Recurrent Event Analysis in the Presence of Real-Time High Frequency Data via Random Subsampling

Walter Dempsey

Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

Abstract
Digital monitoring studies collect real-time high frequency data via mobile sensors in the subjects’ natural environment. This data can be used to model the impact of changes in physiology on recurrent event outcomes such as smoking, drug use, alcohol use, or self-identified moments of suicide ideation. Likelihood calculations for the recurrent event analysis, however, become computationally prohibitive in this setting. Motivated by this, a random subsampling framework is proposed for computationally efficient, approximate likelihood-based estimation. A subsampling-unbiased estimator for the derivative of the cumulative hazard enters into an approximation of log-likelihood. The estimator has two sources of variation: the first due to the recurrent event model and the second due to subsampling. The latter can be reduced by increasing the sampling rate; however, this leads to increased computational costs. The approximate score equations are equivalent to logistic regression score equations, allowing for standard, “off-the-shelf” software to be used in fitting these models. Simulations demonstrate the method and efficiency-computation tradeoff. We end by illustrating our approach using data from a digital monitoring study of suicidal ideation. Supplementary materials for this article are available online.

1. Introduction
Advancement in mobile technology has led to the rapid integration of mobile and wearable sensors into behavioral health (Free et al. 2013). Take HeartSteps, for example, a mobile health (mHealth) study designed to increase physical activity in sedentary adults (Klasnja et al. 2019). Here, a Jawbone sensor is used to monitor step count every minute of the participant’s study day. Of interest in many mHealth studies is the relation of such real-time high frequency sensor data to an adverse, recurrent event process. In a smoking cessation mHealth study (Spring 2019), for example, the relation between a time-varying sensor-based measure of physiological stress and smoking lapse is of scientific interest. In a suicidal ideation mHealth study (Kleiman et al. 2018), the relation of electrodermal activity (EDA) and accelerometer with self-identified moments of suicidal ideation is of scientific interest.

The goal of this article is to construct a simple, easy-to-implement method for parameter estimation and inference. To do so, we introduce a random subsampling procedure that has several benefits. First, the resulting inference is unbiased; however, there is a computation-efficiency tradeoff. In particular, a higher sampling rate can decrease estimator variance at the cost of increased computation. We show via simulations that the benefits of very high sampling rates is often negligible, as the contribution to the variation is small relative to the variation in the underlying stochastic processes. Second, derived estimating equations are optimal, implying loss of statistical efficiency is only due to subsampling procedure and not the derived methodology. Finally, implementation can leverage existing, standard software for functional data analysis and logistic regression, leading to fast adoption by domain scientists.

1.1. Related Work
The use of wearable devices to passively monitor patients has led to a rapid increase in the number of studies with high-frequency sensors that can be conceptualized as studies where measurements are functions. The development of such applications have been accompanied by intense methodological development in regression models with functional covariates (James 2002; James and Silverman 2005; Muller 2005; Ramsay and Silverman 2005b; Reiss and Ogden 2007; Crainiceanu, Staicu, and Di 2009; Kokoszka and Reimherr 2017).

Recently, modeling time-to-event data with functional covariates has received a fair bit of attention. The functional linear Cox regression model (FLCRM) considers the association between a time-to-event and a set of functional and scalar predictors (Kong et al. 2018). In this setting, the functional and scalar predictors are observed at baseline and the hazard function satisfies the proportional hazards assumption and involves a nonparametric baseline hazard with the exponential adjustment including both a functional linear model and a linear model for the baseline scalar predictors. Since the linear assumption may be too restrictive, Cui, Crainiceanu, and Leroux (2021) consider a more flexible additive functional
Cox model for multivariate functional predictors in which the hazard depends on an unspecified bivariate twice differentiable function.

In our setting, the high frequency sensor process is an outcome and must therefore be jointly modeled with the recurrent event process. Recently, Li, Xiao, and Luo (2022) proposed a functional approach to joint modeling of multivariate longitudinal biomarkers and a time-to-event outcome (Tsikritsis and Davidian 2004; Rizopoulos 2010). Here, as in traditional joint models, the longitudinal biomarker is observed at only a few observation times (typically on the order of tens of observations) and is modeled via a functional principal component analysis. The eigenvalues are the shared parameters across the longitudinal and survival submodels. A similar joint model was proposed in Dong et al. (2021). Inference proceeds by Monte Carlo EM which can require very large Monte Carlo sample size to ensure good performance, and naive implementation may not recover EM’s ascent property (Caffo, Jank, and Jones 2005).

In this article, we also consider joint modeling of recurrent events and a high frequency sensor process. In contrast to traditional longitudinal biomarkers, sensor processes are observed multiple times per second leading to thousand and/or millions of observations per individual. To address this increase in scale, a design-based perspective is taken in which we subsample nonevent times to avoid some of the computationally demanding aspects of joint models. Subsampling in the context of functional data analysis for massive data has recently been investigated (Liu, You, and Cao 2021) in which a functional L-optimality criterion is derived. Subsampling is different in our current context, as our goal is to subsample from the individual sensor processes to produce a design-unbiased estimate of the cumulative hazard function. Finally, traditional joint modeling often consider the hazard function to depend on the current or cumulative value of the health process (Rizopoulos 2010; Li, Xiao, and Luo 2022). For high frequency sensor processes, we expect the hazard function to only depend on the recent sensor history and use recent history functional linear models (Kim, Şentürk, and Li 2011) to account for this local dependence. We combine these models with advances in longitudinal functional data analysis (Staicu, Crainiceanu, and Carroll 2010) and penalized functional regression (Goldsmith, Zipunnikov, and Schrck 2015) to perform scalable inference using off-the-shelf software.

2. Main Contributions and Outline

The main contributions of this article are as follows: We define a joint model using historical functional linear models and demonstrate how techniques from design-based inference can be used to circumvent computational challenges of joint modeling in Section 3. We then discuss longitudinal functional principal components analysis and demonstrate in Lemma 4.1 that the resulting approximate score equations are equivalent to score functions for logistic regression with binary response and an offset related to the subsampling rate meaning our proposed approach can be fit using off-the-shelf software. An asymptotic analysis is presented in Section 4.4, with an accompanying discussion of the novel computational versus statistical efficiency tradeoff in Section 4.5. A simulation study in Section 5 demonstrates the impact of subsampling rates, the window length parameter, and a computational comparison to existing methods. A multivariate extension and method to handle missing data are presented in Section 6. We end by illustrating our approach using data from a digital monitoring study of suicidal ideation in Section 7.

3. Recurrent Event Process and Associated High Frequency Data

Suppose $n$ subjects are independently sampled with observed event times $T_i = \{T_{ij,1}, \ldots, T_{ij,k}\}$ over some observation window $[0, t_i]$ for each subject $i = 1, \ldots, n$. Assume the event times are ordered, that is, $T_{ij} < T_{ij'}$ for $j < j'$. The observation window length, $t_i$, is the censoring time and is assumed independent of the event process. Let $N_i(t)$ denote the associated counting process of $T_i$; that is, $N_i(t) = \sum_{j=1}^{k} I[T_{ij} < t]$. In this section, we assume a single-dimensional health process $x_i = (x_i(s))_{0 \leq s < t_i}$, for each participant is measured at a dense grid of time points. Accelerometer, for example, is measured at a rate of 32Hz (i.e., 32 times per second). Electrodermal activity (EDA), on the other hand, is measured at a rate of 4Hz (i.e., 4 times per second). Given the high frequency nature of sensor data, this article assumes the process is measured continuously. See Appendix A for a notation glossary.

Let $H_{i,t}^{NX} = H_{i,t}^{N} \otimes H_{i,t}^{X}$ be the σ-field generated by all past values $(N_i(s), x_i(s))_{0 \leq s < t_i}$. In this article, the instantaneous risk of an event at time $t$ is assumed to depend on the health process, time-in-study, and the event history through a fully parametric conditional hazard function:

$$h_i(t \mid H_{i,t}^{NX}; \theta) = \lim_{\delta \to 0} \delta^{-1} \Pr(N_i(t + \delta) - N_i(t) > 0 \mid H_{i,t}^{NX}),$$

(1)

where $\theta$ is the parameter vector. For high frequency physiological data, we assume that current risk is log-additive and depends on a linear functional of the health process over some recent window of time and pre-specified features of the counting process; that is,

$$h_i(t \mid H_{i,t}^{NX}; \theta) = h_0(t; \gamma) \exp\left(\frac{\gamma}{\alpha} + \int_{t-\Delta}^{t} x_i(s) \beta(s) ds\right)$$

(2)

where $h_0(t; \gamma)$ is a parameterized baseline hazard function, $\Delta$ is an unknown window-length, and $g(\hat{H}_{i,t}^{N}) \in \mathbb{R}^q$ is a $p$-length feature vector summarizing the event-history and time-in-study information. The final term $\int_{t-\Delta}^{t} x_i(s) \beta(s) ds$ reflects the unknown linear functional form of the impact of the time-varying covariate on current risk.

An alternative to (2) would be to construct features from the sensor data history $f_i(H_{i,t}^{X}) \in \mathbb{R}^q$ and incorporated these features in the place of the final term. Our current approach builds linear features of $H_{i,t}^{X}$ directly from the integrated history, avoiding the feature construction problem—a highly nontrivial issue for high frequency time-series data. The main caveat is the additional parameter $\Delta$; however, as long as the estimated $\hat{\Delta}$ exceeds $\Delta$, then resulting estimation is unbiased albeit at a loss of efficiency. Moreover, sensitivity analysis can be performed to determine how choice of $\hat{\Delta}$ affects inference. One limitation of the approach presented here is that only fully parametric hazard
models may be fit to the data. However, a spline model for the log baseline hazard affords sufficient model flexibility.

3.1. Likelihood Calculation

For the sake of notational simplicity, we leave the dependency of the conditional hazard function on $H_{i,t}^{NX}$ implicit, and write $h_i(t; \theta)$. In our current setting, we assume the health process $(x_i(t))_{0 \leq t < t_i}$ is directly observed. Therefore, we can consider the event process $T_i$ conditional on the health process $x_i$, which results in the log-likelihood related to the event process being then given by

$$L_n(\theta) = \sum_{i=1}^{n} \left( \sum_{j=1}^{k_i} \log \left( h_i \left( T_{ij}; \theta \right) \right) - H_i \left( T_{ij}; \theta \right) \right)$$

where $H_i(\tau; \theta) = \int_{0}^{\tau} h_i(t; \theta) dt$ is the cumulative hazard function. See Appendix B for additional arguments in favor of our proposed approach. Solving the associated score equations $U_n(\theta) = 0$ yields the maximum likelihood estimator $\hat{\theta}$,

$$U_n(\theta) = \sum_{i=1}^{n} \left( \sum_{j=1}^{k_i} h_i^{(1)} \left( T_{ij}; \theta \right) \right) - \sum_{i=1}^{n} \left( \sum_{j=1}^{k_i} h_i^{(1)} \left( T_{ij}; \theta \right) \right)$$

where $h_i^{(1)}(T_{ij}; \theta)$ and $H_i^{(1)}(\tau; \theta)$ are derivatives with respect to $\theta$.

In classical joint models (Henderson, Diggle, and Dobson 2000; Tsiatis and Davidian 2004), time-varying covariates $x_i(t)$ are observed only intermittently at appointment times. In our current setting, maximizing the likelihood is computationally prohibitive since for any $\theta$ we must compute the cumulative hazard functions $H_i(\tau; \theta)$ which requires integration of $h_i(t; \theta)$ given by (2) which itself depends on the integral $\int_{t}^{\tau} x_i(s) \beta(s) ds$ that is a function of the unknown functional parameter $\beta(\cdot)$. That is, the risk model now depends on an integrated past history of the time-varying covariate which leads to severe increase in computational complexity.

To circumvent these computational difficulties, we will derive approximate score equations based on design-based inference of point processes. Design-based inference is common for spatial point processes (Waagepetersen 2008) where the spatial varying covariate is observed at a random sample of locations. It is common in mobile health where ecological momentary assessments (Rathbun 2012; Rathbun and Shifman 2016) are used to randomly sample individuals at various time-points to assess their emotional state. In the current setting, we will leverage these ideas to form a subsampling protocol that can substantially reduce computationally complexity. Therefore, the purpose is quite different. Moreover, the dependence of the intensity function on the recent history of sensor values leads to additional complications that must be addressed.

3.2. Probabilistic Subsampling Framework

To solve the computational challenge we employ a point-process subsampling design to obtain unbiased estimates of the derivative of the cumulative hazards for each subject. The subsampling procedure treats the collected sensor data as a set of potential observations. Suppose covariate information is sampled at times drawn from an independent inhomogeneous Poisson point process with known intensity $\pi_i(t)$. At a subsampled time $t$, the windowed covariate history $(x_i(t - s))_{0 \leq s \leq \Delta}$ and counting process features $g_i(H_{i,t}^{NX})$ are observed. Optimal choice of $\pi_i(t)$ is beyond the scope of this article; however, simulation studies have suggested setting the subsampling rate proportional to the hazard function $h_i(t; \theta)$.

An estimator is design-unbiased if its expectation is equal to that parameter under the probability distribution induced by the sampling design (Cassel, Särndal, and Wretman 1977). Let $D_i \subset [0, t_i]$ denote the random set of subsampled points. Note, by construction, this random set is distinct from the set of event times with probability one, that is, $\text{pr}(T_i \cap D_i = \emptyset) = 1$. Under subsampling via $\pi_i(t)$, one may compute a Horvitz-Thompson estimator of the derivative of the cumulative hazard $\hat{H}_i^{(1)}(\tau; \theta) = \sum_{u \in D_i} \frac{h_i^{(1)}(u; \theta)}{\pi_i(u) + h_i(u; \theta)}$. An alternative design-unbiased estimator of the derivative of the cumulative hazards is given by

$$\hat{H}_i^{(1)}(\tau; \theta) = \sum_{u \in (T_i \cup D_i)} \frac{h_i^{(1)}(u; \theta)}{\pi_i(u) + h_i(u; \theta)}$$

Equation (3) is the estimator suggested by Waagepetersen (2008). This estimator depends on the superposition of the event and subsampling processes. Lemma 4.7 shows the estimator for $\theta$ associated with using (3) is the most efficient within a suitable class of estimators for the derivative of the cumulative hazard function (including the Horvitz-Thompson estimator). Therefore, we restrict our attention to (3) for the remainder of this article. Letting

$$w_i(t; \theta) = \frac{\pi_i(t)}{\pi_i(t) + h_i(t; \theta)},$$

the resulting approximate estimating equations can be rewritten as

$$\hat{U}_n(\theta) = \sum_{i=1}^{n} \sum_{u \in T_i} \frac{w_i(u; \theta) h_i^{(1)}(u; \theta)}{\pi_i(u) + h_i(u; \theta)} - \sum_{u \in D_i} w_i(u; \theta) \frac{h_i^{(1)}(u; \theta)}{\pi_i(u)}$$

Equation (5) represents the approximate score functions built via plug-in of the design-unbiased estimator of the derivative of the cumulative hazard given in (3).

4. Longitudinal Functional Principal Components Within Event-History Analysis

Probabilistic subsampling converts the single sensor stream $x_i$ into a sequence of functions observed repeatedly at sampled times $D_i$ and event times $T_i$ over windows of length $\Delta$. Such a data structure is commonly referred to as longitudinal functional data (Xiao, Li, and Ruppert 2013; Goldsmith, Zipunnikov, and Schrck 2015). Given the large increase in longitudinal functional data in recent years, corresponding analysis has received much recent attention (Morris et al. 2003; Morris and Carroll 2006; Baladandayuthapani et al. 2008; Di et al. 2009; Greven et al. 2010; Staicu, Crainiceanu, and Carroll 2010; Chen and Müller 2012; Li and Guan 2014). Here, we combine work by Park and
Staicu (2015) and Goldsmith et al. (2011) to construct a computationally efficient penalized functional method for solving the estimation equations \( \hat{U}_n(\theta) \).

### 4.1. Estimation of the Windowed Covariate History

We start by defining \( X(t, s) = x(t - s) \) to be the sensor measurement \( 0 \leq s \leq \Delta \) time units prior to time \( t \in T \cup D \). The sandwich smoother (Xiao, Li, and Ruppert 2013) is used to estimate the mean \( \mu_y(t, s) = \mathbb{E}_y[X(t, s)] \) where the expectation is indexed by whether \( t \) is an event \( (y = 1) \) or subsampled \( (y = 0) \) time respectively. Alternative bivariate smoothers exist, such as the kernel-based local linear smoother (Hastie, Tibshirani, and Friedman 2009), bivariate tensor product splines (Wood 2006), and the bivariate penalized spline smoother (Marx and Eilers 2006). The sandwich smoother was chosen for its computational efficiency and estimation accuracy. We then define \( \hat{X}(t, s) = X(t, s) - \hat{\mu}_y(t, s) \) to be the mean-zero process at each time \( t \in T \cup D \).

As in Park and Staicu (2015), define the marginal covariance by

\[
\Sigma_y(s, s') = \int_0^T c_y((T, s), (T, s'))f_y(T)dT
\]

for \( 0 \leq s, s' \leq \Delta \), where \( c_y((t, s), (t, s')) \) is the covariance function of the windowed covariate history \( X(t, .) \), \( \tau \) is the observation window length of the process, and \( f_y(T) \) is the intensity function for event \( (y = 1) \) and subsampled \( (y = 0) \) times, respectively. Estimation of \( \Sigma_y \) occurs in two steps. For simplicity, we present the steps for subsampled times (i.e., \( y = 0 \)) but the steps are the same for event times as well. First, the pooled sample covariance is calculated at a set of grid points:

\[
\hat{\Sigma}_0(s_r, s_{r'}) = \left( \sum_{i=1}^n |D_i| \right)^{-1} \left( \sum_{i=1}^n \sum_{t \in D_i} \hat{X}(t, s_r)\hat{X}(t, s_{r'}) \right).
\]

As we assume the health process \( x_i \) is directly observed, the diagonal elements of \( \hat{\Sigma}_0 \) are not inflated. Second, the estimator \( \hat{\Sigma} \) is further smoothed again using the sandwich smoother (Xiao, Li, and Ruppert 2013). Note Park and Staicu (2015) smooth the off-diagonal elements, while here we smooth the entire pooled sample covariance matrix. All negative eigenvalues are set to zero to ensure positive semi-definiteness. The result is used as an estimator \( \hat{\Sigma} \) for the pooled covariance \( \Sigma_0 \).

Next, we take the spectral decomposition of the estimated covariance function; let \( \hat{\psi}_k(0, s), \hat{\lambda}_k \) be the resulting sequence of eigenfunctions and eigenvalues. The key benefit of the marginal covariance approach is that it allows us to compute a single, time-invariant basis expansion; this reduces the computational burden by avoiding the three dimensional covariance function (i.e., covariance depends on \( t \)) and associated spectral decomposition in methods considered by Chen and Müller (2012). Using the Karhunen-Loève decomposition, we can represent \( \hat{X}(t, s) \) for \( t \in T \cap D \) by

\[
X(t, s) = \hat{\mu}_y(t, s) + \sum_{k=1}^{\infty} \hat{c}_k(t) \hat{\psi}_k(s) \approx \hat{\mu}_y(t, s) + \hat{c}_k(t) \hat{\psi}_k(s)
\]

where \( \hat{c}_k(t) = \int_{t-\Delta}^{t} \hat{X}(t, s) \hat{\psi}_k(s) ds \), \( \hat{c}_k(t) = (\hat{c}_{1,k}(t), \ldots, \hat{c}_{k,k}(t)) \), \( \hat{\psi}_k(s) = (\hat{\psi}_1(s), \ldots, \hat{\psi}_{k}(s))^T \), and \( K < \infty \) is the truncation level of the infinite expansion. Following Goldsmith et al. (2011), we set \( K_s \) to satisfy identifiability constraints (see Section 4.2 for details). In subsequent sections, we leave the dependence on \( y \) (i.e., whether \( t \in T \) or \( t \in D \)) implicit unless required for notational simplicity.

### 4.2. Estimation of \( \beta \)

The next step of our method is modeling \( \beta(t) \). Here, we leverage ideas from the penalized spline literature (Ruppert, Wand, and Carroll 2003; Wood 2003). Let \( \phi(t) = \{\phi_1(t), \ldots, \phi_{K_b}(t)\} \) be a spline basis and assume that \( \beta(t) = \sum_{j=1}^{K_b} b_j \phi_j(t) = \beta(t)b \) where \( b = [b_1, \ldots, b_{K_b}]^T \). Thus, the integral in (2) can be restated as

\[
\int_{t-\Delta}^{t} X(t, s)\beta(t)ds \approx \int_{t-\Delta}^{t} \left[ \hat{\mu}(t, s) + c(t)^T \hat{\psi}_s(s) \right] \times [\phi(s)b] ds
\]

where \( M_t = (M_{t,1}, \ldots, M_{t,K_b}) \), \( M_{t,j} = \int_{t-\Delta}^{t} \hat{\mu}(t, s) \phi_j(s)ds \), and \( \phi_s(b) \) is a \( K_b \times K_b \) dimensional matrix with the \( (k,l) \)th entry is equal to \( \int_{0}^{\Delta} \hat{\psi}_k(s) \phi_l(s)ds \) (Ramsay and Silverman 2005a).

Given the basis for \( \beta(t) \), the model depends on choice of both \( K_b \) and \( K_x \). We follow Ruppert (2002) by choosing \( K_b \) large enough to prevent under-smoothing and \( K_x \geq K_b \) to satisfy identifiability constraints. While our theoretical analysis considers truncation levels that depend on \( n \), in practice, we follow the simple rule of thumb and set \( K_b = K_x = 35 \). As long as the choices of \( K_b \) and \( K_x \) are large enough, their impact on estimation is typically negligible. Below, we will exploit a connection between (5) and score equations for a logistic regression model. Before moving on, we introduce some additional notation. Define

\[
h_i(t | H_i^{N[X]}; \gamma) \approx \exp \left( Z_i^\gamma + g_i(H_i^{N[X]}) \alpha + M_{t,i}^T b + C_{i,t} I_{\phi,\psi} b \right)
\]

where \( \theta = (\gamma, \alpha, b) \) and \( \exp(Z_i^\gamma) = h_0(t) \) is the parameterized baseline intensity function. We write \( \hat{U}_n(\theta) \) to denote the approximate score function when substituting in (6) for (2).

### 4.3. Connection to Logistic Score Functions

We next establish a connection between the above approximate score equations \( \hat{U}_n(\theta) \) and the score equations for a logistic regression model. We can then exploit this connection to allow the model to be fit robustly using standard mixed effects software (McCulloch and Searle 2001; Ruppert 2002).

**Lemma 4.1.** Under weights (4) and the log-linear intensity function (6), the approximate score function \( \hat{U}_n(\theta) \) is equivalent to

\[
\sum_{i=1}^{n} \sum_{t \in T_i \cup D_i} \left[ 1[t \in D_i] - \frac{1}{1 + \exp\left(-\left(\hat{W}_{i,t}^\top \theta + \log \pi_i(t)\right)\right)} \right] \hat{W}_{i,t}
\]
where $\tilde{W}_{i,t} = -W_{i,t}$. This is the score function for logistic regression with binary response $Y_i(t)$ for $t \in T_i \cup D_i$ and $i \in [n]$ where $Y_i(t) = 1$ if $t \in T_i$, offset log $\pi_i(t)$, and covariates $\tilde{W}_{i,t}$.

The connection established by Lemma 4.1 between our proposed methodology and logistic regression allows us to leverage “off-the-shelf” software. The main complication is pre-processing of the functional data; however, these additional steps can also be taken care of via existing software. Therefore, the entire data analytic pipeline is easy-to-implement and requires minimal additional effort by the end-user. To see this, we briefly review the proposed inference procedure.

**Remark 4.2 (Inference procedure review).** Given observed recurrent event and high frequency data \{\(T_i, X_i\)\}_i^{n} = 1,

(a) For each $i \in [n]$, sample nonevent times as a time-inhomogeneous Poisson point process with intensity according to $\pi_i(t)$.

(b) Estimate mean $\mu_i(\theta)$ for $0 \leq s \leq \Delta$ at all event times $t \in \bigcup_{i=1}^{n} T_i$ and sampled nonevent times $t \in \bigcup_{i=1}^{n} D_i$.

(c) Compute marginal covariance across event times, $\Sigma_1$, and nonevent times, $\Sigma_0$.

(d) Compute eigendecomposition \{\(\tilde{\psi}_k, \lambda_k\)\} of marginal covariance $\Sigma_y$.

(e) Use the eigendecomposition to construct $\tilde{W}_{i,t}$ for all $i \in [n]$ and $t \in D_i \cup T_i$.

(f) Perform logistic regression with binary outcome \(\{Y_i(t)\}\) and offset of log $\pi_i(t)$.

Before demonstrating the methodology via simulation in Section 5 and a worked example in Section 7, we provide a theoretical analysis of our current proposal.

### 4.4. Theoretical Analysis

Our theoretical analysis requires assumptions regarding the subsampling procedure, the event process, and the functional data. We state these assumptions and then our main theorems. We start by assuming there exists $\tau < \infty$ such that all individuals are no longer at risk (i.e., $\tau_i < \tau$ for all $i$). Moreover, define $R_i(t)$ to be the at-risk indicator for participant $i$, that is $R_i(t) = 1$ for $t \in (0, \tau_i)$. Asymptotic theory provided in Lemma 4.6 will be proven under regularity conditions A-E in Andersen et al. (1993), pp. 420–421 along with the following additional assumptions:

**Assumption 4.3 (Event process assumptions).** We assume the following holds:

(E.1) The subsampling rate is both lower and upper bounded at all at-risk times; that is, $0 < L < \pi_i(t) < U < \infty$ for all $i = 1, 2, \ldots$ and $t \in [0, \tau]$ such that $R_i(t) = 1$.

(E.2) There exists a nonnegative definite matrix $\Sigma(\theta)$ such that

$$n^{-1} \Sigma_n(\theta) = n^{-1} \sum_{i=1}^{n} \int_{0}^{\tau} w_i(t; \theta)$$

$$\times \begin{bmatrix} h_i^{(1)}(t; \theta) h_i^{(1)}(t; \theta)^\top \end{bmatrix}$$

$$\times \frac{\hat{h}_i(t; \theta)}{h_i(t; \theta)} R_i(t) dt \overset{p}{\to} \Sigma(\theta).$$

(E.3) There exists $M$ such that $|W_{i,j,t}| < M$ for all $(i, j, t)$.

(E.4) For all $j, k$

$$n^{-1} \sum_{i=1}^{n} \int_{0}^{\tau} \left| \frac{d^2}{dt^2} \frac{\hat{h}_i(t; \theta_0)}{R_i(t)} \right|^2 R_i(t) dt \overset{p}{\to} C < \infty$$

as $n \to \infty$.

We also require several assumptions due to the truncation of the Karhunen-Loève decomposition that represents $X(t, s)$.

**Assumption 4.4 (Functional assumptions (Park and Staicu 2015)).** The following assumptions are standard in prior work on longitudinal functional data analysis (Yao, Müller, and Wang 2005; Chen and Müller 2012; Park and Staicu 2015):

(A.1) $X = \{X(t, s) : (t, s) \in T \times S\}$ is a square integrable element of $L^2(T \times S)$.

(A.2) The subsampling and conditional intensity rate functions $f_j(T)$ are continuous and $\sup|f_j(T)| < \infty$.

(A.3) $\mathbb{E}X(t, s)X(t', s')X(t', s') < \infty$ for each $s, s' \in [0, \Delta]$ and $0 < t, t' < \tau$.

(A.4) $\mathbb{E}||X(t, s)||^4 < \infty$ for each $0 < t < \tau$.

Finally, for simplicity, we assume that there exists $b^*$ such that $\beta(t) = \phi(t)b^*$; that is, the true function $\beta(t)$ sits in the span of the spline basis expansion.

**Remark 4.5 (Practical consequences of Assumptions 4.3 and 4.4).** Assumptions 4.3 and 4.4 contain as a special case the scenario where individuals are independent and identically distributed, the functional process is bounded (i.e., $|X(t, s)| < M$ for some $M < \infty$), and the subsampling rate is both lower and upper bounded at all risk times. As such bounds are likely to be true in most practical settings, this demonstrates the reasonableness of our assumptions for applied settings.

**Lemma 4.6.** Under Assumptions 4.3, 4.4, and $\Delta$ known, for large $n$ the estimator $\hat{\theta}_n$ is consistent; moreover,

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{D}{\to} N(0, \Sigma(\theta)^{-1}),$$

where $\overset{D}{\to}$ is convergence is distribution and

$$\Sigma(\theta) = \int_{0}^{\tau} w(s; \theta) \times \frac{\pi^{(1)}(s; \theta) \times \pi^{(1)}(s; \theta)^\top}{\pi^{(1)}(s; \theta)} ds.$$ and $\tau$ is the random censoring time of the event process.

Proof of Lemma 4.6 is presented in Appendix C.2. A design-unbiased estimator for $\Sigma(\theta)$ is

$$\hat{\Sigma}(\theta) = n^{-1} \sum_{i=1}^{n} \sum_{t \in T_i \cup D_i} w_i(t; \theta) (1 - w_i(t; \theta)) \frac{h_i^{(1)}(t; \theta)}{\hat{h}_i(t; \theta)}$$

$$\times \frac{\hat{h}_i^{(1)}(t; \theta)}{h_i(t; \theta)}.$$

For the log-linear intensity model, the sampling-unbiased estimator for $\hat{\Sigma}(\theta)$ is equivalent to the Fisher information for the previously described logistic regression model. This implies that
subsampling from an inhomogeneous Poisson process, standard logistic regression software can be used to fit the recurrent event model by specifying an offset equal to log \( \pi(t) \). Based on this, we can leverage existing inferential machinery to obtain variance-covariance estimates of model parameters. That is, if \( \Sigma_{bb} \) is the \( K_b \times K_b \) dimensional matrix obtained by plugging in the estimates of variance components into the formula for the variance of \( \hat{b} \), then the standard error for estimate at time \( t_0 \) — that is, \( \hat{\beta}(t_0) = \phi(t_0) \hat{b} \) — is given by \( \sqrt{\phi(t_0) \Sigma_{bb} \phi(t_0)^\top} \). Then the approximate 95% confidence interval can be constructed as \( \hat{\beta}(t_0) \pm 1.96 \sqrt{\phi(t_0) \Sigma_{bb} \phi(t_0)^\top} \). We acknowledge an important limitation of confidence intervals obtained via this approach. Specifically, we ignore the variability inherent in the longitudinal functional principal component analysis; that is, our estimates ignore the variability in estimation of eigenfunctions \( \psi \) as well as the coefficients \( \hat{c}_{ik}(t) \). Joint modeling could be considered as in Crainiceanu and Goldsmith (2010), however, this is beyond the scope of this article.

Lemma 4.7 shows (4) is optimal within a class of weighted estimating equations. The result ensures the only loss of statistical efficiency is due to subsampling and not using a suboptimal estimation procedure given subsampling. Here, weights \( w^* \) are considered optimal if the difference between the asymptotic variance \( V(\hat{\beta}(t_0); w^*) \) and the asymptotic variance under any other choice of weights \( w \), \( V(\hat{\beta}(t_0); w) \) is positive semidefinite, that is, any linear contrast has smaller asymptotic variance under weight \( w^* \) than under weight \( w \).

**Lemma 4.7.** If the event process is an inhomogeneous Poisson point process with intensity \( h(t; \theta) \) and subsampling occurs via an independent, inhomogeneous Poisson point process with intensity \( \pi(t) \), then \( \hat{U}_n(\theta) \) are optimal estimating functions (i.e., most efficient) in the class of weighted estimating functions given by (5) replacing (4) by any weight function \( w_j(t; \theta) \). This class includes the Horvitz-Thompson estimator under \( w(s; \theta) = 1 \).

Proof of Lemma 4.7 is presented in Appendix D.

### 4.5. Computation Versus Statistical Efficiency Tradeoff

We next consider the statistical efficiency of our proposed estimator when compared to complete-data maximum likelihood estimation. While subsampling introduces additional variation, it may significantly reduce the overall computational burden. It is this tradeoff that we next make precise. In particular, we consider the following choice of subsampling rate, \( \pi(t) = c \times h(t; \theta) \) for \( c > 0 \). That is, the subsampling rate is proportional to the intensity function with time-independent constant \( c > 0 \). Under this subsampling rate, the weight function (4) is equal to \( c/(c + 1) \). Under Lemma 4.6,

\[
\Sigma(\theta) = \frac{c}{c + 1} \int_0^\tau h^{(1)}(t; \theta) h^{(1)}(t; \theta)^\top dt = \frac{c}{c + 1} \Sigma(\theta)
\]

where \( \Sigma(\theta) \) is the Fisher information of the complete-data maximum likelihood estimator. Therefore, the relative efficiency is \( c/(1 + c) \). For an upper bound \( H = \max_{t \in (0, \tau)} h(t; \theta) \), if

**Table 1.** Data reduction (total # of measurements divided by expected number of subsampled measurements) given sensor rate, subsampling constant and an upper bound on the intensity rate.

| Sensor Rate | Subsampling Constant (c) | Upper Bound on Intensity Rate per Hour | Statistical Efficiency |
|-------------|--------------------------|----------------------------------------|------------------------|
|             |                          | 0.5                                    | 1                      | 3                      | 5                      | 10                     |
| 4Hz (EDA)   | 5                        | 5760                                   | 2880                   | 960                    | 576                    | 288                    | 0.833                  |
|             | 10                       | 2880                                   | 1440                   | 480                    | 288                    | 144                    | 0.909                  |
|             | 100                      | 288                                    | 144                    | 48                     | 29                     | 14                     | 0.990                  |
| 32Hz (AI)   | 5                        | 46080                                  | 23040                  | 7680                   | 4608                   | 2304                   | 0.833                  |
|             | 10                       | 23040                                  | 11520                  | 3840                   | 2304                   | 1152                   | 0.909                  |
|             | 100                      | 2304                                  | 1152                   | 384                   | 230                   | 115                    | 0.990                  |

we set \( \pi(t) = c \times H \), then the relative efficiency can be lower bounded by \( c/(c + 1) \).

Sensor measurements occur multiple times per second. Suppose the intensity rate is bounded above by 1 and the unit time scale was hours. If we then subsample the data at a rate of 10 times per hour, then we have a lower bound on the efficiency of 0.909. For a 4Hz sensor, this reduces the number of samples per hour from \( 4 \times 60 \times 60 = 14,400 \) per hour to on average 10 per hour. While the computational complexity of logistic regression is linear in the number of samples, we get 1440 times reduction in the data size at the cost of a 0.909 statistical efficiency. If we sample 100 times per hour, then the efficiency loss is only 0.999, with a 144 times reduction in data size. Table 1 provides additional examples for a 4Hz and 32Hz sensor rate respectively. The data reduction depends on this rate; however, the lower bound on statistical efficiency does not because the subsampling rate only depends on the upper bound of the intensity function. In particular, if the events are rare then subsampling rate can be greatly reduced with no impact to statistical efficiency.

### 4.6. Penalized Functional Regression Models

Recall theoretical results were proven under the assumption that there exists \( b^* \) such that \( \beta(t) = \phi(t)b^* \). To make this assumption plausible, we set \( K_b \) large enough (but less than \( K_x \) to ensure identifiability) to ensure the spline basis expansion is sufficiently expressive. However, in practice, such a choice of \( K_b \) may lead to overfitting the data. Following Goldsmith et al. (2011), we choose the linear spline model and set \( \phi(t) = b_0 + b_1 t + \sum_{j=2}^{K_b} (t - \kappa_j)_+ \) where \( \kappa_j \) are the chosen knots and assume \( \{b_j\}_{j=2}^{K_b-1} \sim N(0, \sigma^2 I) \) to induce smoothness on the spline model. Combining the penalized spline formulation with Lemma 4.1 establishes a connection between our approximate score equations and solving a generalized mixed effects logistic regression with offset. Given the connection with generalized mixed effects models, the inferential machinery to obtain variance-covariance estimates. As we leverage the standard R package "glmnet," the smoothness parameter is chosen via cross-validation. In this context, we acknowledge another limitation. Specifically, penalization may lead to confidence intervals that perform poorly in regions where \( \hat{\beta}(t) \) is over-smoothed.

### 5. Simulation Study

We next assess the proposed methodology via a simulation study. Here, we assume each individual is observed over five
days where each day is defined on the unit interval [0, 1] with 1000 equally spaced observation times per day. We define $X(t)$ at the grid of observations as a mean-zero Gaussian process with covariance

$$
\Sigma(t, t') = \frac{\sigma^2}{\Gamma(v)2^{v-1}} \left( \frac{\sqrt{2v} |t - s|}{\rho_1} \right)^v K_v \left( \frac{\sqrt{2v} |t - s|}{\rho_1} \right)
$$

where $K_v$ is the modified Bessel function of the second kind. We set $v = 1/2$, $\sigma^2 = 1$, and $\rho = 0.3$ as well as set $K_0 = K_\pi = 35$. For simplicity, we assume $\Sigma$ is known in computation of the eigendecompositions. Given $X(t)_{t_j \leq t \leq t_{j+1}}$ for a given user-day, we generate event times according a chosen hazard function $h(t; \theta)$. To mimic our real data, we set

$$h(t; \theta) = \exp \left( \theta_0 + \int_0^\Delta X(t - s) \beta(s) ds \right).$$

We set $\Delta$ to mimic a 30-min window for a 12-hr day. We set $\theta_0 = \log(5/1000)$ to set a baseline risk of approximately 5 events per day. We consider two choices of $\beta(s)$: (1) $\beta_0 + \exp(-\beta_1 s)$ which decays to 0 as $s$ approaches $\Delta$ from below, and (2) $\beta_1 \sin(2\pi \frac{s}{6}) - \pi/2$ which is significantly different from 0 as $s$ approaches $\Delta$ from below.

We generate 1000 datasets, each consisting of 500 user-days. For a given simulated user-day, we randomly sample non-event times using a Poisson process with rate of every 5 min. We use the proposed methodology to construct the estimate $\hat{\beta}_{i,12}(t)$ for the ith simulated dataset; we then subsample the sampled non-event times with thinning probability $1/3$, $1/6$, $1/12$, and $1/24$. This results in randomly sampled non-event times given by a Poisson process with rates of every 15 min, half-hour, hour, and 2 hr. We can construct the corresponding estimates $\hat{\beta}_{i,4}(t)$, $\hat{\beta}_{i,12}(t)$, $\hat{\beta}_{i,1}(t)$, and $\hat{\beta}_{i,0.05}(t)$, respectively. Subsampling allows us to compare the variance due to subsampling as compared to the variance due to sampling fewer non-event times.

Since we are primarily interested in accuracy, we report the mean integrated squared error (MISE) defined as

$$\frac{1}{1000} \sum_{i=1}^{1000} \int_0^\infty (\hat{\beta}_{i,j}(s) - \beta(s))^2 ds$$

for each $j$ where $\beta(s) \equiv 0$ for $s > \Delta$. The MISE is defined in this manner to account for settings where $\Delta$ is unknown. Next, let $\hat{\beta}_{i,j}(t) = \frac{1}{1000} \sum_{i=1}^{1000} \hat{\beta}_{i,j}(t)$ denote the average estimate for $j = 0.5, 1, 2, 4$. Then the squared bias is given by $\frac{1}{1000} \sum_{i=1}^{1000} \int_0^\infty (\hat{\beta}_{i,j}(s) - \beta(s))^2 ds$ and the subsampling variance is defined as $\frac{1}{1000} \sum_{i=1}^{1000} \int_0^\infty (\hat{\beta}_{i,j}(s) - \hat{\beta}_{i,12}(s))^2 ds$. Table 2 show the MISE decomposed into the variance and squared biased as well as the subsampling variance. To allow fair comparisons across the two choices of $\beta(s)$, all reported numbers are scaled by the integrated square of the true function $\int_0^\infty \beta(s)^2 ds$. The relative MISE (RMISE) to the lowest sampling rate and the average runtime (in seconds) are also reported.

**Remark 5.1** (Time-complexity and run-time of maximum likelihood estimation). The complexity of logistic regression with $n$ observations and covariates of dimension $d$ is $O(n d)$. Maximum likelihood estimation can be well approximated by subsampling at a very high rate; therefore, under a 4Hz sensor the time-complexity of the maximum likelihood estimate is $(4 \times 60 \times \pi \cdot n \cdot d$ compared to the expected time-complexity under subsampling with rate $c$ per minute of $(c \cdot \pi \cdot n \cdot d$. This means using the sensor at its observation frequency of four times per second will take approximately $2^{13}$ times as long as subsampling at a rate of once every half-hour. In our first example, using the average run times at each sampling rate to project run time at four times per second yields an approximately 2.0 hr run time. In our second example, we project a run time at four times per second of 6.2 hr. In both instances, the relative efficiency gain would be negligible suggesting a huge computational increase for minimal relative information gain.

### 5.1. Impact of $\Delta$

A concern with the proposed approach is the selection of window-length, $\Delta$. Here we investigate the impact of misspecification of the window length for $\beta(t) = \beta_1 \cdot \sin(2\pi t/\Delta - \pi/2)$ and the true window length ($\Delta^*$) is set to 32-min. See Section E.1 of the supplementary materials for a similar discussion for $\beta(t) = \beta_0 \exp(-\beta_1 s)$. As in the previous simulation, we generate 1000 datasets per condition each consisting of 500 user-days. For each simulation, we analyze the data using window lengths $\Delta \in (26, 29, 32, 35, 37)$.

When the window length is too large, that is, $\Delta > \Delta^*$, then asymptotically the estimation is unbiased as $\beta(t) \equiv 0$ for $t > \Delta$; however, we incur a penalty in finite samples, especially for settings where the function is far from zero near $t = \Delta$. We find the MISE increases as the absolute error $|\Delta - \Delta^*|$ increases. While the MISE increased for $\Delta < \Delta^*$, the pointwise estimation error remains low for $t < \Delta$. This does not hold for $\Delta > \Delta^*$, where instead we see parameter attenuation, that is, a bias toward zero in the estimates at each $0 < t < \Delta$. To capture this, we define a partial MISE as $\frac{1}{\Delta} \int_0^\Delta (\hat{\beta}_{i,j}(s) - \beta(s))^2 ds$ where $\Delta = \min(\Delta, \Delta^*)$, which is the MISE on the subset $0 < t < \min(\Delta, \Delta^*)$ and scaled for comparative purposes.

### 5.2. Selection of Bandwidth

While above we explore the bias-variance tradeoff under bandwidth misspecification, here we explore a data-adaptive method for bandwidth selection. This is a critical consideration for recent history functional linear models (Kim, Şentürk, and Li 2011). Given the model complexity does not change as a function of $\Delta$ in our simulations, our proposal is to compare AIC across a range of bandwidths and choose that which maximizes the criterion. Table 4 presents AIC-based selection across the 500 simulated datasets analyzed in the prior section. We see markedly distinct behavior of the selection criterion in the two settings. The AIC-based selection works well for the sinusoidal effect across subsampling rates, but performs poorly in the exponential setting. Recall that the sinusoidal effect is significantly different from 0 at $\Delta^*$ and then 0 for all $s > \Delta^*$.
Table 2. Mean-integrated squared error, variance, squared bias, and subsampling (SubS) variance for \( \beta(s) \) given by (1) and (2), respectively.

|           | Sampling rate (per hour) |          |          |          |          |
|-----------|--------------------------|----------|----------|----------|----------|
| SubS. Variance |                          | 12       | 4.0      | 2.0      | 1.0      | 0.5      |
|            | –                        | 1.2 \times 10^{-5} | 2.0 \times 10^{-5} | 3.5 \times 10^{-5} | 8.6 \times 10^{-5} |
| Variance   | 1.3 \times 10^{-2}      | 1.5 \times 10^{-2} | 2.1 \times 10^{-2} | 2.7 \times 10^{-2} | 3.8 \times 10^{-2} |
| Squared Bias| 6.3 \times 10^{-2}      | 6.0 \times 10^{-2} | 6.1 \times 10^{-2} | 6.1 \times 10^{-2} | 6.7 \times 10^{-2} |
| MISE       | 7.6 \times 10^{-2}      | 7.5 \times 10^{-2} | 8.2 \times 10^{-2} | 8.8 \times 10^{-2} | 1.1 \times 10^{-1} |
| RMISE      | –                       | 0.99     | 1.08     | 1.16     | 1.39     |
| Avg. runtime (secs) |                  | 356      | 108      | 51       | 24       | 17       |

Table 3. Mean-integrated squared error (MISE), variance, and partial MISE as a function of \( \Delta^* \) and sampling rate when true \( \Delta = 32 \) min.

| \( \Delta \) | Sampling rate (per hour) |          |          |          |          |
|-------------|--------------------------|----------|----------|----------|----------|
| 26          |                          | 4.0      | 2.0      | 1.0      | 0.5      |
| MISE        | 0.390                    | 0.395    | 0.397    | 0.413    |          |
| Variance    | 5.0 \times 10^{-2}      | 5.3 \times 10^{-2} | 5.5 \times 10^{-2} | 6.8 \times 10^{-2} |          |
| P-MISE      | 0.210                    | 0.216    | 0.217    | 0.235    |          |
| MISE        | 0.220                    | 0.225    | 0.224    | 0.238    |          |
| 29          |                          |          |          |          |          |
| Variance    | 3.7 \times 10^{-2}      | 4.4 \times 10^{-2} | 4.3 \times 10^{-2} | 5.5 \times 10^{-2} |          |
| P-MISE      | 0.084                    | 0.090    | 0.088    | 0.103    |          |
| MISE        | 0.061                    | 0.062    | 0.061    | 0.072    |          |
| 32          |                          | 2.7 \times 10^{-2} | 2.9 \times 10^{-2} | 3.0 \times 10^{-2} | 4.2 \times 10^{-2} |
| P-MISE      | 0.061                    | 0.062    | 0.061    | 0.072    |          |
| MISE        | 0.246                    | 0.247    | 0.249    | 0.255    |          |
| 35          |                          | 1.9 \times 10^{-2} | 2.1 \times 10^{-2} | 2.3 \times 10^{-2} | 2.9 \times 10^{-2} |
| P-MISE      | 0.154                    | 0.154    | 0.154    | 0.160    |          |
| MISE        | 0.360                    | 0.359    | 0.353    | 0.361    |          |
| 37          |                          | 2.6 \times 10^{-2} | 3.3 \times 10^{-2} | 3.7 \times 10^{-2} | 4.4 \times 10^{-2} |
| P-MISE      | 0.256                    | 0.255    | 0.252    | 0.259    |          |

Table 4. Rate of selection of \( \hat{\Delta} \) across 500 simulations when true \( \Delta = 32 \) min.

|               | Exponential | Sinusaloid |
|---------------|-------------|------------|
| \( \Delta \)  | 4.0         | 2.0        | 1.0       | 0.5      |
| 26            | 0.38        | 0.35       | 0.30      | 0.26     |
| 29            | 0.22        | 0.20       | 0.17      | 0.12     |
| 32*           | 0.11        | 0.12       | 0.13      | 0.18     |
| 35            | 0.13        | 0.13       | 0.14      | 0.15     |
| 37            | 0.14        | 0.20       | 0.26      | 0.28     |

NOTE: Bold: most likely bandwidth given a chosen subsampling rate.

6. Extensions

In this section, we demonstrate the flexibility of the proposed approach by exploring extensions in several important directions to ensure these methods are robust for practical use with high frequency data. This section will continue to leverage the connection to generalized functional linear models provided by Lemma 4.1.

6.1. Multivariate Extensions

In this section, we extend our model to the case of multiple functional regressors. That is, suppose \( L \) health processes, that is, \( x_i = \{x_{i,t}(t) = (x_{i,1}(t), \ldots, x_{i,L}(t))_{0 \leq t < T_i}, \text{ for each participant is measured at a dense grid of time points.} \) In the suicidal ideation case study, for example, accelerometer is measured at a rate of 32Hz while electrodermal activity (EDA) is measured at a rate of 4Hz. A multivariate extension of our model (1) is given by

\[
h_i(t | H_{x_{i,t}}^{NL}; \theta) = h_0(t; \gamma) \exp \left( \sum_{l=1}^{L} \psi_l(\alpha) + \int_{t-\Delta_l}^{t} x_{i,l}(s) \beta_l(s) ds \right). \tag{8}
\]

The approach given in Section 4.3 extends naturally to the multivariate functional setting. For each functional regressor, we estimate the pooled sample covariance \( \Sigma_{x_{i,l}} \) for \( y \in \{0,1\} \) and \( l = 1, \ldots, L \) as in Section 4.1. Let \( \sum_{l=1}^{L} \sum_{j=0}^{L} \lambda_{k,l}^{(y)} \hat{\lambda}_{k,l}^{(y)}(\cdot) \hat{\psi}_{k,l}(\cdot) \) be the spectral decomposition of \( \Sigma_{x_{i,l}} \). Then \( x_{i,l}(t) \) is approximated using a truncated Karhunen-Loeve \( \int_{t-\Delta_l}^{t} X_l(t,s) \beta_l(s) ds \).

6.2. Missing Data

Sensor data can often be missing for intervals of time due to sensor wearing issues. In the suicidal ideation case study, for example, there are 2139 self-identified moments of distress across all 91 participants. Of these, 1289 event times had complete data for the prior 30 min, 1984 had fraction of missing data on a fine grid less than 30%, and 1998 had fraction of missing data on a fine grid less than 10%.
Missing data is a critical issue because \( c_{i,h}(t) \) cannot be estimated if \( X(s,t) \) is not observed for all \( s \in [0, \Delta] \). Moreover, standard errors should reflect the uncertainty in these coefficients when missing data is prevalent. Goldsmith et al. (2011) suggest using best linear unbiased predictors (BLUP) or posterior modes in the mixed effects model to estimate \( c_{i,h}(t) \); however, this is ineffective when there is substantial variability in these estimates. To deal with this, Crainiceanu and Goldsmith (2010) take a full Bayesian analysis. Yao, Müller, and Wang (2005) introduced PACE as an alternative frequentist method. Petrovich, Reimherr, and Daymont (2019) shows that for sparse, irregular longitudinal, the imputation model should not ignore the outcome variable \( Y_i(t) \).

Here we present an extension of Petrovich, Reimherr, and Daymont (2019) to our setting by leveraging Lemma 4.1 and the marginal covariance estimation procedure to construct a multiple imputation procedure. Let \( x_{i,t} \) denote incomplete sensor data at time \( t \) (i.e., at times \( s_{i,t}^{(b)} \) in \([0, \Delta]\)). Then

\[
\mathbb{E}[X_i(t) | Y_i(t) = y, x_{i,t}(t)] = \mu_y(s,t) + \mathbf{a}_{i,t}^\top \mathbf{b}_{i,t}\]

for

\[
\mathbf{a}_{i,t} = \begin{pmatrix}
\Sigma_i(s_{i,1}, s) \\
\cdot \\
\Sigma_i(s_{i,k}, s)
\end{pmatrix}
\]

and

\[
\mathbf{b}_{i,t} = \begin{pmatrix}
\Sigma_i(s_{i,1}, s_{i,1}) & \Sigma_i(s_{i,1}, s_{i,2}) & \cdots \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\Sigma_i(s_{i,k}, s_{i,1}) & \Sigma_i(s_{i,k}, s_{i,2}) & \cdots
\end{pmatrix},
\]

where we have

\[
\mathbf{a}_{i,t}^\top = \begin{pmatrix}
\Sigma_i(s_{i,1}, s) \\
\cdot \\
\Sigma_i(s_{i,k}, s)
\end{pmatrix};
\]

\[
\mathbf{b}_{i,t}^{-1} = \begin{pmatrix}
\Sigma_i(s_{i,1}, s_{i,1}) & \Sigma_i(s_{i,1}, s_{i,2}) & \cdots \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\Sigma_i(s_{i,k}, s_{i,1}) & \Sigma_i(s_{i,k}, s_{i,2}) & \cdots
\end{pmatrix},
\]

\[
\mathbf{a}_{i,t}(s) = \mathbb{E}[x_{i,t}(s) | Y_i(t) = y] = \{\mu_y(s,t), t\}_{t=1}^T
\]

and \( \mu_y(s,t) \) is the mean of \( X(s,t) \) from group \( y \), and \( \Sigma_i(s, s') \) is the covariance between \( X(s', t) \) and \( X(s, t) \) for \( s', s \in [0, \Delta] \) and \( t \in \mathbb{R}_+ \).

Note that Petrovich, Reimherr, and Daymont (2019) requires modeling the mean and covariance functions separately for \( y = 0 \) and \( y = 1 \).

### 6.2.1. Multiple Imputation Under Uncongeniality

Multiple imputation yields valid frequentist inferences when the imputation and analysis procedure are congenial (Meng 1994); the above procedure is derived for function-on-scalar multiple imputation for binary outcomes, which ignored the joint nature of recurrent event analysis in the presence of high frequency sensor data. The main advantage of the above imputation framework is its simplicity and approximate congeniality when events are rare and the sampling rate is low. The main disadvantage is that the above framework is uncongenial under many events and/or high sampling rates. Meng (1994) defines congeniality between an imputation procedure and an analyst’s complete (and incomplete) analysis procedure if there exists a unifying Bayesian model which embeds the imputer’s imputation model and the analyst’s complete data procedure. A recent discussion of congeniality can be found Bartlett and Hughes (2020). Congeniality ensures good frequentist coverage properties. In some uncongenial settings, standard multiple imputation can be biased downwards leading to under-coverage of confidence intervals. A key question is whether we can use the above imputation methods within a general procedure to handle uncongeniality.

To address this, we use the recommendation from Bartlett and Hughes (2020) and consider a method that first bootstraps a sample from the dataset and then apply multiple imputation to each bootstrap sample. This general approach was originally proposed by Shao and Sitter (1996) and Little and Rubin (2002). We suppose \( B \) bootstraps and \( M \) imputations per bootstrap; let \( \hat{\theta}_{b,m} \) denote the estimator for the \( m \)th imputation of the \( b \)th bootstrap. The point estimator is given by \( (B^{-1})\sum_{b=1}^B \hat{\theta}_b \) where \( \hat{\theta}_b = M^{-1}\sum_{m=1}^M \hat{\theta}_{b,m} \). To construct the confidence interval, we require mean sum of squares with and between bootstraps, that is, \( MSW = \frac{1}{B(M-1)} \sum_{b=1}^B \sum_{m=1}^M (\hat{\theta}_{b,m} - \hat{\theta})(\hat{\theta}_{b,m} - \hat{\theta})^\top \) and \( MSB = \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_b - \hat{\theta})(\hat{\theta}_b - \hat{\theta})^\top \), respectively. Then the estimator of the variance-covariance matrix of \( \hat{\theta} \) is given by \( \hat{\Sigma}_{BM} = \left( \frac{B+1}{BM} \right) MSB - MSW/M \). We obtain the variance for \( \beta(t) \) by \( \phi(t) \hat{\Sigma}_{BM} \phi(t)^\top \). We follow Bartlett and Hughes (2020) and construct confidence intervals based on Satterthwaiate’s degrees of freedom, which here is given by

\[
\hat{\nu} = \left[ \left( \frac{B+1}{BM} \right) MSB - MSW/M \right]^{1/2} \left( B - 1 \right)^{-1} \frac{MSW^2}{BM(M-1)}.
\]

The bootstrap followed by multiple imputation procedure has been studied extensively by Bartlett and Hughes (2020) and is robust to uncongeniality. The main disadvantage of this approach is its considerable computational intensity. Recall likelihood calculations were computationally prohibitive by themselves, so combining with bootstrap and MI would further increase this large scale computation. The random subsampling framework thus simplifies handling of missing data via connections to function-on-scalar multiple imputation by Petrovich, Reimherr, and Daymont (2019) as well as to bootstrap to handle uncongeniality by Bartlett and Hughes (2020). Ignoring the computational time of bootstrap sampling, the computational time for the first choice in the simulation study with \( B = 200 \) bootstraps and \( M = 2 \) imputations per bootstrap leads to 35 hr for a sampling rate of 0.5 compared to 7 hr for a sampling rate of 4, which highlights the benefits of the proposed framework.

### 6.3. Multilevel Models

The approach can be extended to multilevel models with functional regressors, which are critical in mobile health where a high degree of individual variation is often observed. Let \( b_i \sim N(0, \sigma_b^2) \), then the multilevel extension of (1) is

\[
h_i(t | H_i^{NX}, \theta, b_i) \approx \exp \left( Z_i^\top \gamma + g_i(H_i^{NX})^\alpha \right.
\]

\[
+ \left[ M_i^\top + C_i \tilde{I}_{\tilde{\phi}^\top} \right] (\beta + b_i) \right)
\]

\[
= \exp \left( W_i^\top \theta + Z_i^\top b_i \right),
\]

where \( Z_i = M_i + C_i \tilde{I}_{\tilde{\phi}^\top} \). Lemma 4.1 implies that the random subsampling framework applied to (11) leads to a penalized logistic mixed-effects model. As far as the authors are aware, the
combination of mixed-effects and $L_2$-penalization on a subset of parameters is not available in existing software. Given the paper’s focus on “off-the-shelf” software implementations, multilevel models are considered important future work.

7. A Worked Example: Adolescent Psychiatric Inpatient mHealth Study

During an eight month period in 2018, 91 psychiatric inpatients admitted for suicidal risk to Franciscan Children’s Hospital were enrolled in an observational study. Study data were graciously provided by Evan Kleiman and his study team (https://kleimanlab.org). Each study participant wore an Empatica E4 (Garbarino et al. 2014), a medical-grade wearable device that offers real-time physiological data. On each user-day, participants were asked to self-identify moments of suicidal distress. At these times, the participant was asked to press a button on the Empatica E4 device. The timestamp of the button press was recorded. One of the primary study goals was to assess the association between sensor-based physiological measurements and self-identified moments of suicidal distress. In particular, the scientific question is whether there are early indicators of escalating distress by monitoring physiological correlates.

A key concern is whether all moments of suicidal distress are recorded. To ensure this, clinical staff interviewed participants in the evening who were then asked to review their button press activity. Any events that were identified as incorrect button press activity were removed. At the end of the 30-day study period, the average number of button presses per day was 2.42 with a standard deviation of 2.62. Investigation of the button press data shows low button-press counts prior to 7 a.m. and a sharp drop off by 11 p.m. This demonstrates an additional concern: events can only occur when the individual is at-risk, that is, (A) the individual is wearing the Empatica E4 and (B) is currently awake. To deal with (A) and (B), here we define each study day to begin at 9 a.m. and end at 8 p.m.

Figure 1 visualizes button presses versus time since study entry for each user. Day 30 is assumed to censor the observation process. A black mark signals dropout before day 30. Figure 1 shows the potential heterogeneity in button press rates between users and study days. To assess whether there is between user or between study day variation, Table 5 presents a two-way ANOVA decomposition of the button press counts as a function of participant and day in study. The ANOVA decomposition demonstrates high variation with day in study and across users.

Here, we focus on two physiological processes—(1) electrodermal activity (EDA), a measure of skin conductance measured at 4Hz, and (2) activity index (AI) (Bai et al. 2016), a coordinate-free feature built from accelerometer data collected at 32Hz that measures physical movement. Electrodermal activity can be significantly impacted by external factors (e.g., room temperature). To account for the high between user-day variation, we analyze EDA standardized per study-day and device. The individual EDA and AI trajectories are highly variable which tends to obscure patterns and trends. In Figure 2, the mean trajectories of EDA and AI are plotted in reverse time from button press timestamps, which shows sharp changes in EDA and AI in the 30 min prior to button presses. In Figure 2 in Appendix E.3, the mean trajectories of EDA and AI are plotted in reverse time for the sampled nonevent times in the 30 min prior to the nonevent time. The distinct mean behavior motivates our desire to model these two processes separately as discussed in Remark 4.2.

7.1. Complete-Case Analysis

Inspection of Figure 2(a) and (b) suggest setting $\Delta$ between 5 and 30 min. Here, we investigate three potential window lengths, $\Delta = 5, 15,$ and $30$. To ensure minimal loss of efficiency, the subsampling rate was set to once every 15 min. Given the daily button press rate, this ensures an average of 44 nonevents to 2.5 events per day. Based on Table 1, this ensures we can achieve a substantial data reduction at a minimal loss of efficiency. After sampling nonevent times, complete-case analyses are performed, that is, sampled times where sensors included in the model have any level of missing data are ignored. Table 6 presents both AIC and BIC criteria on the complete-case data, normalized by the number of observations to ensure fair comparison. We find that $\Delta = 15$ is adequate to capturing the proximal impact of EDA and AI on the risk of a button press.

| Delta | AIC | BIC |
|-------|-----|-----|
| 5     | 0.96| 0.99|
| 15    | 0.98| 1.01|
| 30    | 0.95| 0.99|

Table 6. Normalized AIC and BIC for different choices of $\Delta$ using complete-case data.

NOTE: For $\Delta = 5$, $K_0 = K_5 = 31$ due to amount of data in shorter window-length, while $K_0 = K_5 = 35$ for $\Delta = 15$ and 30.
In Section E.2 of the supplementary materials, Figures 1(a) and (b) present the per-patient average fraction with missing EDA and AI in 30-min windows, respectively. For EDA, we have a wide range of missingness—from 0% to over 40% across individuals with most between 5% and 30% average fraction missing. For AI, the missingness is less pronounced—from 0% to 15% across individuals with most between 0% and 10% average fraction missing.

We analyzed activity index (AI) and electrodermal activity jointly in a multivariate model as in (8). To account for times when the participants may not be wearing the device and thus not be at-risk for a button press, we limited the analysis to data collected between 9 a.m. and 8 p.m. In the analysis, we assumed a constant baseline hazard and included a binary indicator of whether the participant had an event in the past 12 hr to account for patient heterogeneity in the number of events as seen in Figure 1. Figure 3(a) and (b) presents the estimates from the joint analysis and their associated 95% confidence intervals using the bootstrap and multiple imputation strategy. We highlight in gray the statistically significant regions. Here, we see that standardized EDA is not associated with increased risk of button press, while activity index sees a positive association in the final few minutes prior to a button press.

### 7.2. Sensitivity Analysis for Window Length

We next investigate whether the results are sensitive to window length. Specifically, we re-analyzed activity index (AI) and electrodermal activity jointly in a multivariate model as in equation (8) with $\Delta = 5$ and 30 min. Figure 4(a) and (b) presents...
the estimates from the joint analysis for the activity index (AI) and their associated 95% confidence intervals using the bootstrap and multiple imputation strategy. We highlight in gray the statistically significant regions which remain similar across all three choices of window length. We do not present results for the standardized EDA as it continues to be not associated with increased risk of button press.

8. Discussion

In this article, we have presented a methodology for translating a difficult functional analysis with recurrent events problem into a traditional logistic regression. The translation leveraged subsampling and weighting techniques, specifically the use of weights suggested by Waagepetersen (2008), along with flexible functional data analysis methods of Goldsmith et al. (2011) with marginal covariance methods for longitudinal functional data of Park and Staicu (2015). The proposed methodology abides by the idea that we should make data as small as possible as fast as possible. Subsampling and weighting converts the problem to well-known territory which allows us to leverage existing software. We show limited loss of efficiency when the subsampling is properly tuned to the event rates. Important extensions to an online sampling algorithm, optimal weighting when the Poisson point process assumption does not hold, and nonlinear functional data methods are all considered important future work.

Supplementary Materials

The supplementary materials contains a notation glossary, technical arguments, additional simulations, and case study details.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

References

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models based on Counting Processes*. New York: Springer. [529]

Bai, J., Di, C., Xiao, L., Evenson, K. R., LaCroix, A. Z., Crainiceanu, C. M., and Buchner, D. M. (2016), “An Activity Index for Raw Accelerometry Data and its Comparison with Other Activity Metrics,” *PloS One*, 11, 1–14. DOI:10.1371/journal.pone.0160644. [534]

Balandandayuthapani, V., Mallick, B. K., Young Hong, M., Lupton, J. R., Turner, N. D., and Carroll, R. J. (2008), “Bayesian Hierarchical Spatially Correlated Functional Data Analysis with Application to Colon Carcinogenesis,” *Biometrics*, 64, 64–73. [527]

Bartlett, J. W., and Hughes, R. A. (2020), “Bootstrap Inference for Multiple Imputation Under Uncongeniality and Misspecification,” *Statistical Methods in Medical Research*, 29, 3533–3546. [533]

Caffo, B. S., Jank, W., and Jones, G. L. (2005), “Ascent-based Monte Carlo Expectation–Maximization,” *Journal of the Royal Statistical Society*, Series B, 67, 235–251. DOI:10.1111/j.1467-9868.2005.00499.x. [526]

Cassel, C., Särndal, C.-E., and Wretman, J. H. (1977), *Foundations of Inference in Survey Sampling*, New York: Wiley. [527]

Chen, K., and Müller, H.-G. (2012), “Modeling Repeated Functional Observations,” *Journal of the American Statistical Association*, 107, 1599–1609. [527,528,529]

Crainiceanu, C., and Goldsmith, J. (2010), “Bayesian Functional Data Analysis Using Winbugs,” *Journal of Statistical Software*, 32, 1–33. [530,533]

Crainiceanu, C. M., Staiku, A.-M., and Di, C.-Z. (2009), “Generalized Multilevel Functional Regression,” *Journal of the American Statistical Association*, 104, 1550–1561. DOI:10.1198/jasa.2009.tm08564. [525]

Cui, E., Crainiceanu, C. M., and Leroux, A. (2021), “Additive Functional Cox Model,” *Journal of Computational and Graphical Statistics*, 30, 780–793. DOI:10.1080/10618600.2020.1853550. [525]

Di, C. Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009), “Multilevel Functional Principal Component Analysis,” *Annals of Applied Statistics*, 3, 458–488. [527]

Dong, J., Cao, J., Gill, J., Miles, C., and Plumb, T. (2021). “Functional Joint Models for Chronic Kidney Disease in Kidney Transplant Recipients,” *Statistical Methods in Medical Research*, 30, 1932–1943. DOI:10.1177/09622802211009265. [526]

Free, C., Phillips, G., Galli, L., Watson, L., Felix, L., Edwards, P., Patel, V., and Haines, A. (2013), “The Effectiveness of Mobile-Health Technology-
