A Bayesian functional approach to test models of life course epidemiology over continuous time

Julien Bodelet[1,*], Cecilia Potente[1], Justin Chumbley[1], Hira Imeri[1], Scott Hofer[2], Kathleen Mullan Harris[3], Graciela Muniz Terrera[4,5], and Michael Shanahan[1]

[1]Jacobs Center for Productive Youth Development, University of Zurich, Zurich Switzerland
[2]Department of Psychology, University of Victoria, Victoria, Canada
[3]University of North Carolina at Chapel Hill, Carolina Population Center, Chapel Hill, NC, USA
[4]Center for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK
[5]Ohio University Heritage College of Osteopathic Medicine, Ohio University, Athens, OH, USA

[*]Corresponding author. Jacobs Center for Productive Youth Development, University of Zurich, Andreastrasse 15, Zurich CH8050, Switzerland. E-mail: julien.bodelet@jacobscenter.uzh.ch

Abstract

Background: Life course epidemiology examines associations between repeated measures of risk and health outcomes across different phases of life. Empirical research, however, is often based on discrete-time models that assume that sporadic measurement occasions fully capture underlying long-term continuous processes of risk.

Methods: We propose (1) the functional Relevant Life Course Model (fRLM), which treats repeated, discrete measures of risk as unobserved continuous processes, and (2) a testing procedure to assign probabilities that the data correspond to conceptual models of life course epidemiology (critical period, sensitive period, and accumulation models). The performance of the fRLM is evaluated with simulations, and the approach is illustrated with empirical applications relating body mass index to mRNA-seq signatures of chronic kidney disease, inflammation, and breast cancer.

Results: Simulations reveal that fRLM identifies the correct life course model with three to five repeated assessments of risk and 400 subjects. The empirical examples reveal that chronic kidney disease reflects a critical period process, while inflammation and breast cancer likely reflect sensitive period mechanisms.

Conclusions: The proposed fRLM treats repeated measures of risk as continuous processes and, under realistic data scenarios, the method provides accurate probabilities that the data correspond to commonly-studied models of life course epidemiology. fRLM is implemented with publicly-available software.

Keywords— Life course models, Bayesian statistics, Functional data analysis.
Word count: 3705

Key Messages

- Models of life course epidemiology typically use discrete-time models whereby a limited number of repeated measures of risk are assumed to capture continuous exposure to risk.
- We propose a model that uses discrete data to test life course hypotheses over continuous time.
- Simulation studies reveal that the correct life course model can be identified with high probability with 3 to 5 repeated assessments of risk and 400 subjects.
- The method and software are illustrated with examples involving BMI trajectories from adolescence to mid-adulthood predicting mRNA-seq signatures of chronic health challenges.

1 Introduction

Life course epidemiology often focuses on exposures to repeated risks and their consequences for health over many decades of life[1]. Empirical studies are typically guided by three nested conceptual models: accumulation, which posits that all exposures to a repeated risk factor meaningfully predict the outcome; sensitive period, according to which more than one, but not all, exposures are predictive; and critical period, meaning that only one exposure matters[2]. Although additional models are recognized[3], methodological research has focused on analytic strategies to determine which of these three models best corresponds to the observed data[4,5,6]. The analytic task has been to (1) estimate the association between exposure to risk and the outcome at each measurement occasion and then (2) decide which conceptual model is best supported by these estimates.

Madathil and colleagues proposed a Relevant Life course Model (RLM) for continuously-scaled repeated exposures measured in successive waves of a panel study to estimate weights associated with each measurement occasion and then select the most apt life course model based on these weights[7]. First, for each subject i, the relevant life course exposure is conceptualized as the product between the continuously-scaled repeated risk x_t and a weight reflecting its relevance at each of the measurement occasions. The outcome y_i is then assumed to depend linearly on the sum of the relevant life exposure,

$$y_i = \delta \sum_{t=1}^{T} x_{i,t} w_t + \boldsymbol{C}_i' \boldsymbol{\alpha} + \epsilon_i. \qquad (1)$$

where $w_t \geq 0$, $\sum_{t=1}^{T} w_t = 1$, are weights, C_i are covariates and ϵ_i random errors. The parameter δ represents the effect of the relevant life exposure $\sum_{t=1}^{T} x_{i,t} w_t$. Closely-spaced, discrete time points and T large, parametric shapes[8] and nonparametric shapes[9,10,11] for w_t have previously been considered. In the RLM framework, the reference weights for the accumulation model refer to the case where $w_t = 1/T$ for all t, the critical period model to the case where $w_t = 1$ for one period and 0 for the others, and sensitive models to any other combinations. Second, Madathil and colleagues select the life course conceptual model based on the distance between the reference weights and the mean of the posterior distribution of weights[12,13,14].

Drawing on the RLM, Chumbley and colleagues proposed a different strategy for deciding which life course model is most descriptive[15]. The proposed method tests life course hypotheses by sequentially partitioning the simplex to identify the most credible ranking among the weights (e.g. that $w_1 > w_2 > w_3$ [a full ranking] or that $w_1, w_2 > w_3$ [a partial ranking]). We refer to this method as the Sequential Partitioning Test (SPT). SPT uses the greatest difference among the weights as test statistics to define regions of practical equivalence (ROPEs) for each of the three conceptual models.

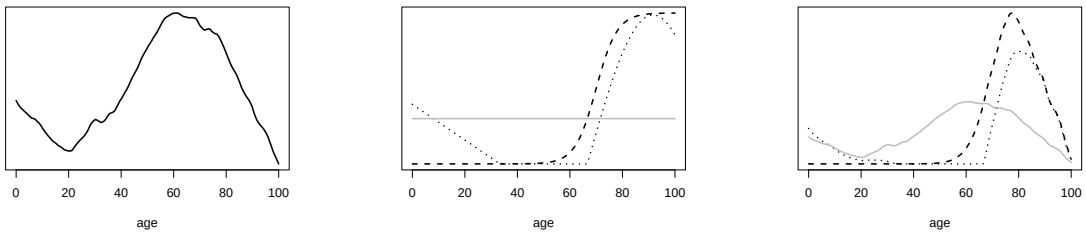
The posterior probability of each model is then estimated by the fraction of posterior Markov Chain Monte Carlo (MCMC) samples falling into the corresponding ROPEs. For models not falling into the accumulation and critical period regions, post hoc decompositions then determine the most likely ranks for a sensitive period model.

Although discrete-time models such as the RLM correspond to the repeated assessments of risk that are often available in cohort studies, they assume that (1) the association between risk and health involves discrete jumps corresponding to the time of measurement; (2) the risk factors and health outcomes are measured at the moment corresponding to these jumps; and (3) the measurement occasions include all relevant times of exposure to risk[16]. Yet these assumptions may well be problematic in cases involving continuous processes of risk exposure. For example, addictive behaviors (such as consumption of tobacco or alcohol) are ongoing physiological assaults. Recently, substantial efforts have been made in the field of epidemiology to address these issues through the development of new functional approaches[17,18,19,20].

In this paper, we propose the functional Relevant Life Course Model (fRLM), an extension of the RLM, which takes into account that the observed risk data are only discrete measurements of an unobserved process changing continuously over time. Specifically, the fRLM assumes that the outcome depends on a weighted integral of the exposure

$$y_i = \alpha + \delta \int X_i(t)\omega(t)dt + \boldsymbol{C}_i' \boldsymbol{\alpha} + \epsilon_i, \qquad (2)$$

where $X_i(t)$ are random functions observed at a finite number of discrete time locations, $\omega(t)$ is a continuous positive weight function, and t now refers to the exact age. An example of the fRLM for the different life course conceptual models is provided in Figure 1 with the relative importance $\omega(t)$ based on discrete measures of risk. The number of measurements is allowed to vary across subjects (i.e., subject specific), and the fRLM is well-suited to panel studies that begin with an age-heterogeneous group. Note that (1) can be also be seen as a particular case of (2) when $X_i(t)$ are



(a) Life course pattern of exposure, $X_i(t)$, for a given subject.

(b) Relative importance of life course exposure, $\omega(t)$.

(c) Relevant Life course exposure, $\omega(t)X_i(t)$.

Figure 1: Simulated examples of continuous life course models representing under the accumulation (gray line), critical (dashed line), and sensitivity (dotted line) hypotheses.

step functions. We consider the estimation of the whole curve $\omega(t)$ and we show how to apply the SPT to the fRLM in order to test the different life course hypotheses.

The paper is structured as follows. We first describe the model and present an estimation method. We assess the performance of the model in identifying the most descriptive conceptual model of life course epidemiology given plausible data scenarios. Drawing on data from the National Longitudinal Study of Adolescent and Adult Health (Add Health), we then consider empirical examples that examine repeated assessments of body mass index (BMI) between ages 12 and 43 and gene expression (mRNA-seq) signatures representing the molecular underpinnings of chronic kidney disease (CKD), inflammation, and breast cancer. The discussion subsequently highlights the advantages and drawbacks of our methodology. Additional simulation experiments are considered in the Supplementary Materials to compare the proposed method with alternative estimation procedures. fRLM and SPT are implemented in R with software available on GitHub.

2 Methods

2.1 The model

We consider data for which, for each subject indexed by $i \in \{1, 2, \ldots, n\}$, one observes a scalar outcome variable, y_i, along with repeated measurements of a time-varying risk exposure variable, $x_{i,j}$, observed at different time locations $t_{i,j}$, where $j \in \{1, 2, \ldots, N_i\}$. Note that both the number of measurement occasions N_i and their specific timing $t_{i,j}$, may vary across subjects. We assume that the $x_{i,j}$ are discrete measurements of smooth functions of the continuous time, $X_i(t)$, specific to each subject. The functions $X_i(t)$ are not observed except at the specified time locations $t_{i,j}$, where we have $X_i(t_{i,j}) = x_{i,j}$. The time t could be the age of subjects or the elapsed time after a lifetime event for example, and lie in a specific time interval $[a, b]$.

We assume that the data is generated by the following functional regression model,

$$y_i = \delta \int_a^b X_i(t)\omega(t)dt + \boldsymbol{C}_i'\boldsymbol{\alpha} + \epsilon_i, \qquad (3)$$

where the functional parameter ω is a positive twice differentiable function that satisfies $\int_a^b \omega(t)dt = 1$. The errors ϵ_i are assumed to be independently and identically normally distributed with mean 0 and variance σ^2; $\boldsymbol{C}_i$ are p-dimensional non-functional covariates with $\boldsymbol{\alpha} \in \mathbb{R}^p$ being the corresponding covariate effects; and δ is a scalar parameter that represents the life-time effect. The function $\omega(t)$ can be interpreted as a density, and the relative importance of a given period T can be computed as the integral $\int_T \omega(t)dt$.

2.2 Estimation method

The model reflects a two step procedure: (1) the prediction of each of the curves $X_i(t)$ based on the samples $x_{i,1}, x_{i,2}, \ldots, x_{i,N_i}$; (2) and then a Bayesian functional regression is estimated using the curves derived from step (1), $\hat{X}_i(t)$.

In the first step, for predicting the individual curves, we make certain assumptions about their prior distribution. The random functions are assumed to be Gaussian Processes with different mean and covariance kernels for each subject in order to allow for variability in the sample curves. The distribution of a Gaussian Process is fully specified by a mean function and covariance function (or covariance kernel). Specifically, we assume that $X_i(t)$ are Gaussian Processes with unknown mean $E[X_i(t)] = \mu_i(t)$ and covariance function

$k_i(t,s) = Cov(X_i(t), X_i(s))$. The parameterization of k_i involves specific behaviors for the random functions, and we select the exponential covariance function, $k_i(t,s) = \nu_i^2 \exp\{-(t-s)^2/\kappa_i\}$, in order to ensure smooth patterns. Here ν_i and κ_i are subject-specific hyperparameters called signal-variance and length-scale, respectively. The hyperparameters can be optimized through maximum likelihood for each subject. For each subject, the curves are predicted with the Gaussian Process regression method. Specifically, given realizations $x_{i,1}, x_{i,2}, \ldots, x_{i,N_i}$, each curve is predicted at any time point t, by its conditional expectation, $\hat{X}_i(t) = E[X_i(t)|X_i(t_{i,1}) = x_{i,1}, X_i(t_{i,2}) = x_{i,2}, \ldots, X_i(t_{i,N_i}) = x_{i,N_i}]$. In the literature on functional regression, alternative methods have been proposed for estimating $X_i(t)$, such as the functional Principal Component Analysis, used in the PACE method[21], and mixture of B-splines[22]. In the Supplementary Material we provide a simulation experiments to compare the performance of these two estimation methods.

In the second step, we estimate a Bayesian functional regression on the predicted risk curves $\hat{X}_i(t)$, and prior distributions for the parameters have to be specified. Establishing a suitable prior for the functional parameter $\omega(t)$ requires care. Mixture of (B-)splines are flexible, effective prior distributions used in nonparametric Bayesian statistics. Specifically we model the functional parameter as a linear combination of B-splines, i.e. $\omega(t) = \sum_{l=1}^{L} \beta_l \phi_l(t)$. In this framework, $\omega(t)$ has to be positive with integral being one. To meet these two constraints, we used ϕ_l as density B-splines[23] (i.e. rescaled B-splines satisfying $\int_a^b \phi(t)dt = 1$), and constrain the parameters β_l to belong to a simplex (i.e we restrict them to be positive and to sum up to one, $\sum_{l=1}^{L} \beta_l$). The Dirichlet distribution is thus proposed, which is a natural distribution over the simplex and satisfies these constraints. A non-informative prior on the coefficient β_l would be $\text{Dir}(1, 1, \ldots, 1)$. Finally, the Bayesian functional regression can then be estimated by computing the integrals $Z_{i,l} := \int_a^b \hat{X}_i(t)\psi_l(t)dt$, and using them as regressors in a linear Bayesian regression model,

$$y_i = \delta \sum_{l=1}^{L} Z_{il}\beta_l + \boldsymbol{C}_i'\boldsymbol{\alpha} + e_i, \qquad (4)$$

where the β_l have Dirichlet priors. The posterior distribution is obtained through MCMC simulations.

2.3 Testing for models of life course epidemiology

The SPT procedure is then used to test which of the models of life course epidemiology best corresponds to the estimates: the accumulation, critical, or sensitive period models[15]. Although the SPT was proposed in the context of the linear RLM (1), the strategy applies to the fRLM as well. In the context of the fRLM, the user defines specific time periods of interest $T_1, T_2, ..., T_J$, such that they form a partition of the unit interval $[0, 1]$. The specification of the time periods should be defined in the specific research context, but might include, for example, age-based categories, or processes before, during, and after events (e.g., the pubertal transition). The user-defined time periods do not necessarily depend on the specifically-timed measurements occasions, which is a distinct advantage vis-á-vis the discrete RLM , according to which the time periods must coincide with the specific measurements.

The relative importance of the measurement occasions, w_j, for the period T_j is then the integral of the weight function ω over the period. That is, $w_j := \int_{T_j} \omega(t)dt$. As $\{T_j, j = 1, 2, \ldots, J\}$ is a partition, $(w_1, w_2, \ldots, w_J)$ belongs to a simplex. Thus, the SPT can be applied to w_j. The distribution of w_j is obtained by integrating the functions ω obtained across the MCMC samples.

3 Evaluation of the fRLM with simulations

3.1 Goals of the simulation

The fRLM and SPT are evaluated over a range of plausible data scenarios. Specifically, we consider the impact of the following on the ability of the model to recover simulated ground-truths:

(i) the underlying model of life course epidemiology (accumulation, and critical and sensitive period models);

(ii) the sample size ($n = \{100, 400\}$); and

(iii) the number of measurement occasions: a sparse scenario, (where N_i is uniformly distributed over $\{3, 4, 5\}$), a moderately sparse scenario, (N_i is uniformly distributed over $\{6, 7, 8\}$), and a scenario with completely observed trajectories (denoted by $N_i = \infty$). For the first two scenarios, we generated random observed time points by $t_{i,j} = \sum_{k=1}^{j} U_{i,k} / \sum_{k=1}^{N_i+1} U_{i,k}$, where

$(U_{i,1}, \ldots, U_{i,N}, U_{i,N+1})$ are generated randomly from standard uniform variables for each simulation scheme and each subject. This allows us to obtain random time points satisfying $0 < t_{i,1} < t_{i,2} < ... < t_{i,N_i} < 1$.

We expect that with increasing sample size and number of observed time points, N_i, the performance of the estimates will improve.

3.2 Parameters of the simulation

A functional regression model (3) was simulated with an intercept $C_i = 1$ and errors from a normal distribution with variance $\sigma^2 = 2$, and $\delta = 3$ and $\alpha = 1$. The curves $X_i(t)$ were generated as Gaussian processes with mean=0 and variance=1, and correlation kernel $k_i(t,s) = \exp(-\kappa_i(t-s)^2)$, where κ_i was randomly generated from an exponential distribution with mean 1.

The data were generated from three different models:

1. An accumulation model where $\omega(t) = 1$;
2. A critical period model where $\omega(t) = \frac{10}{3(1+e^{-25(t-0.7)})}$; and
3. A sensitive period model where

$$\omega(t) = \begin{cases} 1.32(1-3t) & t \leq 1/3 \\ 0 & 1/3 < t \leq 2/3 \\ 3.3\sin(2\pi t - 4\pi/3) & t > 2/3. \end{cases} \quad (5)$$

For the accumulation model, ω is simply set to a constant. For the critical period model we parameterize ω as a sigmoid function, which is used to yield a smooth transition between the non-critical period and the critical period. This allows ω to meet the smoothness condition. For the sensitivity model, a general function is selected that is sparse over the interval $[1/3, 2/3]$. The three functions are illustrated in Figure 1b.

3.3 Numerical implementation

For the sparse and moderately sparse scenarios, the curves are estimated using maximum likelihood estimation for Gaussian processes. Regarding the choice of L, the selection of the number of B-splines bases is not crucial, as long as it is large enough to represent the complexity of the regression function[24].

In this regard, taking into account the model complexity, the number of splines is set to $L = 4, 6, 7$ for accumulation, critical and sensitive period models, respectively. We used the following prior distributions for the parameters:

$$\begin{aligned} \boldsymbol{\beta} &\sim Dir(1, ..., 1) \\ \delta &\sim \mathcal{N}(0, 10) \\ \alpha &\sim \mathcal{N}(0, 10) \\ \sigma &\sim \log\mathcal{N}(0, 1) \end{aligned}$$

Posterior distributions are obtained with MCMC simulations. To examine the properties of the fRLM to correctly identify the underlying life course model, the time interval is divided, for purposes of illustration, into three periods of equal lengths: $T_1 = [0, 1/3]$, $T_2 = (1/3, 2/3]$, and $T_3 = (2/3, 1]$. We then compute the posterior probability of the vector (w_1, w_2, w_3), where $w_j = \int_{T_j} \omega(t)dt$. The analyst could change these based on theoretical considerations. Integrals are computed using Riemann approximations for each MCMC sample.

3.4 Results of the simulation

Results of the simulation are reported in Table 1. We also report a summary of the convergence statistics and diagnostics in Table 3 in the Supplementary Material. The performance of the estimators is evaluated with the mean squared error between the estimates and the true underlying values for ω and δ:

$$mse_\omega = \int_0^1 (\hat{\omega}(t) - \omega(t))^2 dt, \quad mse_\delta = |\hat{\delta} - \delta|.$$

Following the SPT procedure, we report the posterior probability of the life course hypotheses for the omnibus test, $Pr(model|y)$. For results indicating a sensitive model, we report the best sequence of nested sub-models of the sensitive model and their posterior probability.

Table 1 reveals, as expected, that performance of the fRLM improves with n but also with the average number of time points N_i. The mse_ω and mse_δ decrease, for each life course model, from a sample size of 100 and 3-5 measurement occasions to a sample of 400 with completely observed trajectories. The probabilities associated with identifying the correct life course model suggest that 100 cases are insufficient, but probabilities exceed .90 in all situations involving 400 cases. The correct identification of the full rank submodel (i.e., $w2 < w1 < w3$) is achieved with 400 cases and three to five measurement occasions ($p = 0.995$).

4 Empirical Data Example

We use data from the National Longitudinal Study of Adolescent to Adult Health (Add Health),

Table 1: Performance metrics for the fRLM over 100 replications (median and Median Absolute Deviation (MAD))

n	Setup	mse_ω		mse_δ		$Pr(model\|y)$		$Pr(w_2 < w_1 < w_3\|y)$	
	Accumulation model								
100	3-5	0.051	(0.044)	0.116	(0.106)	0.514	(0.150)		
	6-8	0.041	(0.031)	0.073	(0.070)	0.666	(0.168)		
	∞	0.041	(0.030)	0.072	(0.063)	0.669	(0.156)		
400	3-5	0.034	(0.030)	0.092	(0.056)	0.940	(0.061)		
	6-8	0.020	(0.016)	0.044	(0.038)	0.992	(0.010)		
	∞	0.015	(0.013)	0.038	(0.035)	0.995	(0.006)		
	Critical model								
100	3-5	0.194	(0.153)	0.076	(0.067)	0.674	(0.216)		
	6-8	0.186	(0.130)	0.063	(0.042)	0.760	(0.176)		
	∞	0.174	(0.117)	0.068	(0.043)	0.752	(0.172)		
400	3-5	0.125	(0.107)	0.069	(0.042)	0.921	(0.085)		
	6-8	0.070	(0.039)	0.027	(0.027)	0.973	(0.030)		
	∞	0.064	(0.030)	0.032	(0.026)	0.974	(0.029)		
	Sensitive model								
100	3-5	0.192	(0.151)	0.158	(0.101)	0.724	(0.292)	0.749	(0.212)
	6-8	0.170	(0.145)	0.063	(0.061)	0.851	(0.172)	0.883	(0.129)
	∞	0.165	(0.126)	0.072	(0.050)	0.857	(0.176)	0.879	(0.132)
400	3-5	0.148	(0.069)	0.121	(0.057)	0.935	(0.090)	0.992	(0.012)
	6-8	0.128	(0.065)	0.043	(0.033)	0.972	(0.037)	0.999	(0.001)
	∞	0.116	(0.076)	0.030	(0.029)	0.980	(0.027)	1	(0.000)

Table 2: Omnibus test for posterior probability of the correct life course model

Signature	Accumulation	Sensitive	Critical
CKD	0	0.020	0.980
Inflammation	0.090	0.910	0.0005
Breast cancer	0.079	0.680	0.241

which is a nationally representative longitudinal study of US adolescents in grades 7-12 in 1994-1995 (age range 12-18) who were followed into adulthood over five waves of data collection[25]. The BMI trajectory was measured from: Wave I (12-18 years), Wave II (14-20 years), Wave III (18-26 years), Wave IV (24-32 years), and Wave V (33-43 years). During waves II, III, IV, and V, field examiners collected height and weight measurements for each respondent. Self-reported height and weight were available for waves I and V (measured height and weight were also collected during wave V). Wave V includes mRNA-seq abundance data from peripheral blood samples (for details of data collection protocol and the pre-processing of the data[26]).

We examine the association between BMI trajectories and three gene expression mRNA signatures: chronic kidney disease (CKD) (70 genes[27]), inflammation (751 genes[28]), and, for women only, breast cancer (BC) (44 genes[29,30]). We used Principal Component Analysis to reduce the dimensionality of each signature. The first principal component of each signature was used as the outcome.

We estimated model (3) with Bayesian Hamilton Monte Carlo Markov Chains. For participant i, the assessments of BMI are denoted as $x_{t_{i,1}}, x_{t_{i,2}}, \ldots, x_{t_{i,N_i}}$ performed at age $t_{i,1}, t_{i,1}, \ldots, t_{i,N_i}$. Participants whose weights were missing for more than 3 waves were excluded from the analysis and we thus have $3 \leq N_i \leq 5$. The resulting sample sizes were $n = 3708$ for CKD and Inflammation and $n = 2233$ for Breast cancer. Covariates include biological sex, age at wave V, number of hours fasting prior to blood draw, plate, use of anti-inflammatory medicines in the past four weeks, count of common clinical symptoms in the past four weeks (e.g., cold, fever, flu), count of common infectious and inflammatory diseases in the past four weeks (e.g., active infection, seasonal allergy) with correction for batch using ComBat[31]. The BMI trajectories $X_i(t)$ were predicted using Gaussian Process regression for each subject. To estimate the functional model, we used $L = 7$ density B-splines. The priors were

the same as in the simulations.

Table 3: Best sequence of partial rankings for the sensitive models for Inflammation and Breast cancer.

Signature	Ranking	Probability
Inflammation	3|1|2	0.570
	3|1,2	0.917
Breast cancer	1|2|3	0.781
	1,2|3	0.936

For the testing procedure, we selected $J = 3$ periods for illustrative purposes: T_1 = adolescence (age 12-18), T_2 = early-adulthood (age 19-29), and T_3 = mid-adulthood (age 30-40). The relative estimated importance of each period was computed by integrating the estimated weights, $w_j = |T_j|^{-1} \int_{T_j} \hat{\omega}(t)dt$. The ROPEs for the test statistics are selected as $[0, 0.2], (0.2, 0.8), [0.8, 1]$ for the accumulation, sensitive and critical model respectively. The results of the omnibus test and post hoc decompositions are described in Table 2 and 3. We also report a summary of the convergence statistics and diagnostics in Table 4 in the Supplementary Material.

Table 2 reports the Bayesian omnibus test of the three composite models as the posterior probability of the true composite model, i.e., $Pr(model|y)$, where $model \in \{\text{accumulation, critical, sensitive}\}$. For the sensitive model, Table 3 reports the probabilities of the finest credible rankings.

For CKD, the omnibus test unambiguously identifies the critical period variant as the correct model (probability = 0.98). Figure 2a indicates that time period 3, middle adulthood, corresponds to the critical period. Nevertheless, because of the design of the study, this conclusion is tentative because middle-adulthood may not be critical (i.e., an age period of heightened vulnerability) but rather it reflects recency, meaning that the last measurement occasion, no matter what age range it might cover, would produce the same result.

The omnibus test for inflammation points to a sensitive period model (probability = 0.91). The post hoc decomposition (Table 3) also reveals an unambiguous conclusion: that BMI in time periods 1 and 2 is a more powerful predictor of inflammation than BMI in time period 3 (probability = 0.917). This conclusion is further supported by Figure 2b. Inflammation in middle adulthood is thus predicted by BMI in adolescence and early adulthood.

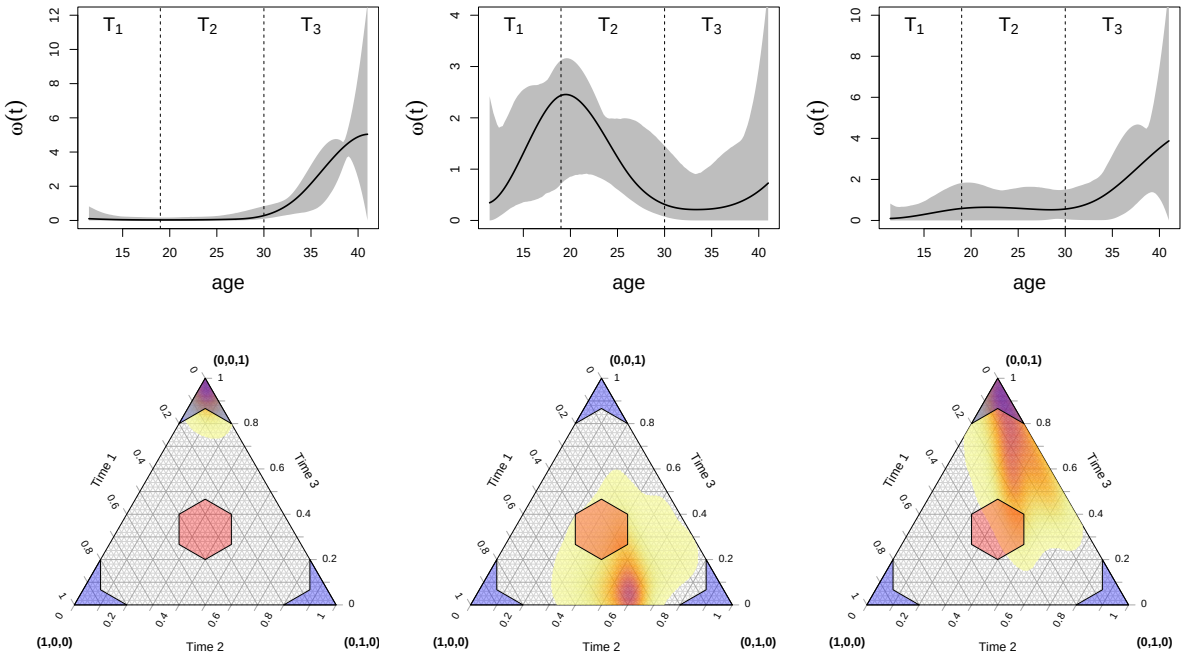
Table 2 reveals uncertainty, however, about the correct model for breast cancer, although the most warranted model is, once again, sensitive period ($p = 0.68$). The post hoc decomposition shows that effect of BMI is greatest at time 3 and the partial ranking of $1, 2|3$ is most supported (probability = 0.936). The plotted $\hat{\omega}(t)$ in Figure 2c may suggest a critical period for time 3, but the accompanying ternary plot shows considerable dispersion of the posterior distribution of weights beyond the ROPE. Thus, the breast cancer signature reflects BMI in middle adulthood, but the effects associated with adolescence and young adulthood are not negligible.

Finally, Figure 3 illustrates the patterns of BMI observed for 4 people and reveals considerable diversity in BMI trajectories: two subjects experienced precipitous increases in BMI but the other two people experienced positive and negative fluctuations. The relative importance of BMI for inflammation is shown in Figure 2b, and Figure 3b shows the relevant exposure, which is the product of the BMI trajectories and $\hat{\omega}(t)$ in (3). The relevant exposure shows relatively similar patterns, i.e. a bimodal configuration. However some subjects exhibit much higher relevant risk than other subjects, depending on the shape of their BMI trajectories.

5 Discussion

We propose the functional Relevant Life Course Model (fRLM), which considers discrete, sparse measurements as unobserved processes occurring in continuous time. This analytic goal is appropriate when the risk factors being studied reflect continuous processes (e.g., substance use, poverty or income trajectories, blood glucose). The fRLM defines the total lifetime exposure to risk as an integral (2) according to which exposures are assumed to be unobserved smooth functions. Because t refers to the exact age of the person, the fRLM is best suited to panel studies that begin with an age-heterogeneous group, although the model can also be applied to birth cohort studies. We also test life course hypotheses by applying Chumbley et al's SPT procedure[15] to our framework.

Simulations show that the performance of the fRLM improves with the number of repeated measurement occasions as expected, and the method is able to identify the correct life course model when $n = 400$ at least, even for very sparse designs with 3 repeated measurements per subject. Finally, the method is illustrated with three instructive empirical examples that examine the relationship between BMI trajectories from adolescence to middle adulthood and mRNA-seq expression signatures for chronic kidney disease, inflammation,



(a) CKD (b) Inflammation (c) Breast cancer

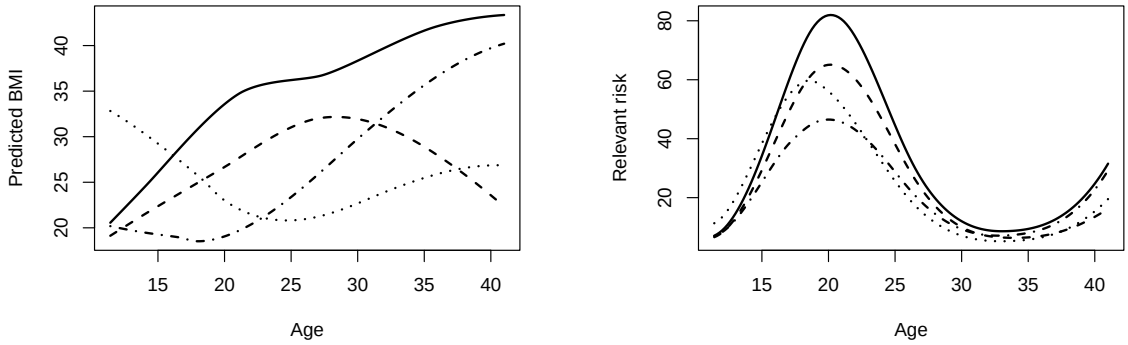
Figure 2: Estimation results for CKD, Inflammation, and breast cancer gene signatures. Upper panels show the estimated relative importance $\hat{\omega}(t)$ (black lines), confidence bands (gray regions), and separation of time periods (dashed lines). Lower panels show the posterior distribution of the weights $w_j = |T_j|^{-1} \int_{T_j} \omega(t)dt$ as well as the ROPEs for the accumulation model (red region) and critical model (blue region).

and breast cancer.

Note that the proposed model extends the RLM but also differs from approaches[9,10,11] that consider nonparametric estimation of the weights by regarding the observations as realizations of a continuous underlying process.

The closest model may be in the context of survival analysis[19] where a functional regression with a weight function satisfies $\int \omega(t)dt = 1$ but this is allowed to be negative. The implementation is frequentist and is performed by first estimating $\beta(t) = \delta\omega(t)$ and then identifying $\omega(t)$ by rescaling. The advantage of the present Bayesian implementation of the fRLM model is that it flexibly constrains parameters (i.e., the weight function is constrained to belong to a set of distributions by defining the appropriate prior distribution on the B-splines coefficients, the Dirichlet prior distribution). Prior models also consider densely, regularly spaced time points, whereas the fRLM allows for sparse and irregularly time points. In this way the time index t need not correspond to the timing of measurement occasions and is allowed to represent meaningful milestones based on the exact age of the subjects. Also, the model uses all available data and thus avoids limitations of methods for missing data. Finally, as discussed in the Supplementary Materials, other methods can be used to estimate the fRLM[21,22].

Franklin and his colleagues' review of suicidal behaviors notes several requisites for a successful empirical study of risk[32], which represent strategic opportunities to extend the fRLM. First, the fRLM can accommodate multiple risk factors as a straightforward additive functional linear model and interactions among different risks are also possible. Second, repeated assessments can also be modelled in dynamic terms, implemented with, for example, a function-on-function regression. Third, improvements in efficiency can be made by considering other processes (e.g., log-Gaussian process for positive data). Fourth, empirical studies of risk offer the promise of an increasingly personalized approach to health by providing people with a risk score, but such scores do not reflect the changing nature of risk across life (e.g., the Framingham risk score[33] and the CAIDE [Cardiovascular Risk Factors, Aging, and Incidence of Dementia] score to predict dementia[34]). The fRLM offers a method by which risk scores could reflect



(a) Life Course Exposure

(b) Relevant Life Course Exposure

Figure 3: For four randomly selected subjects, predicted pattern of Life Course Exposure, $\hat{X}_i(t)$, and estimated Relevant Life Course Exposure, $\hat{X}_i(t) * \hat{\omega}(t)$, for the Inflammation gene signature.

the changing nature of risk across the life course by, for example, reflecting the estimated relevant life course exposure, $\hat{X}_i(t) * \hat{\omega}(t)$.

Nevertheless, the fRLM has several limitations. First, the risk exposure is modeled as a Gaussian Process, which excludes modeling of data with binary or discretely-scaled risk exposures. Thus, the health outcome and repeated risk factor must be continuously scaled, which rules out, for example, the study of caseness defined by, for example, clinical cut-offs. Second, the fRLM can not test chain-of-risk models (e.g., a Markov autoregressive model with an earlier risk factor predicting its later value, which in turn predicts the outcome). Chain-of-risk models are intrinsically discrete-time, however, in contrast to the fRLM's depiction of risk as a continuous process. Finally, although we develop a broad framework for continuous risk exposure, some efficiency could be gained by setting priors that are more specific to the risk. For example, BMI is always positive, so it may improve the inference to set a positive prior distribution.

Despite these limitations, the fRLM offers a method by which discrete data can be used to model the experience of risk across many decades of life as a continuous process. Especially in the context of life course epidemiology, many risks are chronic and thus the focus on continuous process is likely more realistic than discrete-time models.

A package including all the functions to perform the analyses is included on GitHub at the following address https://github.com/jbodelet/fRLM. We also provide the simulations for reproducibility.

Funding

This work was supported by the Jacobs Foundation and by NIH Grants R01- HD087061 (MPIs K.H. Harris and M. J. Shanahan) specifically for the present analyses), P30-AG017265, R01-AG043404, and R01-AG033590; by the Swiss National Science Foundation (10531C-197964 to Shanahan); and by the Jacobs Center for Productive Youth Development (University of Zürich). Scott Hofer was supported by the National Institute on Aging (1R01AG067621). This research uses data from Add Health, a program directed by Robert Hummer and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris (University of North Carolina at Chapel Hill). The Add Health program is funded by Grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations (https://www.cpc.unc.edu/projects/addhealth/about/funders).

Acknowledgements

The authors thank Shawn Bauldry and Charles Driver for helpful comments.

Conflict of interest

None declared.

References

[1] Lynch J, Smith GD. A life course approach to chronic disease epidemiology. Annual review of public health. 2005;26(1):1-35.

[2] Kuh D, Ben-Shlomo Y, Lynch J, Hallqvist J, Power C. Life course epidemiology. Journal of epidemiology and community health. 2003;57(10):778.

[3] Zhu Y, Simpkin AJ, Suderman MJ, Lussier AA, Walton E, Dunn EC, et al. A structured approach to evaluating life-course hypotheses: moving beyond analyses of exposed versus unexposed in the-omics context. American journal of epidemiology. 2021;190(6):1101-12.

[4] Mishra G, Nitsch D, Black S, De Stavola B, Kuh D, Hardy R. A structured approach to modelling the effects of binary exposure variables over the life course. International journal of epidemiology. 2009;38(2):528-37.

[5] Smith AD, Heron J, Mishra G, Gilthorpe MS, Ben-Shlomo Y, Tilling K. Model selection of the effect of binary exposures over the life course. Epidemiology (Cambridge, Mass). 2015;26(5):719.

[6] Smith AD, Hardy R, Heron J, Joinson CJ, Lawlor DA, Macdonald-Wallis C, et al. A structured approach to hypotheses involving continuous exposures over the life course. International Journal of Epidemiology. 2016;45(4):1271-9.

[7] Madathil S, Joseph L, Hardy R, Rousseau MC, Nicolau B. A Bayesian approach to investigate life course hypotheses involving continuous exposures. International journal of epidemiology. 2018;47(5):1623-35.

[8] Vacek PM. Assessing the effect of intensity when exposure varies over time. Statistics in medicine. 1997;16(5):505-13.

[9] Hauptmann M, Wellmann J, Lubin JH, Rosenberg PS, Kreienbrock L. Analysis of exposure-time-response relationships using a spline weight function. Biometrics. 2000;56(4):1105-8.

[10] Madathil S, Rousseau MC, Joseph L, Coutlée F, Schlecht NF, Franco E, et al. Latency of tobacco smoking for head and neck cancer among HPV-positive and HPV-negative individuals. International Journal of Cancer. 2020;147(1):56-64.

[11] Sylvestre MP, Abrahamowicz M. Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. Statistics in medicine. 2009;28(27):3437-53.

[12] Potente C, Harris KM, Chumbley J, Cole SW, Gaydosh L, Xu W, et al. The Early Life Course of Body Weight and Gene Expression Signatures for Disease. American journal of epidemiology. 2021;190(8):1533-40.

[13] Madathil S, Blaser C, Nicolau B, Richard H, Parent MÉ. Disadvantageous socioeconomic position at specific life periods may contribute to prostate cancer risk and aggressiveness. Frontiers in oncology. 2018;8:515.

[14] Madathil SA, Rousseau MC, Durán D, Alli B, Joseph L, Nicolau B. Life course tobacco smoking and risk of HPV-negative squamous cell carcinomas of oral cavity in two countries. Frontiers in oral health. 2022:16.

[15] Chumbley J, Xu W, Potente C, Harris KM, Shanahan M. A Bayesian approach to comparing common models of life-course epidemiology. International journal of epidemiology. 2021;50(5):1660-70.

[16] Driver CC. Inference With Cross-Lagged Effects-Problems in Time and New Interpretations. 2022.

[17] Bhadra D, Daniels MJ, Kim S, Ghosh M, Mukherjee B. A Bayesian semiparametric approach for incorporating longitudinal information on exposure history for inference in case–control studies. Biometrics. 2012;68(2):361-70.

[18] Yang H, Li R, Zucker RA, Buu A. Two-stage model for time varying effects of zero-inflated count longitudinal covariates with applications in health behaviour research. Journal of the Royal Statistical Society Series C: Applied Statistics. 2016;65(3):431-44.

[19] Li X, Chang CCH, Donohue JM, Krafty RT. A competing risks regression model for the association between time-varying opioid exposure and risk of overdose. Statistical Methods in Medical Research. 2022;31(6):1013-30.

[20] Wang C, Liu H, Gao S. A penalized cox proportional hazards model with multiple time-varying exposures. The Annals of Applied Statistics. 2017;11:185-201.

[21] Yao F, Müller HG, Wang JL. Functional linear regression analysis for longitudinal data.

The Annals of Statistics. 2005;33(6):2873-903.

[22] Goldsmith J, Bobb J, Crainiceanu CM, Caffo B, Reich D. Penalized functional regression. Journal of computational and graphical statistics. 2011;20(4):830-51.

[23] Cai B, Meyer R. Bayesian semiparametric modeling of survival data based on mixtures of B-spline distributions. Computational statistics & data analysis. 2011;55(3):1260-72.

[24] Li Y, Ruppert D. On the asymptotics of penalized splines. Biometrika. 2008;95(2):415-36.

[25] Harris KM, Halpern CT, Whitsel EA, Hussey JM, Killeya-Jones LA, Tabor J, et al. Cohort profile: The national longitudinal study of adolescent to adult health (Add Health). International Journal of Epidemiology. 2019;48(5):1415-1415k.

[26] Shanahan MJ, Cole SW, Ravi S, Chumbley J, Xu W, Potente C, et al. Socioeconomic inequalities in molecular risk for chronic diseases observed in young adulthood. Proceedings of the National Academy of Sciences. 2022;119(43):e2103088119.

[27] Scherer A, Günther OP, Balshaw RF, Hollander Z, Wilson-McManus J, Ng R, et al. Alteration of human blood cell transcriptome in uremia. BMC medical genomics. 2013;6(1):1-13.

[28] Loza MJ, McCall CE, Li L, Isaacs WB, Xu J, Chang BL. Assembly of inflammation-related genes for pathway-focused genetic analysis. PloS one. 2007;2(10):e1035.

[29] Dumeaux V, Ursini-Siegel J, Flatberg A, Fjosne HE, Frantzen JO, Holmen MM, et al. Peripheral blood cells inform on the presence of breast cancer: A population-based case–control study. International journal of cancer. 2015;136(3):656-67.

[30] Dumeaux V, Fjukstad B, Fjosne HE, Frantzen JO, Holmen MM, Rodegerdts E, et al. Interactions between the tumor and the blood systemic response of breast cancer patients. PLoS computational biology. 2017;13(9):e1005680.

[31] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27.

[32] Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. Psychological bulletin. 2017;143(2):187.

[33] Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(18):1837-47.

[34] Ngandu T, Lehtisalo J, Solomon A, Levälahti E, Ahtiluoto S, Antikainen R, et al. A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (FINGER): a randomised controlled trial. The Lancet. 2015;385(9984):2255-63.