulative dose and exposure duration provided insights regarding the associations between benzodiazepines and injuries. *J Clin Epidemiol*, *59*(4), 393–403.

Abrahamowicz, M., Beauchamp, M. E., & Sylvestre, M. P. (2012). Comparison of alternative models for linking drug exposure with adverse effects. *Stat Med*, *31*(11-12), 1014–1030.

Abrahamowicz, M., Ciampi, A., & Ramsay, J. O. (1992). Nonparametric Density Estimation For Censored Survival Data: Regression-Spline Approach. *Can J Stat*, *20*(2), 171–185.

Abrahamowicz, M., MacKenzie, T. A., & Esdaile, J. M. (1996). Time-dependent hazard ratio: Modeling and hypothesis testing with application in lupus nephritis. *JASA*, *91*(436), 1432–1439.

Agid, O., Seeman, P., & Kapur, S. (2006). The "delayed onset" of antipsychotic action–an idea whose time has come and gone. *J Psychiatry Neurosci*, *31*(2), 93–100.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans Automat Contr*, *19*(6), 716–723.

Andersen, P. K., & Gill, R. (1982). Cox's Regression Model for Counting Processes : A Large Sample Study. *Ann Stat*, *10*(4), 1100–1120.

Bourne, D. W. (2010). Continuous IV Infusion - Steady State Model (Chapter 6). http://www.boomer.org/c/p4/c06/c06.html (Accessed:2016-10-04).

Breslow, N. E. (1974). Covariance Analysis of Censored Survival Data. *Biometrics*, *30*(1), 89–99.

Burns, D. M. (2003). Epidemiology of smoking-induced cardiovascular disease. *Prog Cardiovasc Dis*, *46*(1), 11–29.

Cavender, J. B., Rogers, W. J., Fisher, L. D., Gersh, B. J., Coggin, C. J., & Myers, W. O. (1992). Effects of smoking on survival and morbidity in patients randomized to medical or surgical therapy in the Coronary Artery Surgery Study (CASS): 10-year follow-up. CASS Investigators. *J Am Coll Cardiol*, *20*(2), 287–294.

Centers for Disease Control and Prevention (2015). Childhood Lead Poisoning Prevention Program - PLPYC 91 Chapter 6. https://www.cdc.gov/nceh/lead/publications/books/plpyc/chapter7.htm (Accessed:2018-03-14).

Centers for Disease Control and Prevention (2016). Smoking and Tobacco Use; Fact Sheet; Smoking Cessation. https://www.cdc.gov/tobacco/data_statistics/fact_sheets/cessation/quitting/index.htm (Accessed:2017-10-04).

Chaiton, M., Diemert, L., Cohen, J. E., Bondy, S. J., Selby, P., Philipneri, A., & Schwartz, R. (2016). Estimating the number of quit attempts it takes to quit smoking successfully in a longitudinal cohort of smokers. *BMJ Open*, *6*(6), e011045.

Coggon, D., Reading, I., Croft, P., McLaren, M., Barrett, D., & Cooper, C. (2001). Knee osteoarthritis and obesity. *Int J Obes*, *25*(5), 622–627.

Cole, S. R., Chu, H., & Greenland, S. (2014). Maximum likelihood, profile likelihood, and penalized likelihood: A primer. *Am J Epidemiol*, *179*(2), 252–260.

Cook, A. (2010). Chapter 3. The Cox Proportional Hazards Model. http://blog.nus.edu.sg/alexcook/files/2010/12/ch31.pdf (Accessed:2017-09-16).

Copeland, K. T., Checkoway, H., Mcmichael, A. J., & Holbrook, R. H. (1977). Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol*, *105*(5), 488–495.

Cox, D. R. (1972). Regression Models and Life-Tables. *J R Stat Soc Ser B*, *34*(2), 187–220.

Cox, D. R., & Reid, N. (1992). A note on the difference between profile and modified profile likelihood.

Critchley, J. A., & Capewell, S. (2003). Mortality Risk Reduction Associated With Smoking Cessation in Patients with Coronary Heart Disease: A Systematic Review. *JAMA*, *290*(1), 86–97.

Cupples, L. A., D'Agostino, R. B., Anderson, K., & Kannel, W. B. (1988). Comparison of baseline and repeated measure covariate techniques in the Framingham Heart Study. *Stat Med*, *7*(1-2), 205–218.

D'Agostino, R. B., Lee, M. L., Belanger, A. J., Cupples, L. A., Anderson, K., & Kannel, W. B. (1990). Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Stat Med*, *9*(12), 1501–15.

Delnevo, C. D., Gundersen, D. A., Hrywna, M., Echeverria, S. E., & Steinberg, M. B. (2011). Smoking-cessation prevalence among U.S. smokers of menthol versus non-menthol cigarettes. *Am J Prev Med*, *41*(4), 357–365.

Efron, B. (1977). The Efficiency of Cox's Likelihood Function for Censored Data. *JASA*, *72*(359), 557–565.

Enea, M., Meiri, R., & Kalimi, T. (2015). Package 'speedglm'. https://cran.r-project.org/package=speedglm.

Felson, D. T., Anderson, J. J., Naimark, A., Walker, A. M., & Meenan, R. F. (1988). Obesity and knee osteoarthritis. The Framingham Study. *Ann Intern Med*, *109*(1), 18–24.

Finkelstein, M. M. (1991). Use of "time windows" to investigate lung cancer latency intervals at an ontario steel plant. *Am J Ind Med*, *19*(2), 229–235.

Flora, G., Gupta, D., & Tiwari, A. (2012). Toxicity of lead: A review with recent updates. *Interdiscip Toxicol*, *5*(2), 47–58.

Gasparrini, A. (2011). Distributed Lag Linear and Non-Linear Models in R: The Package dlnm. *J Stat Softw*, *43*(8), 1–20.

Gasparrini, A. (2014). Modeling exposure-lag-response associations with distributed lag non-linear models. *Stat Med*, *33*(5), 881–899.

Gasparrini, A. (2016). Modelling lagged associations in environmental time series data. *Epidemiology*, *27*(6), 1.

Gasparrini, A., Armstrong, B., & Kenward, M. G. (2010). Distributed lag non-linear models. *Stat Med*, *29*(21), 2224–2234.

Gasparrini, A., & Leone, M. (2014). Attributable risk from distributed lag models. *BMC Med Res Methodol*, *14*(1), 55.

Giovannucci, E., Egan, K. M., Hunter, D. J., Stampfer, M. J., Colditz, G. A., Willett, W. C., & Speizer, F. E. (1995). Aspirin and the risk of colorectal cancer in women. *NEJM*, *333*(10), 609–14.

Giovannucci, E., Rimm, E. B., Stampfer, M. J., Colditz, G. A., Ascherio, A., & Willett, W. C. (1994). Aspirin Use and the Risk for Colorectal Cancer and Adenoma in Male Health Professionals. *Ann Intern Med*, *121*(4), 241.

Green, M. S., & Symons, M. J. (1983). A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *J Chronic Dis*, *36*(10), 715–723.

Gulson, B. L., Mahaffey, K. R., Mizon, K. J., Korsch, M. J., Cameron, M. A., & Vimpani, G. (1995). Contribution of tissue lead to blood lead in adult female subjects based on stable lead isotope methods. *J Lab Clin Med*, *125*(6), 703–712.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning The Elements of Statistical LearningData Mining, Inference, and Prediction, Second Edition*. Springer-Verlag New York.

Hauptmann, M., Lubin, J. H., Rosenberg, P. S., Wellmann, J., & Kreienbrock, L. (2000a). The use of sliding time windows for the exploratory analysis of temporal effects of smoking histories on lung cancer risk. *Stat Med*, *19*, 2185–2194.

Hauptmann, M., Wellmann, J., Lubin, J. H., Rosenberg, P. S., & Kreienbrock, L. (2000b). Analysis of exposure-time-response relationships using a spline weight function. *Biometrics*, *56*(4), 1105–1108.

Hertz-Picciotto, I., & Rockhill, B. (1997). Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics*, *53*(3), 1151–6.

Hosmer, D. W., Lemeshow, S., May, S., Ibrahim, J. G., & Muirhead, R. J. (2008). Applied Survival Analysis. *21st Annual Summer Work Northeast Illinois Chapter of the American Statistical Association*, *41*(X), i—-viii.

Howard, G., Wagenknecht, L. E., Burke, G. L., Diez-Roux, A., Evans, G. W., McGovern, P., Nieto, J., & Tell, G. S. (1998). Cigarette smoking and progression of atherosclerosis: The Atherosclerosis Risk in Communities (ARIC) Study. *JAMA*, *279*(2), 119–124.

Jackson, E., & Rubenfire, M. (2015). Cardiovascular risk of smoking and benefits of smoking cessation. https://www.uptodate.com/contents/cardiovascular-risk-of-smoking-and-benefits-of-smoking-cessation (Accessed:2017-02-28).

Kawachi, I., Colditz, G. A., Stampfer, M. J., Willet, W. C., Manson, J. E., Rosner, B., Speizer, F. E., & Hennekens, C. H. (1994). Smoking Cessation and Time Course of Decreased Risks of Coronary Heart Disease in Middle-Aged Women. *Arch Intern Med*, *154*(2), 169–75.

Kleinbaum, D. G., & Klein, M. (2011). *Survival Analysis: A Self-Learning Text, Third Edition (Statistics for Biology and Health)*. Springer-Verlag New York.

Koehler, E., Brown, E., & Haneuse, S. J. P. A. (2009). On the assessment of Monte Carlo error in simulation-based Statistical analyses. *Am Stat*, *63*(2), 155–162.

Lachenmeier, D. W., & Rehm, J. (2015). Comparative risk assessment of alcohol, tobacco, cannabis and other illicit drugs using the margin of exposure approach. *Sci Rep*, *5*, 8126.

Langholz, B., Thomas, D. C., Xiang, A., & Stram, D. (1999). Latency analysis in epidemiologic studies of occupational exposures: Application to the Colorado Plateau uranium miners cohort. *Am J Ind Med*, *35*(3), 246–256.

Law, M. R., & Wald, N. J. (2003). Environmental tobacco smoke and ischemic heart disease. *Prog Cardiovasc Dis*, *46*(1), 31–38.

Leffondré, K., Abrahamowicz, M., Siemiatycki, J., & Rachet, B. (2002). Modeling smoking history: A comparison of different approaches. *Am J Epidemiol*, *156*(9), 813–823.

Mahboubi, A., Abrahamowicz, M., Giorgi, R., Binquet, C., Bonithon-Kopp, C., & Quantin, C. (2011). Flexible modeling of the effects of continuous prognostic factors in relative survival. *Stat Med*, *30*(12), 1351–65.

Mitra, R. (2011). Adverse effects of corticosteroids on bone metabolism: A review. *PM R*, *3*(5), 466–471.

Mohiuddin, S. M., Mooss, A. N., Hunter, C. B., Grollmes, T. L., Cloutier, D. A., & Hilleman, D. E. (2007). Intensive smoking cessation intervention reduces mortality in high-risk smokers with cardiovascular disease. *Chest*, *131*(2), 446–452.

Murphy, S. A., & Van Der Vaart, A. W. (2000). On profile likelihood. *JASA*, *95*(450), 449–465.

Ngwa, J. S., Cabral, H. J., Cheng, D. M., Pencina, M. J., Gagnon, D. R., LaValley, M. P., & Cupples, L. A. (2016). A comparison of time dependent Cox regression, pooled logistic regression and cross sectional pooling with simulations and an application to the Framingham Heart Study. *BMC Med Res Methodol*, *16*(1), 1–12.

of Florida, U. (2000). Useful Pharmacokinetic Equations. http://pharmacy.ufl.edu/files/2013/01/5127-28-equations.pdf (Accessed:2017-01-25).

Prescott, E., Hippe, M., Schnohr, P., Hein, H. O., & Vestbo, J. (1998). Smoking and risk of myocardial infarction in women and men: longitudinal population study. *BMJ*, *316*(7137), 1043.

R Core Team (2017). R: A Language and Environment for Statistical Computing. https://www.r-project.org/.

Rachet, B., Abrahamowicz, M., Sasco, A. J., & Siemiatycki, J. (2003). Estimating the distribution of lag in the effect of short-term exposures and interventions: Adaptation of a non-parametric regression spline model. *Stat Med*, *22*(14), 2335–2363.

Richardson, D. B. (2009). Latency Models for Analyses of Protracted Exposures. *Epidemiology*, *20*(3), 395–399.

Richardson, D. B., Cole, S. R., Chu, H., & Langholz, B. (2011). Lagging exposure information in cumulative exposure-response analyses. *Am J Epidemiol*, *174*(12), 1416–1422.

Rogot, E., & Murray, J. L. (1980). Smoking and causes of death among U.S. veterans: 16 years of observation. *Public Health Reports*, *95*(3), 213–222.

Rosenberg, L., Adams-Campbell, L. L., & Palmer, J. R. (1995). The Black Women's Health Study: a follow-up study for causes and preventions of illness. *J Am Med Womens Assoc*, *50*(2), 56–8.

Rosenberg, L., Kaufman, D. W., Helmrich, S. P., & Shapiro, S. (1985). The Risk of Myocardial Infarction After Quitting Smoking In Men Under 55 Years of Age. *NEJM*, *313*(24), 1511–1514.

Rosenberg, L., Palmer, J. R., & Shapiro, S. (1990). Decline in the Risk of Myocardial Infarction among Women Who Stop Smoking. *NEJM*, *322*(4), 213–217.

Rothman, K. J. (1981). Induction and Latent Periods. *Am J Epidemiol*, *114*(2), 253–259.

Salvan, A., Stayner, L., Steenland, K., & Smith, R. (1995). Selecting an Exposure Lag Period. *Epidemiology*, *6*(4), 387–390.

SAS Institute Inc., Cary, N. (2011). SAS. www.sas.com/pt_br/home.html.

Schwartz, J. (2000). The Distributed Lag between Air Pollution and Daily Deaths. *Epidemiology*, *11*(3), 320–326.

Scranton, R. E., Young, M., Lawler, E., Solomon, D. H., Gagnon, D. R., & Gaziano, J. M. (2005). Statin use and fracture risk: Study of a US veterans population. *Arch Intern Med*, *165*(17), 2007–2012.

Serrano, M., Madoz, E., Ezpeleta, I., San Julián, B., Amézqueta, C., Pérez Marco, J. A., & de Irala, J. (2003). Smoking cessation and risk of myocardial reinfarction in coronary patients: a nested case-control study. *Rev Esp Cardiol*, *56*(5), 445–451.

Sharpsteen, C., & Bracken, C. (2016). *tikzDevice: R Graphics Output in LaTeX Format*. https://cran.r-project.org/package=tikzDevice.

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., & Despouy, P. (2017). *plotly: Create Interactive Web Graphics via 'plotly.js'*. https://cran.r-project.org/package=plotly.

Sprott, D. A. (2000). *Statistical Inference in Science*. Springer Series in Statistics. New York, NY: Springer-Verlag New York, 1 ed.

Strohacker, K., Carpenter, K. C., & McFarlin, B. K. (2009). Consequences of Weight Cycling: An Increase in Disease Risk? *Int J Exerc Sci*, *2*(3), 191–201.

Sylvestre, M.-P., & Abrahamowicz, M. (2009). Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Stat Med*, *28*(27), 3437–3453.

Team, R. (2016). RStudio: Integrated Development for R. *[Online] RStudio, Inc, Boston, MA URL http//www rstudio com*, (pp. RStudio, Inc., Boston, MA).

Teo, K., Ounpuu, S., & Hawken, S. (2006). INTERHEART Study Investigators. Tobacco use and risk of myocardial infarction in 52 countries in the INTERHEART study: a case-control study. *Lancet*, *368*(9536), 647–658.

Therneau, T. M. (2015). A Package for Survival Analysis in S. https://cran.r-project.org/package=survival.

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. New York: Springer.

Thomas, D. C. (1983). Statistical methods for analyzing effects of temporal patterns of exposure on cancer risks. *Scand J Work Environ Health*, *9*(4), 353–66.

Tofler, G. (2017). Psychosocial factors in acute myocardial infarction. https://www.uptodate.com/contents/psychosocial-factors-in-acute-myocardial-infarction (Accessed:2018-01-19).

Toh, S., & Hernandez-Diaz, S. (2007). Statins and fracture risk. A systematic review. *Pharmacoepidemiology and Drug Safety*, *16*(6), 627–640.

Tolstrup, J. S., Hvidtfeldt, U. A., Flachs, E. M., Spiegelman, D., Heitmann, B. L., Bälter, K., Goldbourt, U., Hallmans, G., Knekt, P., Liu, S., Pereira, M., Stevens, J., Virtamo, J., & Feskanich, D. (2014). Smoking and risk of coronary heart disease in younger, middle-aged, and older adults. *Am J Public Health*, *104*(1), 96–102.

U.S. Department of Health and Human Services (1990). The Health Benefits of Smoking Cessation. Tech. Rep. DHHS Publication No. (CDC) 90-8416, U.S. Department of Health and Human Services. Public Health Service. Centers for Disease Control. Center for Chronic Disease Prevention and Health Promotion. Office on Smoking and Health.

U.S. Department of Health and Human Services (2010). How Tobacco Smoke Causes Disease The Biology and Behavioral Basis for Smoking-Attributable Disease A Report of the Surgeon General. Tech. rep., U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, GA.

Van Gaalen, R. D., Abrahamowicz, M., & Buckeridge, D. L. (2015). The impact of exposure model misspecification on signal detection in prospective pharmacovigilance. *Pharmacoepidemiology and Drug Safety*, *24*(5), 456–467.

Van Staa, T. P., Leufkens, H. G., Abenhaim, L., Zhang, B., & Cooper, C. (2000). Use of oral corticosteroids and risk of fractures. *J Bone Miner Res*, *15*(6), 993–1000.

Venzon, D. J., & Moolgavkar, S. H. (1988). A method for computing profile-likelihood-based confidence-intervals. *J R Stat Soc Ser C*, *C 37*(1), 87–94.

Vestergaard, P., Rejnmark, L., & Mosekilde, L. (2008). Fracture risk associated with different types of oral corticosteroids and effect of termination of corticosteroids on the risk of fractures. *Calcif Tissue Int*, *82*(4), 249–257.

WHO (2004). Fact sheet about health benefits of smoking cessation. http://www.who.int/tobacco/quitting/en_tfi_quitting_fact_sheet.pdf (Accessed:2017-02-14).

Wickham, H. (2007). Reshaping data with the reshape package. *J Stat Softw*, *21*(12).

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., Francois, R., Henry, L., & Muller, K. (2017). *dplyr: A Grammar of Data Manipulation*. https://cran.r-project.org/package=dplyr.

Wijnand, H. P. (1988). Pharmacokinetic model equations for the one- and two-compartment models with first-order processes in which the absorption and exponential elimination or distribution rate constants are equal. *J Pharmacokinet Biopharm*, *16*(1), 109–128.

Winter, M. (2004). *Basic Clinical Pharmacokinetics*. Baltimore: Lippincott Williams & Wilkins, 4th ed.

Wynant, W., & Abrahamowicz, M. (2014). Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. *Stat Med*, *33*(19), 3318–37.

Wynant, W., & Abrahamowicz, M. (2016). Flexible estimation of survival curves conditional on non-linear and time-dependent predictor effects. *Stat Med*, *35*(4), 553–65.

Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J., & Lisheng, L. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet*, *364*(9438), 937–952.

Zanobetti, A., Wand, M. P., Schwartz, J., & Ryan, L. M. (2000). Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics*, *1*(3), 279–292.

CURRICULUM VITAE