Statistics
in Medicine

Research Article

Received 28 June 2016, Accepted 3 June 2017 Published online 17 July 2017 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7385

Extension of the association structure in joint models to include weighted cumulative effects

Katya Mauff,[a*†] Ewout W. Steyerberg,[b] Giel Nijpels,[c,d] Amber A.W.A van der Heijden[e] and Dimitris Rizopoulos[a]

Motivated by a study measuring diabetes-related risk factors and complications, we postulate an extension to the standard formulation of joint models for longitudinal and survival outcomes, wherein the longitudinal outcome has a cumulative effect on the hazard of the event, weighted by recency. We focus on the relationship between the biomarker HbA1c and the development of sight threatening retinopathy, since the impact of the HbA1c marker on the risk of sight threatening retinopathy is expected to be cumulative, with the evolution of the HbA1c marker over time contributing to progressively greater damage to the vascular structure of the retina. Opting for a parametric approach, we propose the use of the normal and skewed normal probability density functions as weight functions, estimating the relevant parameters directly from the data. The use of the recency-weighted cumulative effect specification allows us to incorporate differences in the development of the longitudinal profile over time in the calculation of hazard ratios between subjects. The proposed functions provide us with parameters with clinically relevant interpretations while retaining a degree of flexibility. In addition, they also allow answering of important clinical questions regarding the relative importance of various segments of the biomarkers history in the estimation of the risk of the event.

Keywords: joint models; longitudinal outcome; survival outcome; association structures

1. Introduction

Very often in medical studies, data on multiple outcomes of interest are simultaneously recorded. Typically, these data take the form of a longitudinal or repeated measurement outcome and time-to-event records. Individual analyses of these outcomes is possible for simpler hypotheses regarding differing average longitudinal profiles between groups of interest, or the impact of treatment on the risk of an event. However, in instances where the association between the longitudinal and time-to-event outcomes is of specific interest, or where we would like to assess the impact of an endogenous time-varying covariate measured with error on the risk of the event occurrence, joint modelling approaches are preferred.

In the standard formulation of joint models, the current underlying value of the subject-specific marker is assumed to be associated with the risk of an event occurring at the same time t, through an association parameter α. However, this particular functional form may not be adequate in describing the association structure between the outcomes in all settings. To this end, alternative association structures have been proposed [1–3], two of which are as follows: first, to allow the risk of an event to depend on the slope of the longitudinal profile and, second, to postulate that the risk of an event at time t is dependent on the integrated longitudinal profile (a cumulative effect).

Motivated by a study on diabetes research, we focus in this paper on the cumulative effect association structure. More specifically, the data considered are that from the Hoorn Diabetes Care System (DCS)

[a]*Department of Biostatistics, Erasmus MC, PO Box 2040, 3000 CA Rotterdam, The Netherlands*
[b]*Department of Public Health, Erasmus MC, Rotterdam, The Netherlands*
[c]*Department of General Practice, VU Medical Centre, Amsterdam, The Netherlands*
[d]*EMGO Institute for Health and Care Research, VU Medical Centre, Amsterdam, The Netherlands*
[e]*Department of General Practice and Elderly Care Medicine, VU Medical Center, Amsterdam, The Netherlands*
[*]*Correspondence to: Katya Mauff, Department of Biostatistics, Erasmus MC, PO Box 2040, 3000 CA Rotterdam, The Netherlands.*
[†]*E-mail: k.mauff@erasmusmc.nl*

cohort. The DCS cohort represents a complete dataset on the natural course of type 2 diabetes with annual highly protocolised follow-up measurements on diabetes-related risk factors and complications. Of specific interest is the diagnosis of retinopathy, in particular, sight threatening retinopathy (STR), and the impact of HbA1c (glycated haemoglobin) on the development of STR. Retinopathy is characterized by ongoing inflammation and vascular remodelling of the retina over time, with a higher glucose burden contributing to the risk of further microvascular damage, resulting in leakage and bleeding, devascularization, and the formation of new vessel development in other areas of the retina, causing visual loss and eventually the risk of blindness. HbA1c is a form of glycosylated haemoglobin expressing the 3-month average plasma glucose concentration [4]. The test is limited to a 3-month average because the lifespan of a red blood cell is 3 months. It is formed in a non-enzymatic glycation pathway by haemoglobin's exposure to plasma glucose. Higher values of HbA1c indicate poorer control of blood glucose levels, and while specific values of 7.5% or 58 mmol/mol have been identified as important thresholds in the risk of developing retinopathy, the impact of HbA1c on the risk of retinopathy and STR is expected to be cumulative, with the evolution of the HbA1c level over time contributing to progressively greater damage to the vascular structure of the retina.

In the context of the standard Cox model, several authors [5–7] have discussed the use of recency weighted cumulative exposure models, where the cumulative effect of the exposure history is modelled as a weighted sum of all past values, with weights representing the relative importance of the values as a function of the time elapsed since exposure. Abrahamowicz *et al*. [8] considered a parametric time-dependent weight function for inclusion in the extended Cox model, whereby the weighted cumulative exposure is calculated at each time point u rather than only once at the end of the study, but specified the values of the parameters a priori. Further research [9] suggested the use of regression spline-based methods in place of the parametric weight function, as a means of estimating the function directly from the data.

In the analysis of the DCS cohort, we are interested in determining the impact of the entire history of the endogenous HbA1c level on the risk of developing STR, and propose the use of a time-dependent recency-weighted cumulative exposure association structure for the inclusion of the marker in the relative risk model, now within the context of joint models. Our work may be seen as an extension of the proposal by Abrahamowicz *et al*. [8], in which we utilize joint models to account for the endogenous nature of the longitudinal outcome, and we estimate the relevant parameters of two proposed parametric weight functions directly from the data. The motivation behind the choice of the parametric approach is prior knowledge of the often exhibited functional form seen in many longitudinal biomarkers, similar to an exponential decay curve, with more recent biomarker levels being more relevant than early follow-up values. In addition, such weight functions also provide a clinically relevant interpretation of parameters, while retaining flexibility.

The rest of the paper is organized as follows. In Section 2, we define the standard joint model and the model with a weighted cumulative association structure. We then introduce the two potential weight functions in Section 2.2 and discuss the estimation procedure in Section 3. Finally, in Section 4, we present the results of the analysis of the DCS data, together with a further illustration of the proposed methodology using the well-known primary biliary cirrhosis (PBC) data introduced by Murtaugh *et al*. [10], as well as the results of a small scale simulation study.

2. Model formulation

2.1. The joint model with cumulative association structure

Let $D_n = \{Ti, \delta_i, \boldsymbol{y}_i; i = 1, \dots, n\}$ denote a sample from the target population, where T_i^* denotes the true event time for the i-th subject , C_i the censoring time, $T_i = \min(T_i^*, C_i)$ is the corresponding observed event time, and $\delta_i = I(T_i^* \leq C_i)$ is the event indicator, with $I(.)$ being the indicator function that takes the value 1 when $T_i^* \leq C_i$, and 0 otherwise.

Allowing $\boldsymbol{y}_i$ to denote the $n_i \times 1$ longitudinal response vector for the i-th subject, with element y_{ij} denoting the value of the longitudinal outcome taken at time point $t_{ij}, j = 1, \dots, n_i$, we then propose a generalized linear mixed effects model for the longitudinal outcome as follows. We assume that the conditional distribution of $\boldsymbol{y}_i$, given a vector of random effects $\boldsymbol{b}_i$, is a member of the exponential family,

with linear predictor given by

$$g[E\{y_i(t) \mid \boldsymbol{b}_i\}] = \eta_i(t, \boldsymbol{b}_i) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i, \tag{1}$$

where $g(.)$ is a known one-to-one monotonic link function, $y_i(t)$ is the value of the longitudinal outcome for the i-th subject at time point t, and $\boldsymbol{x}_i(t)$ and $\boldsymbol{z}_i(t)$ are the time-dependent design vectors for the fixed-effects $\boldsymbol{\beta}$ and for the random effects $\boldsymbol{b}_i$, respectively. The random effects are assumed to have a multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{D}$.

For the time-to-event outcome, we make the assumption that the risk for an event depends on a function of the subject-specific linear predictor $\eta_i(t, \boldsymbol{b}_i)$. In the simplest version of the joint model, we assume that the hazard for an event at any time t is associated with the current underlying value of the biomarker at the same time point, denoted as $\eta_i(t, \boldsymbol{b}_i)$ and defined in (1) and that the strength of this association is measured by parameter α

$$h_i(t \mid \eta_i(t, \boldsymbol{b}_i), \boldsymbol{w}_i(t)) = \lim_{\Delta t \to 0} \frac{\Pr\{t \leq T_i^* < t + \Delta t \mid T_i^* \geq t,\ \eta_i(t, \boldsymbol{b}_i),\ \boldsymbol{w}_i(t)\}}{\Delta t} \tag{2}$$

$$= h_0(t)\ \exp\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i(t) + \alpha \eta_i(t, \boldsymbol{b}_i)\}, \tag{3}$$

where $h_0(.)$ denotes the baseline hazard function and $\boldsymbol{w}_i$ is a vector of baseline (or exogenous time-varying) covariates with corresponding regression coefficients $\boldsymbol{\gamma}$.

As previously mentioned, this simple formulation may not always be most appropriate in the case of more complex relationships between the longitudinal and time-to-event processes [9]. Here we focus on a specific alternative formulation of the model, which accounts for a cumulative effect of the longitudinal outcome by including the integral of the longitudinal trajectory from baseline up to time t in the linear predictor of the relative risk submodel [11, 12]. This specification may also increase the statistical power of the analyses [8].

Specifically,

$$h_i(t) = h_0(t)\ \exp\left\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i(t) + \alpha \int_0^t \eta_i(s, \boldsymbol{b}_i)\, ds\right\}, \tag{4}$$

where $\int_0^t \eta_i(s, \boldsymbol{b}_i)\, ds$ is defined as the signed area of the region in the xy-plane that is bounded by the graph of $\eta_i(s, \boldsymbol{b}_i)$, the x-axis, and the vertical lines $s = 0$ and $s = t$. The area above the x-axis adds to the total and that below the x-axis subtracts from the total. For any particular time point t, α measures the strength of the association between the risk for an event at time point t and the signed area (hereafter referred to as the area) under the longitudinal trajectory up to the same time t, with the area under the longitudinal trajectory taken as a summary of the whole marker history $H_i(t) = \{\eta_i(s, \boldsymbol{b}_i),\ 0 \leq s < t\}$.

The interpretation of α differs depending on the model specification, whereby previously, under the standard joint model formulation, α indicated the resultant increase in the log hazard ratio of the event for a 1-unit increase in the value of the longitudinal profile at the same time t; it now determines the change in the log hazard ratio for a 1-unit increase in the area under the longitudinal trajectory. This is clinically advantageous, as it allows calculating hazard ratios between patients utilizing their whole longitudinal profile rather than only their current value as in formulation (3). However, a potential limitation of this formulation is that we will not be able to discriminate (with respect to risk) between two patients who may have different longitudinal profiles but the same area under these profiles.

Figures 1 and 2 illustrate the use of the proposed cumulative effect parameterization on the estimation of the hazard at time t and the interpretation of the association parameter α. Figure 1 demonstrates the observed values and estimated longitudinal profiles for two randomly selected subjects from the DCS cohort. The x-axis denotes the follow-up time in years and is truncated at the maximum of the event times for the two subjects. The hazard ratio between the two subjects for any time $t \leq min(T_1, T_2)$ for $i = 1, 2$ (where T_i is the observed event time for subject i as previously defined) may be calculated as in (3), using the corresponding value of the biomarker from the estimated subject-specific longitudinal profiles (as marked by the X's at time $t = 6$ for example). This is in contrast to the Last Value Carried Forward (LVCF) approach seen in the extended Cox model and indicated on the graph by the dotted line of the step-function.

Extending this parameterization to the cumulative effect, the hazard at any time point t is associated with the entire area under the longitudinal profile, as demonstrated in Figure 2(a).

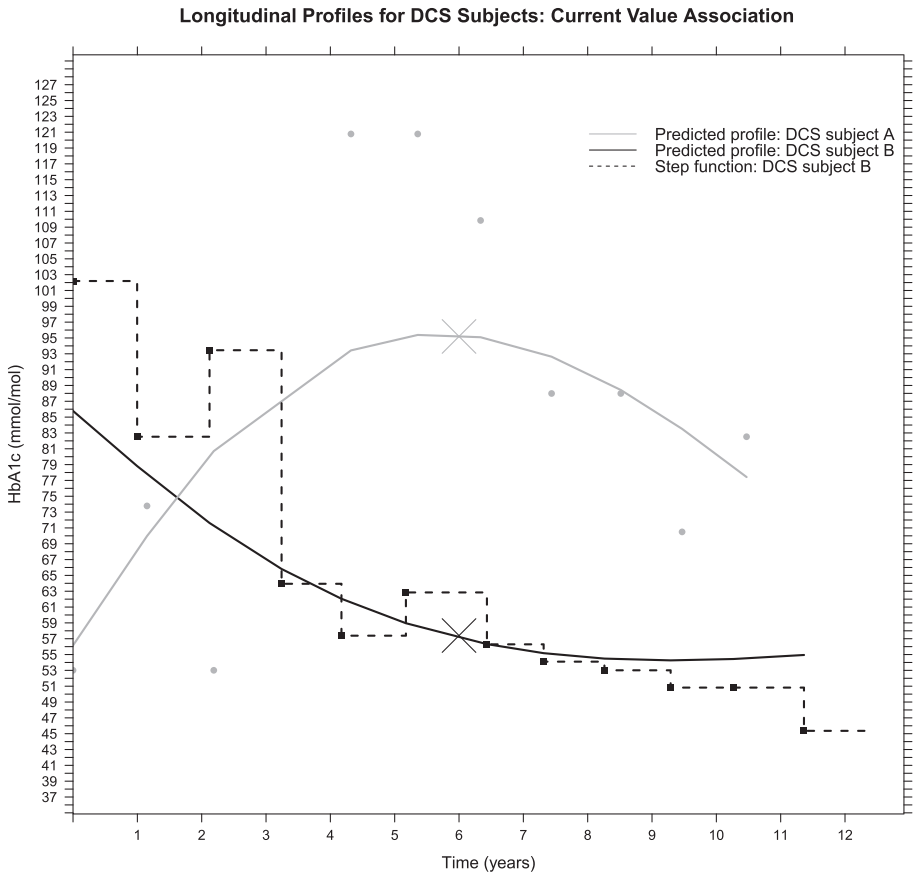

Figure 1. Longitudinal profiles for randomly selected subjects from the DCS cohort.

2.2. *Introduction of weights*

The cumulative effect as defined in (4) assumes that all past values of the biomarker from 0 up to time t are of equivalent importance. This may be realistic for studies with shorter follow-up periods, but for many biomarkers, it may be unreasonable to assume that values at baseline are as important as values after a much longer period of follow-up. To allow for potentially complex behaviour of the biomarker over time and, by extension, the association between the values of the biomarker at various time points and the estimated risk, we require a more flexible approach [6, 13].

For the majority of biomarkers measured in humans, we expect that more recent values may be more relevant in the estimation of the hazard at time t than values further away from this time point, requiring a weight function that is a decreasing function of time. This expectation has been previously noted in the literature, with Vacek [7], suggesting the use of the exponential and normal cumulative distribution functions to assign heavier weights to earlier values, and the use of the normal density function in assigning maximum weight to values at time s, $\forall s \leq t$. The normal density function is used again by Abrahamowicz *et al*. [8], where they define:

$$\varpi(t-s)_+ = \exp\left\{-\frac{(t-s)_+^2}{2\sigma^2}\right\},$$

where $\varpi(.)$ is an appropriately chosen weight function that places different weights at different time points, with $(t-s)_+ = t-s$ when $t > s$ and zero otherwise, and $t-s$ denotes time elapsed since exposure. In both papers, however, the parameters of the density functions were specified a priori; in the case of Abrahamowicz *et al*. [8], two alternative weight functions were considered, in which the scale parameter (given by σ), which controls the rate of change in the weights over time, was selected such that the weight was equal to 0.5 for two specific, clinically relevant values of $t-s$. In this paper, we follow a similar approach but utilize both the normal and skewed normal distributions as weight functions (to allow for greater flexibility), estimating the necessary parameters directly from the data. More specifically, to allow

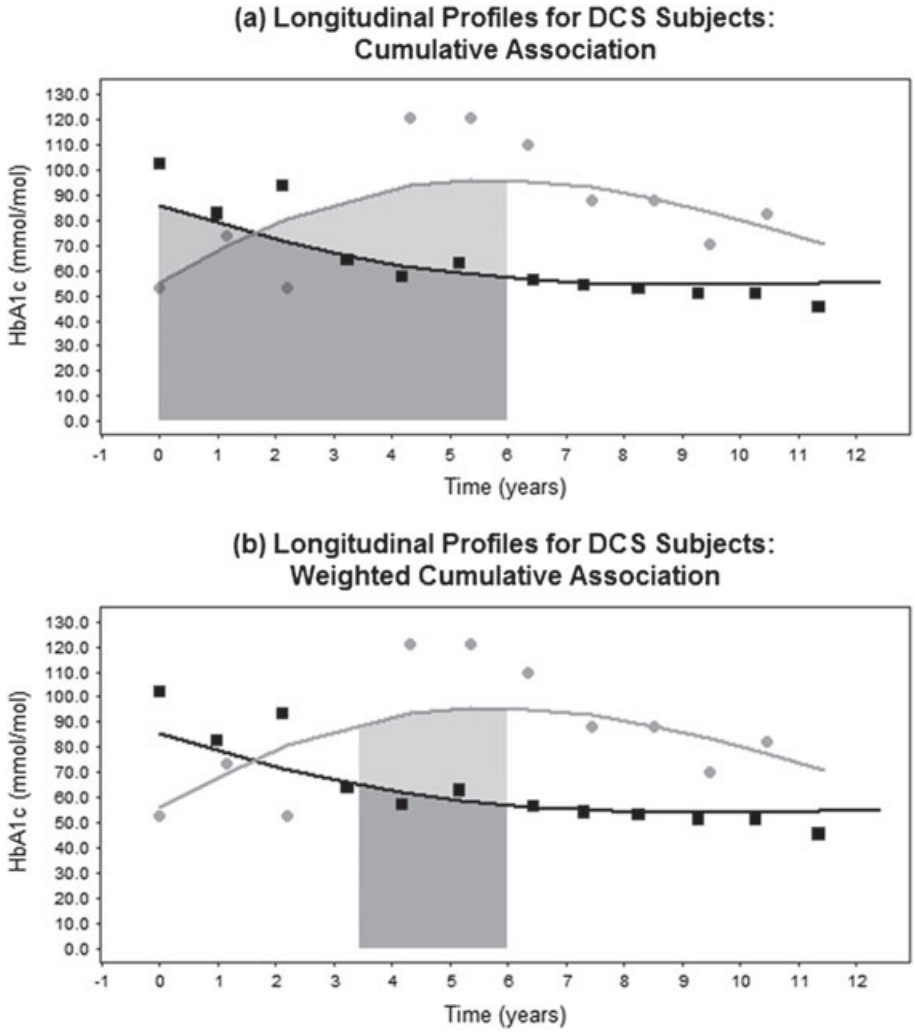

Figure 2. Longitudinal profiles for randomly selected subjects from the DCS cohort with (a) cumulative effect and (b) weighted cumulative effect as estimated using the normal density function.

for differential weights in the cumulative effect formulation, we specify

$$h_i(t) = h_0(t)\,\exp\left\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i(t) + \alpha \int_0^t \varpi(t-s)_+ \eta_i(s, \boldsymbol{b}_i)\,ds\right\}, \tag{5}$$

where $(t-s)_+$ is defined as above and s denotes a time prior to or equal to t. The weight functions that we consider are as follows:

$$\varpi(t-s)_+ = \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp -\{(t-s)_+\}^2/2\sigma^2}{\int_0^{t.max} \frac{1}{\sigma\sqrt{2\pi}} exp-\{x\}^2/2\sigma^2 dx} \tag{6}$$

and

$$\varpi(t-s)_+ = \frac{\frac{1}{v\pi} \exp\{-(t-s)_+/2v^2\} \int_{-\infty}^{\kappa(t-s)_+/v} \exp(-\lambda^2/2) d\lambda}{\int_0^{t.max} \frac{1}{v\pi} \exp\{-x/2v^2\} \int_{-\infty}^{\kappa x/v} \exp\{-\lambda^2/2\} d\lambda dx}, \tag{7}$$

with mean 0 and parameters scale v and shape κ estimated. Figures S1 and S2 show various possible weight functions resulting from different parameter values for the standardized normal (6) and skewed

normal density functions (7), respectively. The solid line in Figure S1 represents the normal density function with $\sigma = 1$, for $t \geq 0$. This has a much steeper decline than the line for which $\sigma = 5$ and a demonstrably shorter period of time wherein the weights are above zero. Figure S2 demonstrates the additional flexibility in the shape of the weight function achieved through the use of the skewed normal density function. Comparison of the three darker lines in Figure S2, representing the κ shape parameter values of 0, 1 and 4, with a scale parameter value of 1, with the solid line in Figure S1, demonstrates specifically a flexibility in the presumed peak or mode of the function, whereby the peak is shifted to the right. Different values of κ thus allow for more flexibility in the structure of the function, altering the linearity and therefore the local rate of decline.

In the case of the normal density weight function, estimation of σ may indicate how much of the history of the longitudinal biomarker may be important in predicting the occurrence of future events, through use of the $68 - 95 - 99.7$ rule. This intuitive interpretation of the scale parameter is no longer applicable when using the skewed normal density function, although the period of interest can still be determined for this more flexible option directly from the estimated weight function itself.

To further facilitate the answering of questions regarding the relative importance of various segments of the biomarkers history, we can calculate the area under the density curve for time intervals of specific interest. Because the CDF of the weight function is equal to 1 by design (through the use of the normalized parameterization of the weight functions), we can define the relative area under the curve (AUC), denoted by *rAUC*, as

$$rAUC_{\{t,(t-s)_+\}} = \mathcal{F}(t-s)_+ - \mathcal{F}(t). \tag{8}$$

Calculating the relative AUC for the two weight functions in Figure S1 corresponding to $\sigma = 1$ and $\sigma = 5$, respectively, the interval from 0 to 0.5 would account for approximately 37.3% (cumulatively) of the relative weight for the standard normal density curve ($\sigma = 1$), where the same interval would account for only 17.1% for the density curve with $\sigma = 5$, and the weights for the latter function would be more equally distributed across the entire time period, a pattern that would indicate a much longer period of clinical relevance.

Revisiting Figure 2(b), assuming a weighted cumulative effect parameterization as in (5), we allow the hazard at time t to depend only on the area under the longitudinal profile for a specific, clinically relevant period of time, obtained via direct estimation of the relevant parameters of the weight function. The association parameter α now determines the change in the log hazard ratio for a 1-unit increase in the area under the longitudinal trajectory for the specific time period for which the weights are non-zero.

3. Estimation

The model is estimated using the Bayesian approach, using Markov Chain Monte Carlo algorithms. We assume conditional independence in the derivation of the posterior distribution of the model parameters; that is, given the random effects that underlie both the longitudinal and survival processes, these processes are independent of one another. The longitudinal responses of each subject are also assumed independent conditional on the random effects for that subject. Formally, we have

$$\begin{aligned} p(\boldsymbol{y}_i, T_i, \delta_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) &= p(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(Ti, \delta_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}), \\ p(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) &= \prod_j p(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\theta}), \end{aligned}$$

where $\boldsymbol{\theta}$ denotes the full parameter vector and $p(.)$ denotes an appropriate probability density function. Under these assumptions, the posterior distribution is analogous to

$$p(\boldsymbol{\theta}, \boldsymbol{b}) \propto \prod_{i=1}^{n} \prod_{j=1}^{n_i} p(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(Ti, \delta_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}),$$

where

$$p(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = \exp\left\{ [y_{ij} \psi_{ij}(\boldsymbol{b}_i) - c\{\psi_{ij}(\boldsymbol{b}_i)\}] / a(\varphi) - d(y_{ij}, \varphi) \right\},$$

with $\psi_{ij}(\boldsymbol{b}_i)$ and φ denoting the natural and dispersion parameters in the exponential family, respectively, and $c(.)$, $a(.)$ and $d(.)$ are known functions specifying the member of the exponential family. For the survival part,

$$p(T_i, \delta_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = h_i(T_i \mid H_i(T_i))^{\delta_i} \exp\left\{ -\int_0^{T_i} h_i(s \mid H_i(s))\, ds \right\},$$

with $h_i(.)$ as defined in (5). The survival function for the cumulative effects parameterization, given by

$$S_i(t \mid \eta_i(t, \boldsymbol{b}_i), \boldsymbol{w}_i(t)) = \exp\left[-\int_0^t h_0(s) \exp\left\{ \boldsymbol{\gamma}^\top \boldsymbol{w}_i(s) + \alpha \int_0^s \varpi(s-u)_+ \eta_i(u)\, du \right\} ds \right], \tag{9}$$

does not have a closed-form solution, and thus, a numerical method must be employed for its evaluation. Standard options are the Gauss–Kronrod and Gauss-Legendre quadrature rules, adaptive Gaussian quadrature rules, which utilize specific Kronrod-points of evaluation. Suitable selection of these points allows abscissas from previous iterations to be reused as part of a new set of points, as opposed to the recomputation of all abscissas at each iteration. Specifically, for the survival function specified in equation (9), we have

$$S(t) = \exp\left[-\int_0^{T_i} h_i(s) ds \right] \tag{10}$$

$$\approx \exp\left[-\frac{T_i}{2} \sum_{m=1}^{15} \pi_m h_0(\tau_m) \exp\left\{ \boldsymbol{\gamma}^\top \boldsymbol{w}_i(\tau_m) + \frac{\alpha \tau_m}{2} \sum_{n=1}^{15} \tilde{\pi}_n \varpi\left(\frac{\tau_m (1 - q_n)}{2} \right) \eta_i \varphi_{mn} \right\} \right] \tag{11}$$

where $\tau_m = \frac{T_i(q_m+1)}{2}$, $\varphi_{mn} = \frac{\tau_m(q_n+1)}{2}$, T_i is the observed failure time, π_m and $\tilde{\pi}_n$ denote prespecified weights and q_m and q_n prespecified abscissas. Derivation of Equation (11) may be found in Section 1 of the Supporting Information as per [14].

For the baseline hazard function $h_0(.)$, we specify a penalized version of the B-spline approximation, with

$$\log h_0(t) = \boldsymbol{\gamma}_{h_0,q} B_q(t, \boldsymbol{v}),$$

where $B_q(t, \boldsymbol{v})$ denotes the q-th basis function of a B-spline with knots $v_1, \ldots, v_Q$ and $\boldsymbol{\gamma}_{h_0}$ the vector of spline coefficients, for which we specify the improper prior [15],

$$p(\boldsymbol{\gamma}_{h0} \mid \tau_h) \propto \tau_h^{\rho(K)/2} \exp\left(-\frac{\tau_h}{2} \boldsymbol{\gamma}_{h_0}^\top \boldsymbol{K} \boldsymbol{\gamma}_{h_0} \right),$$

where τ_h is the smoothing parameter that takes a Gamma$(1, 0.005)$ hyper-prior in order to ensure a proper posterior for $\boldsymbol{\gamma}_{h_0}$, $\boldsymbol{K} = \Delta_r^\top \Delta_r$, Δ_r denotes the r-th difference penalty matrix, and $\rho(\boldsymbol{K})$ denotes the rank of $\boldsymbol{K}$.

For the remaining parameters $\boldsymbol{\theta}$, we take standard prior distributions. Specifically, for the vector of fixed effects of the longitudinal submodel $\boldsymbol{\beta}$, for the regression parameters of the survival model $\boldsymbol{\gamma}$, and for the association parameter α, we use independent univariate diffuse normal priors. For the covariance matrix of the random effects, we assume an inverse Wishart prior, and when fitting a joint model with a normally distributed longitudinal outcome, we take an inverse-Gamma prior for the variance of the error terms σ^2. For the parameters of the weight function, we use non-informative Uniform priors. More details regarding Bayesian estimation of joint models may be found in [16] and [2].

4. Analysis of motivating data sets

4.1. DCS data

We continue with the analysis of the DCS data introduced in Section 1. New patients entered the study cohort each year and were then followed up over time. Clinical information was obtained by protocolised annual follow-up measurements, at which time the presence of microvascular and macrovascular complications was also identified. The year of entry was considered the baseline for each new patient, and retina photograph scans for retinopathy were obtained within at least 2 years of entry into the database.

Sight threatening retinopathy is defined as a retinopathy diagnosis of grades 3–5 on the fundal photograph using the EURODIAB coding system, and its development is dependent mainly on several risk factors, such as the age of the patient, the diabetes duration, HbA1c, blood pressure and early stages of retinopathy. The total glycaemic load plays an important role in the development of STR; a longer duration of diabetes in combination with higher HbA1c values results in a higher glycaemic burden and thus in a higher risk of retinopathy.

Of the 8213 patients in the study period from 1998 to 2010, for whom both longitudinal data on clinical information and retinopathy were available, 6294 patients were considered for inclusion in the analysis presented here. Patients in whom STR was already present at baseline (within 2 years of entry) were excluded (70 (0.9%)), together with those that did not have retina photograph results within 2 years of their baseline year of entry (282 (3.4%)) and those that had no retina photograph results (921 (11.2%)) or no recorded HbA1c values (53 (0.7%)); 593 (7.2%) patients were also excluded due to missing values in the covariates. Baseline characteristics for the 6294 patients are described in Table S1. During a mean of 54.5 months of follow-up, 125 (2%) patients experienced the event of interest (first occurrence of STR); 7.6% of patients presented with grade 1 (401 (6.4%)) or 2 (77 (1.2%)) retinopathy at baseline (within 2 years of entry). The average age and duration of diabetes at baseline (calculated as the time from diagnosis to entry) were 60.5 years (SD = 11.8) and 2.8 years (SD = 4.8), respectively, and the mean baseline systolic and diastolic blood pressure were 142.9 mmHg (SD = 20.7) and 81.1 mmHg (SD = 10.9), respectively.

For ease and speed of computation, a sample of 1125 subjects was selected, including all 125 subjects who experienced the event of interest and a random sample of 1000 without the event. Their baseline characteristics are also provided in Table S1 for comparison with those of the larger sample. The subject-specific longitudinal profiles for HbA1c (mmol/mol) are illustrated for a smaller random sample of patients in Figure S3. Given the apparent non-linearity of these profiles, we include natural cubic splines in both the fixed and random effects parts of the longitudinal model for HbA1c, also controlling for the baseline systolic blood pressure, duration of diabetes, age at entry and presence or absence of lower grade retinopathy at baseline;

$$\begin{aligned} y_i(t) = {} & (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})B_n(t, \lambda_1) + (\beta_2 + b_{i2})B_n(t, \lambda_2) \\ & + \beta_3 \texttt{Baseline Age}_i + \beta_4 \texttt{Duration of Diabetes}_i \\ & + \beta_5 \texttt{Baseline SBP}_i + \beta_6 \texttt{Baseline Retinopathy}_i + \epsilon_i(t), \end{aligned}$$

where $\{B_n(t, \lambda_k) : k = 1, 2\}$ denotes the B-spline basis matrix for a natural cubic spline of time with one internal knot placed at the 50th percentile for the follow-up times, $\epsilon_i(t) \sim N(0, \boldsymbol{R}_i)$ and $\mathbf{b}_i \sim N(\mathbf{0}, \boldsymbol{D})$, with $\boldsymbol{R}_i = \sigma^2_\epsilon \mathbf{I}_{n_i}$ and $\boldsymbol{D}$ an unstructured variance-covariance matrix.

In the Cox model, the weighted integral of the subject-specific linear predictor of the mixed model $\eta_i(t, \boldsymbol{b}_i)$ is included in the linear predictor of the relative risk model. Two alternative weight functions were specified: the probability density function of the standardized normal distribution with mean zero, which has as parameter the standard deviation, and the probability density function of the standardized skewed normal distribution with mean zero and both the scale (v) and shape (κ) as parameters. We also fit a model with an unweighted cumulative effect.

$$h_i(t) = h_0(t) \exp\left\{ \alpha \int_0^t \varpi(t - s)_+ \eta_i(s, \boldsymbol{b}_i)\, ds \right\}$$

with $\varpi(t - s)_+$ specified as in (6) and (7).

All analyses presented in this paper are carried out using the **JMbayes** package in R [17, 18], and an example of the code used to fit the various models may be found in Section 5 in the Supporting Information.

Candidate models for the best fitting model are detailed in Table S2, together with their deviance information criterion (DIC) values. Results from the model with the current value specification and those from the chosen weighted cumulative model are similar, although the weighted cumulative model outperforms the current value specification (DIC values of 71,124 and 71,160 respectively).

The output in Tables I and II hereafter indicates that a 1-unit increase in the value of HbA1c results in a 1.06-fold increase in the risk of developing STR (2.5–97.5% CI: 1.05–1.08), and that a 1-unit increase in the *area* under the HbA1c profile for the relevant period of interest corresponds to a 1.07-fold increase (2.5–97.5% CI: 1.06–1.08) for the skewed normal density function. The model with the unweighted

cumulative effect specification indicates a 1.01-fold increase in the risk of developing STR (2.5–97.5% CI: 1.01–1.02) for every unit increase in the area under the HbA1c profile.

Table I. Parameter estimates and 95% credibility intervals under the joint modelling analysis for sight threatening retinopathy (DCS data).

	Event process			
	Current value		Cumulative effect	
	Log hazard (2.5–97.5%)	*p*-value	Log hazard (2.5–97.5%)	*p*-value
Association parameter (α)	0.06 (0.05–0.07)	0.00	0.01 (0.01–0.02)	0.00
	Weighted normal		Weighted skewed normal	
	Log hazard (2.5–97.5%)	*p*-value	Log hazard (2.5–97.5%)	*p*-value
Association parameter (α)	0.07 (0.05–0.08)	0.00	0.07 (0.06–0.08)	0.00
	Weight function			
	Coefficient (2.5–97.5%)	*p*-value	Coefficient (2.5–97.5%)	*p*-value
Scale parameter (σ/ν)	0.85 (0.48–1.67)	0.00	0.65 (0.45–0.96)	0.00
Shape parameter (κ)			4.84 (0.38–9.68)	0.00

Longitudinal outcome is HbA1c.

Table II. Parameter estimates and 95% credibility intervals under the joint modelling analysis for sight threatening retinopathy (DCS data).

	Longitudinal process			
	Current value		Cumulative effect	
	Coefficient (2.5–97.5%)	*p*-value	Coefficient (2.5–97.5%)	*p*-value
Intercept	55.02 (54.06–55.95)	0.00	55.18 (54.23–56.11)	0.00
ns(years, 2)1	−12.74 (−14.52 to −10.96)	0.00	−12.76 (−14.69 to −10.91)	0.00
ns(years, 2)2	1.54 (0.43–2.67)	0.01	1.53 (0.50–2.56)	0.00
Baseline age	−0.24 (−0.30 to −0.19)	0.00	−0.24 (−0.29 to −0.19)	0.00
Baseline SBP	0.01 (−0.01 to 0.03)	0.47	0.01 (−0.01 to 0.03)	0.48
Baseline retinopathy	6.20 (4.37–8.00)	0.00	5.97 (4.16–7.83)	0.00
Duration of diabetes	0.49 (0.41–0.56)	0.00	0.44 (0.37–0.52)	0.00
σ_ϵ	9.00 (8.80–9.23)	0.00	8.97 (8.76–9.20)	0.00
	Weighted normal		Weighted skewed normal	
	Coefficient (2.5–97.5%)	*p*-value	Coefficient (2.5–97.5%)	*p*-value
Intercept	55.04 (54.09–56.01)	0.00	55.00 (54.06–55.97)	0.00
ns(years, 2)1	−12.74 (−14.62 to −10.92)	0.00	−12.68 (−14.44 to −10.86)	0.00
ns(years, 2)2	1.55 (0.52–2.67)	0.00	1.53 (0.51–2.62)	0.01
Baseline age	−0.24 (−0.29 to −0.19)	0.00	−0.24 (−0.29 to −0.19)	0.00
Baseline SBP	0.01 (−0.01 to 0.03)	0.46	0.01 (−0.01 to 0.03)	0.46
Baseline retinopathy	6.11 (4.24–7.99)	0.00	6.11 (4.24–7.90)	0.00
Duration of diabetes	0.48 (0.40–0.56)	0.00	0.48 (0.41–0.55)	0.00
σ_ϵ	8.99 (8.78–9.21)	0.00	9.00 (8.79–9.22)	0.00

Longitudinal outcome is HbA1c.

We obtain an estimated standard deviation of 0.85 (2.5–97.5% CI: 0.48–1.67) for the normal density weight function, which suggests that measurements from within the last 2.55 years (three times the standard deviation) before t are associated with the risk of an event at t, which period is indicated for two subjects from the DCS cohort in Figure 2(b), assuming an event at time $t = 6$. The estimated weight functions are shown in Figure 3. Despite the similarity in the overall relevant period of interest (more clearly demonstrated in Figure 4), the shapes of the weight functions are quite different, with the highest associated weight (and thus the value most relevant to the estimation of the relative risk), occurring slightly earlier in the case of the skewed normal distribution. Calculating the hazard ratio between the two subjects in Figure 2 at time $t = 6$ and comparing the value for the various model parameterizations, we obtain 9.22, 3.37, 11.02 and 11.04 for the current value, cumulative, weighted normal and weighted skewed normal specifications, respectively, reflecting the smaller estimated value for the association parameter

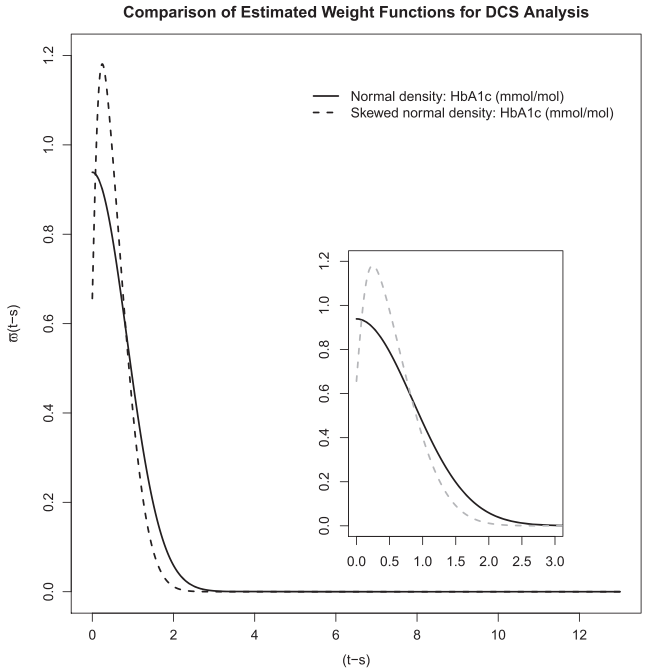

Figure 3. Estimated weight functions under the joint model specification for the development of STR, using the weighted cumulative association parameterization.

in the cumulative effect specification and the similarity between the two individuals in terms of the full area under their respective estimated biomarker profiles.

4.2. PBC data

We further illustrate the capabilities and flexibility of the weighted cumulative effect specification of the joint model using the well-known Primary Biliary Cirrhosis (PBC) data. Collected by the Mayo Clinic from 1974 to 1984 [10], the data involve 312 patients with PBC (a rare autoimmune liver disease), randomized to two treatment groups (placebo and D-penicillamine). The survival outcome of interest is transplantation or death, and several longitudinal biomarkers have been collected, specifically, serum bilirubin levels, cholesterol and presence/absence of hepatomegaly (an enlarged liver condition). Given the differential behaviour of each of these markers, the data provide us a unique opportunity to demonstrate the adaptability of the proposed methodology. Simple descriptive plots for the survival and various longitudinal outcomes are presented in Figures S4–S6. These plots illustrate the survival curves for the two treatment groups and longitudinal profiles for the logged serum bilirubin and square rooted cholesterol outcomes for patients with and without the transplantation/death endpoint.

As previously noted for the HbA1c outcome, in the DCS cohort, the subject-specific profiles over time of the logged serum bilirubin responses appear to be non-linear in shape. We therefore again include natural cubic splines in both the fixed and random effects parts of the longitudinal model for this response. We thus have

$$\eta_i(t, \boldsymbol{b}_i) = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})B_n(t, \lambda_1) + (\beta_2 + b_{i2})B_n(t, \lambda_2) + \epsilon_i(t),$$

with $\{B_n(t, \lambda_k) : k = 1, 2\}$ as previously defined.

The longitudinal models for square rooted serum cholesterol and the dichotomous hepatomegaly outcome include time as a linear effect for the fixed and random effects:

$$g[E\{y_i(t) \mid \mathbf{b}_i\}] = \eta_i(t, \boldsymbol{b}_i) = \beta_0 + b_{i0} + (\beta_1 + b_{i1}) \times \text{time},$$

where the random effects $\mathbf{b}_i$ follow a multivariate normal distribution with mean zero and unstructured variance-covariance matrix $\boldsymbol{D}$, and a standard logit link (logit(p_i)) is used for hepatomegaly. Regression splines are used for the baseline hazard in place of the penalized B-splines in the model for the square rooted serum cholesterol outcome.

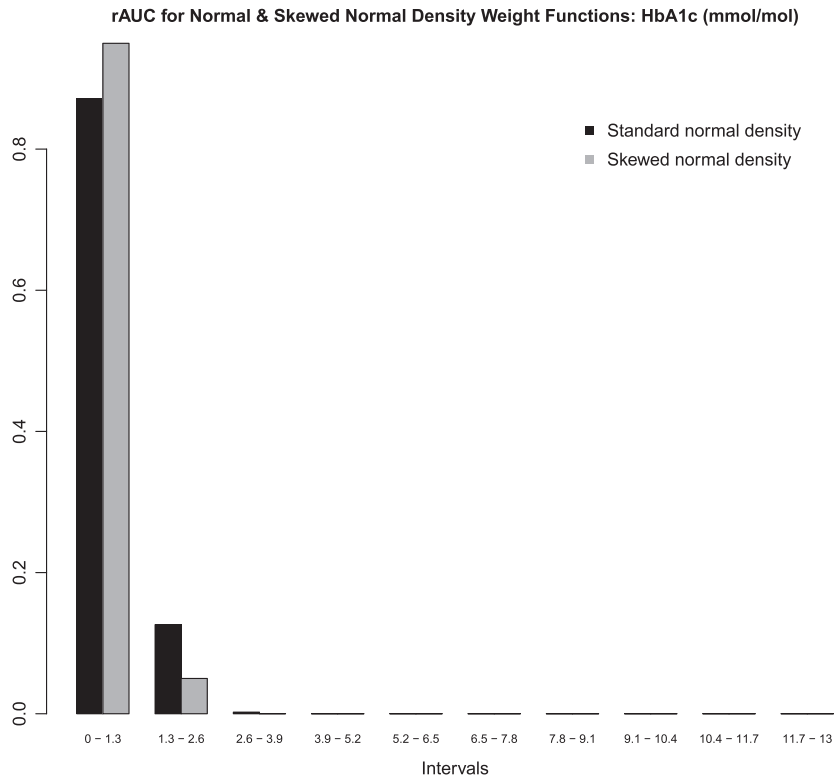

Figure 4. Relative AUC for estimated normal and skewed normal density weight functions using DCS data.

In the Cox model, we control for the effects of treatment and age, allowing for their interaction in the models for the logged serum bilirubin and hepatomegaly. As before, the weighted integral of the subject-specific linear predictor of the mixed model, $\eta_i(t, \boldsymbol{b}_i)$, is included in the linear predictor of the relative risk model, with weight functions as specified in (6) and (7).

$$h_i(t) = h_0(t)\, \exp\left\{ \gamma_1 \texttt{Dpenicil}_i + \gamma_2 \texttt{Age}_i + \gamma_3 (\texttt{Dpenicil}_i \times \texttt{Age}_i) + \alpha \int_0^t \varpi(t-s)_+ \, \eta_i(s, \boldsymbol{b}_i)\, ds \right\}.$$

The results of the models are presented in Tables S3–S8. The weighted cumulative effects specification using the normal density function is the best fitting of the two weighted models for logged serum bilirubin. This model and that using the standard current value specification both suggest similar, strong associations with the risk of transplantation/death, whereby a 1-unit increase in the value of log serum bilirubin corresponds to a 4.1-fold increase of the risk of transplantation/death (2.5–97.5% CI: 3.42–4.95), and a 1-unit increase in the area under the logged serum bilirubin profile for the relevant period of interest corresponds to a 3.94-fold increase of the risk (2.5–97.5% CI: 3.22–4.89). For the model with the cumulative effect parameterization, a 1-unit increase in the area under the logged serum bilirubin profile corresponds to a 1.2-fold increase of the risk (2.5–97.5% CI: 1.2–1.3). The estimated scale parameter for the normal density weight function is 0.1 (95% CI: 0.04–0.19). This implies that only the serum bilirubin measurements within the last 0.3 years (3.6 months) before t are associated with the risk of an event at the same time t, which suggests that measurements beyond the current value are not relevant in the risk estimation at time t. This is illustrated in Figure S7 for two randomly selected subjects, where the periods in which the weights from the normal density function are non-zero are indicated by vertical lines.

Calculating the hazard ratio for time $t = 9.68$ between the two subjects shown in Figure S7 from each of the models for logged serum bilirubin, we obtain similar values of 4.34, 3.95 and 4.05 from the models with the current value, normal and skewed normal weighted cumulative specifications, respectively. These ratios reflect an increased risk of transplantation/death for the patient represented in grey, who is 11 years older than the patient shown in black and has higher values of serum bilirubin at time $t = 9.68$. The hazard ratio calculated from the cumulative effects specification however is 0.75. As demonstrated in Figure S8, this parameterization takes into account the full 'signed area' under the longitudinal trajectory,

as previously defined, and since a portion of the curve for the subject depicted in grey is below the x-axis, and as such must be subtracted, the resulting cumulative effect for this subject is smaller than for the subject for whom we have only positive serum bilirubin values.

There does not appear to be any association with risk as estimated by the model with the current value parameterization for square rooted serum cholesterol. The cumulative parameterization indicates that a 1-unit increase in the area under the profile corresponds to a 1.01-fold increase of the risk (2.5–97.5% CI: 1.0–1.02), which is slightly smaller than the 1.16-fold increase of the risk estimated for the (best fitting) normal density function in the weighted cumulative parameterizations, with (2.5–97.5% CI: 1.07–1.27). The estimated scale parameter for the normal density weight function is 3.13 (95% CI: 1.62–4.42), which implies that the serum cholesterol measurements within the last 9.39 years before t are associated with the risk of an event at the same time t, which is almost the entire period (since maximum follow-up is 14.03 years).

The estimated association parameters from the current value specification and best fitting skewed normal weighted cumulative effect model for hepatomegaly are again similar, although the value from the weighted model is slightly larger. The weighted model suggests an increase of 0.41 (2.5–97.5% CI: 0.29–0.58), in the log hazard ratio value for every unit increase in the relevant area under the logit probability of having hepatomegaly. The estimated scale parameters for the two weight functions are 2.11 (95% CI: 0.49–4.02) and 1.82 (95% CI: 0.02–3.54) for the normal and skewed normal models, respectively, which indicates (from the normal density model) that hepatomegaly measurements within the last 6.33 years before t are associated with the risk of an event at t, again a much larger time interval than that seen with serum bilirubin. The model with the (unweighted) cumulative effect specification indicates an increase of 0.78 (2.5–97.5% CI: 0.05–0.10) in the log hazard ratio value for every unit increase in the area under the logit probability of having hepatomegaly. Visual representations of the weight functions for each of the logged serum bilirubin, square rooted serum cholesterol and hepatomegaly responses are shown in Figures S9–S13, demonstrating the degree of flexibility in the estimated weight function, with differing rates of decline and thus different periods of relevance for the different biomarkers. For hepatomegaly, as previously seen in the case of the HbA1c marker for the DCS data, the best fitting model indicates a maximal weight for values slightly earlier than the current time.

Calculating the relative AUC for each of the estimated weight functions, we then obtain Figures S14–S16, which show the relative importance of each of the biomarkers over several specific intervals of the follow-up period. We see here that the interval from 0 to 1.5 accounts for 100% (cumulatively) of the relative weight for logged serum bilirubin, where the same interval accounts for only 52% and 36.6% for hepatomegaly and square rooted serum cholesterol, respectively, and that the weights for hepatomegaly and serum cholesterol are far more equally distributed across the entire time period.

4.3. Simulation study

In order to better assess the performance of the proposed methodology, we ran a small scale simulation study. Given the complexity of the models and the associated computational time, we constrained the simulation to 100 data sets of 500 participants each.

We assumed a continuous longitudinal outcome of the following form:

$$y_i(t) = \eta_i(t, \boldsymbol{b}_i) + \epsilon_i(t) = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})B_n(t, \lambda_1) + (\beta_2 + b_{i2})B_n(t, \lambda_2) + (\beta_3 + b_{i3})B_n(t, \lambda_3) + \epsilon_i(t),$$

assuming natural cubic splines for both the fixed and random parts of the model, with $\epsilon_i(t) \sim N(0, \sigma^2)$, and $\boldsymbol{b} = (b_{0i}, b_{1i}, b_{2i}, b_{3i}) \sim MVN(\boldsymbol{0}, \boldsymbol{D_{diag}})$ and $\{B_n(t, \lambda_k) : k = 1, 2, 3\}$ as previously defined. Time was simulated from a uniform distribution between 0 and 14.3. For the survival outcome, adjusting only for treatment group (drug) for simplicity, we used

$$h_i(t) = h_0(t) \exp\left\{\gamma_1 \texttt{Drug}_i + \alpha \int_0^t \varpi(t-s)_+ \eta_i(s, \boldsymbol{b}_i)\, ds\right\},$$

with $\varpi(t-s)_+$ specified as a standard normal distribution. Baseline risk was simulated from a Weibull distribution $h_0(t) = \phi t^{\phi-1}$, with $\phi = 0.8223$. A uniform censoring distribution with mean $\mu_c = 2.6$ was chosen for the censoring times, such that the censoring rate was approximately 80%. Full details of the input parameters and results of this simulation study are presented in Section 4 of the Supporting Information, in Tables S9 and S10. Overall performance of the methodology appears to be good, with

small bias and root mean square error (RMSE) error values, with the exception of parameter D_4, the variance-covariance parameter for the random effects corresponding to the third and final interval of the natural cubic spline. This is most likely due to an insufficient number of repeated measurements in this interval.

5. Discussion

Motivated by the clinical question regarding the association between HbA1c and the risk of developing STR, we proposed an extended joint modelling framework, whereby we use a recency-weighted cumulative effect specification, with parametric weight functions. Two alternative but similar weight functions were explored, specifically, the normal and skewed normal density functions, for which the scale and shape parameters were estimated directly from the data. We further illustrated the use and behaviour of the proposed weight functions using the more well-known PBC data and using a small scale simulation study. The use of the recency-weighted cumulative effect specification allows us to more accurately determine the nature of the relationship between an endogenous time-varying covariate and the relative risk of an event of interest, allowing the calculation of the hazard function at time t to depend on a cumulative effect for the most relevant period in the history of the biomarker and for the estimation of that period of interest. Both the current value and cumulative effect specification may be seen as special cases of the weighted cumulative effect specification.

The proposed weight functions were selected based on a clinically plausible assumption regarding the functional behaviour of many biomarkers, where more recent values are expected to be more relevant in determining any associated effects.

Possible extensions would be the development of a more general family of weight functions and an increase in the number of parameters we are able to estimate. From a clinical perspective, it would be interesting to adapt and extend the methodology to allow for recurrent events or progression and also to incorporate competing events, as persons with (the same) elevated risk factors might have died before they were able to develop STR, leading to underestimation of the effect size. The estimated period of interest of approximately 2.55 years in the DCS cohort is of great interest, as it speaks to the assumed duration of glycaemic memory. This will require further research.

References

1. Ye W, Lin X, Taylor J. A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data. *Statistics and its interface* 2008; **1**:33–45.
2. Brown E, Ibrahim J, DeGruttola V. A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics* 2005; **61**:64–73.
3. Rizopoulos D, Hatfield L, Carlin B, Takkenberg J. Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. *Journal of the American Statistical Association* 2014; **109**:1385–1397.
4. Peterson KP, Pavlovich JG, Goldstein D, Little R, England J, Peterson CM. What is hemoglobin A1c? An analysis of glycated hemoglobins by electrospray ionization mass spectrometry. *Clinical Chemistry* 1998; **44**:1951–1958-8.
5. Breslow NE, Lubin JH, Marek P, Langholz B. Multiplicative models and cohort analysis. *Journal of the American Statistical Society* 1983; **78**(381):1–12.
6. Thomas DC. Models for exposure time response relationships with applications to cancer epidemiology. *Annual Reviews of Public Health* 1988; **9**:451–482.
7. Vacek PM. Assessing the effect of intensity when exposure varies over time. *Statistics in Medicine* 1997; **16**:505–513.
8. Abrahamowicz M, Bartlett G, Tamblyn R, du Berger R. Modeling cumulative dose and exposure duration provided insights regarding the associations between benzodiazepines and injuries. *Journal of Clinical Epidemiology* 2006; **59**(4):393–403.
9. Sylvestre MP, Abrahamowicz M. Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Statistics in Medicine* 2009; **28**:3437–3453.
10. Murtaugh P, Dickson E, Van Dam G, Malincho M, Grambsch P, Langworthy A, Gips C. Primary biliary cirrhosis: prediction of short-term survival. *Hepatology* 1994; **20**:126–134.
11. Brown E. Assessing the association between trends in a biomarker and risk of event with an application in pediatric HIV/AIDS. *The Annals of Applied Statistics* 2008; **3**:1163–1182.
12. Rizopoulos D. *Joint Models for Longitudinal and Time-to-Event Data with Applications in R*. Chapman and Hall/CRC Biostatistics Series: Boca Raton, 2012.
13. Fisher L, Lin D. Time-dependent covariates in the cox proportional-hazards regression model. *Annual Review of Public Health* 1999; **20**:145–157.
14. Gray R. Advanced statistical computing. *BIO 248 cd Course Notes* 2001:211–219.
15. Lang S, Brezger A. Bayesian P-splines. *Journal of computational and graphical statistics* 2004; **13**:183–212.
16. Ibrahim J, Chen M, Sinha D. *Bayesian Survival Analysis*. Springer-Verlags: New York, 2001.
17. Rizopoulos D. JMbayes: Joint Modeling of Longitudinal and Time-to-Event Data under a Bayesian Approach. R package version 0.7-2, 2015. https://CRAN.R-project.org/package=JMbayes [Access on: 3 May 2017].

18. Rizopoulos D. The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data using MCMC. ArXiv e-prints 2014; 1404.7625. http://adsabs.harvard.edu/abs/2014arXiv1404.7625R.

Supporting information

Additional supporting information may be found online in the supporting information tab for this article.