**Research Article**

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

# Gaussian process regression for survival data with competing risks

**James E. Barrett[a]*† and Anthony C. C. Coolen[a]**

**We apply Gaussian process (GP) regression, which provides a powerful non-parametric probabilistic method of relating inputs to outputs, to survival data consisting of time-to-event and covariate measurements. In this context, the covariates are regarded as the 'inputs' and the event times are the 'outputs'. This allows for highly flexible inference of non-linear relationships between covariates and event times. Many existing methods, such as the ubiquitous Cox proportional hazards model, focus primarily on the hazard rate which is typically assumed to take some parametric or semi-parametric form. Our proposed model belongs to the class of accelerated failure time models where we focus on directly characterising the relationship between covariates and event times without any explicit assumptions on what form the hazard rates take. It is straightforward to include various types and combinations of censored and truncated observations. We apply our approach to both simulated and experimental data. We then apply multiple output GP regression, which can handle multiple potentially correlated outputs for each input, to competing risks survival data where multiple event types can occur. By tuning one of the model parameters we can control the extent to which the multiple outputs (the time-to-event for each risk) are dependent thus allowing the specification of correlated risks. Simulation studies suggest that in some cases allowing for potential dependence between different risks can lead to more accurate risk-specific predictions.** 

**Keywords:** Gaussian process regression; survival analysis; accelerated failure time model; competing risks

## 1. Introduction

In this work we will develop an accelerated failure time model where the event times are written as an unknown (and noise corrupted) function of the covariates. Gaussian process (GP) regression [1] is used to infer the unknown function in a flexible and non-parametric manner. By specifying different *kernels* in the GP prior we can probabilistically infer a wide range of qualitatively different functions. From this point of view the event times are considered 'outputs' and the covariates 'inputs' in a regression model. We argue that this approach is a more direct way of connecting the quantities that we have experimental access to, namely the covariates and the event times. Many existing methods of analysing survival data focus on the hazard rate. Cox's proportional hazards model [2] is arguably the most popular such approach. These methods typically assume that the hazard rate splits into two components, one that captures the time effects and one that captures the covariate effects. Cox's model further assumes that the covariate effects combine linearly. It is not obvious however that this factorisation is always appropriate. Hazard rate models take a more indirect route that needs to capture both the time and covariate effects on survival outcomes whereas our approach need only capture the covariate effects, and consequently fewer assumptions are required.

The event times are transformed such that they take negative and positive values. Model parameters consist of the 'noise-free' function values and these are inferred in a Bayesian manner. We compute the maximum a posteriori (MAP) solution

[a] *Institute for Mathematical and Molecular Biomedicine, King's College London, London, U.K.*
* *Correspondence to: James E. Barrett, Institute for Mathematical and Molecular Biomedicine, Hodgkin Building, Guy's Campus, King's College London, London, SE1 1UL, U.K.*
† *E-mail: james.j.barrett@kcl.ac.uk*

by numerically maximising the posterior density over parameters. The hyperparameters control qualitative features of the kernel function and the overall noise level. We construct the Laplace approximation of the marginal likelihood and use that to numerically compute the MAP solution for hyperparameters.

Our model can incorporate any type of censored and truncated observations relatively easily. In addition, we obtain estimates of when the event would have occurred to individuals that were censored. We perform several simulation studies which illustrate the model's ability to infer non-monotonic relationships between the covariates and event times. We compare our model to more traditional models such as the Cox proportional hazards model, the Weibull proportional hazards model and a third model that is also based on GP regression but assumes a hazard rate similar to the Cox model but with non-linear covariate effects. We also apply our approach to experimental gene expression data.

We extend our model to the competing risks scenario by using multiple output GP regression [3]. Multiple output GP regression was originally developed for situations where multiple outputs are available corresponding to given inputs. The outputs may be statistically dependent. Again, we regard the time-to-event for different risks as the 'multiple outputs' and the covariates as the 'inputs'. In general, multiple output GP regression can be applied to data where each input has corresponding measurements of all (or some) of the outputs. There are two features of competing risks data that are interesting in this regard. Firstly, at most one output is available for each individual (since we only measure one event time). Secondly, once one of the outputs is observed we know that the remaining outputs must be greater than the observed output. This is because we know that remaining events would have occurred after the first reported event time. Despite these differences we will show that multiple output GP regression performs well on competing risks data.

The model can assume either independent risks or dependent risks by tuning the value of one hyperparameter. Of course, the identifiability problem [4] means we cannot conclude whether the risks are truly independent or not in reality. Nevertheless, within the framework of the model we will infer the value of the parameter that best explains the observed data. If the assumption of dependence has a higher probability then the model will follow this, and as we will show, exploit it to potentially make more accurate predictions. Consider, for example, two strongly dependent risks. If there is a region of the covariate space where only the first event has been observed we can still make accurate predictions of when the second type of event would occur for new individuals. This is because we know the second risk will behave similarly to the first risk. We also examine the issue of what happens in the hypothetical scenario where we 'disable' or 'switch off' one or more risks. The joint event time density takes a particularly convenient form in our model since the event times are conditionally independent given the underlying noise-free function values. Quantities such as the marginal survival probabilities are straightforward to compute.

Existing approaches to analysing competing risks survival data commonly assume parametric or semi-parametric cause specific hazard rates. This is useful to establish whether a certain covariate is associated with a particular risk. It may be less clear, however, how a covariate is related to overall survival probabilities in the presence of competing risks since the survival function is a function of all the cause specific hazard rates. An alternative approach is to model the cumulative incidence function, using for example a form similar to a proportional hazard model [5]. Shared frailty models [6], random effects models [7], and the concepts of pseudo-values [8] and relative survival [9] are other ways to analyse competing risks data. However, all of these approaches contain some parametric or semi-parametric components. Our approach differs from these strategies since we essentially focus on modelling the joint event time density in a non-parametric fashion. Similarly to the case of a single risk this avoids imposing unnecessary structural assumptions on what form the data take.

This rest of this paper is structured as follows. In Section 2 we apply GP regression to survival data with a single risk and independent censoring. We will develop our GP model, outline how to infer parameters and hyperparameters, and explain how to make predictions. This model is then applied to interval censored data. In Section 3 we extend our model to competing risks data. We apply our model to both experimental and simulated data and present the results in Section 4. We finish with discussion in Section 5.

## 2. GP regression with a single risk and independent censoring

We firstly define a general non-linear transformation model from which several existing models can be recovered under different assumptions. This will serve as a natural starting point for our GP regression model and offer an intuitive way to compare it to existing approaches. Survival data are $D = \{(\tau_i, \Delta_i)\}_{i=1\ldots,N}$ where $\tau_i > 0$ is the time until the first event for individual $i$, the indicator variable $\Delta_i = 1$ means the primary event occurred first whereas $\Delta_i = 0$ indicates individual $i$ was censored, and $N$ is the total number of individuals. In addition, we acquire a vector of covariate measurements $\mathbf{x}_i \in \mathbb{R}^d$ for each individual. A general transformation model assumes

$$\phi(\tau_i) = f(\mathbf{x}_i) + \xi_i \quad \text{for } i = 1, \ldots, N \tag{1}$$

where $\phi$ is a monotonically increasing transformation of the event times, $f(\mathbf{x}_i)$ is some function of the covariates, and $\xi_i$ is a noise random variable with a probability density function $p_\xi$.

Under different assumptions of $\phi$, $f$ and $p_\xi$ several existing models, including our GP model, can be derived as special cases of (1). For example, linear transformation models [10] assume $\phi$ is unspecified and $f(\mathbf{x}) = \boldsymbol{\beta} \cdot \mathbf{x}$ where $\boldsymbol{\beta}$ is a vector of regression weights. Various procedures for estimating the regression parameters in such models have been proposed in [11] and [12]. Recently [13] considered the case where $f(\mathbf{x})$ is an unspecified smooth function and proposed a boosting estimation method based on the marginal likelihood.

If we pick $p_\xi(s) = \exp(s - e^s)$ and $\phi(\tau) = \log \Lambda_0(\tau)$, where $\Lambda_0(\tau)$ is the integrated baseline hazard rate, we recover models with a hazard rate similar to Cox's model:

$$\pi(\tau) = \lambda_0(\tau) e^{-f(\mathbf{x})}. \tag{2}$$

The baseline hazard rate is $\lambda_0(\tau)$. When $f(\mathbf{x}) = -\boldsymbol{\beta} \cdot \mathbf{x}$ we recover Cox's original proportional hazards model. Frailty models [14] can be retrieved by assuming $f(\mathbf{x}) = -\boldsymbol{\beta} \cdot \mathbf{x} + w$ where $w$ is a frailty term. Generalised additive models [15] assume $f(\mathbf{x}) = \boldsymbol{\beta} \cdot \mathbf{x} + \sum_{\mu=1}^{d} g_\mu(x_\mu)$ where $g_\mu$ are non-linear functions of the covariates. See [16] and [17] for recent implementations of such models. Alternatively, a GP prior can be assumed for $f(\mathbf{x})$ as shown by [18] and [19]. Viewed in this order these models seek to accommodate increasingly complicated covariate effects through more flexible and sophisticated functions of the covariates. For completeness we note that accelerated failure time models can be recovered by assuming $\phi(\tau) = \log(\tau)$ and $f(\mathbf{x}) = \boldsymbol{\beta} \cdot \mathbf{x}$. Assuming different distributions for $p_\xi$ results in a wide variety of accelerated failure time models (see Section 2.6 of [20]).

### *2.1. The GP accelerated failure time model*

We let $t = \phi(\tau)$ denote the transformed event times. We could choose the traditional $t = \log(\tau)$ but instead we choose

$$t = \phi(\tau) = \log(e^{\tau/\gamma} - 1). \tag{3}$$

This transformation has some desirable features. Provided $\gamma < \min_i(\tau_i)$ then the transformation is effectively linear. A $t = \log(\tau)$ transformation would be non-linear and this will become particularly apparent for large $\tau$ since we may have two large values of $\tau$ that once transformed are rather similar to each each other. This may make it difficult for the model to make accurate inferences for large values of $\tau$. Since we will assume Gaussian noise the uncertainty associated with large event times will be the same as for short event times but with a non-linear transformation this is not desirable. Therefore (3) is preferable. The distortion due to the non-linear component of the transformation (when $\tau < \gamma$) becomes apparent only during predictions. When $t$ takes negative values they are 'squashed' towards the positive half of the real line. A plot of the transformation is given in the Supporting Information.

The transformation of the output variables in GP regression has been explored by [21]. They examine a variety of parameterised monotonic transformations and regard any transformation parameters as hyperparameters to learn during training. Their procedure infers the most appropriate transformation such that the transformed outputs can be modelled using a Gaussian process. It might be useful to apply this method in future work.

GP regression provides a powerful non-parametric probabilistic method for relating inputs $\mathbf{x}$ to outputs $t$. It is assumed that any finite collection of the noise-free outputs $f(\mathbf{x})$ are Gaussian distributed. For compactness we write $f_i = f(\mathbf{x}_i)$. The covariance is given by the *kernel* function, $k(\mathbf{x}_i, \mathbf{x}_j) = \langle (f_i - \langle f_i \rangle)(f_j - \langle f_j \rangle) \rangle$, which roughly tells us how 'similar' $\mathbf{x}_i$ and $\mathbf{x}_j$ are. We will also require the mean function $\langle f_i \rangle = m(\mathbf{x}_i)$. An excellent introduction to GP regression can be found in [1]. We can construct a GP regression model from (1) by assuming a GP prior for the noise-free function values:

$$p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) = \frac{e^{-\frac{1}{2}(\mathbf{f}-\boldsymbol{\eta}) \cdot \mathbf{K}^{-1}(\mathbf{f}-\boldsymbol{\eta})}}{(2\pi)^{N/2} |\mathbf{K}|^{1/2}} \tag{4}$$

where $[\mathbf{f}]_i = f_i$, $[\boldsymbol{\eta}]_i = \eta$ with $\eta$ constant, $[\mathbf{K}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel matrix, $\boldsymbol{\theta}$ is a vector of hyperparameters, and $\mathbf{X}$ denotes the set of $\mathbf{x}_i$ for $i = 1, \ldots, N$. In this work we have used the squared exponential kernel which is defined by $k(\mathbf{x}_i, \mathbf{x}_j) = \sigma \exp(-(\mathbf{x}_i - \mathbf{x}_j) \cdot \mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)/2)$ where $\sigma > 0$ is a hyperparameter controlling the variance of the outputs. The matrix $\mathbf{L} = \mathrm{diag}(\mathbf{l})$ where the components of $\mathbf{l} = (l_1^{-2}, \ldots, l_d^{-2})$ are known as *automatic relevance determination* (ARD) parameters and roughly tell us how important each covariate is. This is because $l_\mu$ defines a characteristic length scale over which the output associated with covariate $\mu$ varies. If the output varies a lot with a particular covariate then it is 'important'. These hyperparameters are analogous to the weights in a linear regression model or the regression coefficients in Cox regression.

For the noise variable in (1) we pick $p(\xi) = \mathcal{G}(0, \beta^2)$, where $\mathcal{G}(0, \beta^2)$ denotes a Gaussian distribution with mean 0 and variance $\beta^2$. It follows that the event time density for individual $i$ is

$$p(t_i|f(\mathbf{x}_i)) = \mathcal{G}(f(\mathbf{x}_i), \beta^2). \tag{5}$$

This is convenient since the conditional event time density has a simple form with all of the non-linear covariate effects captured by $p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})$. From this we can derive the survival function and hazard rate:

$$S(\tau) = \int_{\tau}^{\infty} \mathrm{d}s\, p(s|f_i) \quad \text{and} \quad \pi_i(\tau) = \frac{p(\tau|f_i)}{\int_{\tau}^{\infty} \mathrm{d}s\, p(s|f_i)}. \tag{6}$$

For the present section we will consider only right censoring. Interval censoring will be considered in Section 2.2. We infer the function values $\mathbf{f} \in \mathbb{R}^N$ using Bayes' theorem:

$$p(\mathbf{f}|\mathbf{X},D,\boldsymbol{\theta}) = \frac{p(D|\mathbf{f},\boldsymbol{\theta})p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})}{\int \mathrm{d}\mathbf{f}'\, p(D|\mathbf{f}',\boldsymbol{\theta})p(\mathbf{f}'|\mathbf{X},\boldsymbol{\theta})} \tag{7}$$

with $p(D|\mathbf{f},\boldsymbol{\theta}) = \prod_{i=1}^{N} P(t_i,\Delta_i|f_i)$ where $P(t_i,\Delta_i|f_i)$ is the likelihood contribution made by individual $i$ and depends on what type of censoring or truncation has occurred [20, Section 3.5]. In this case non-censored individuals contribute with the event time density evaluated at the reported event time, $P(t_i,\Delta_i=1|f_i) = p(t_i|f_i)$, and a censored individual contributes the probability that the event occurred after the reported event time, $P(t_i,\Delta_i=0|f_i) = S(t_i|f_i)$. We determine the maximum a posteriori (MAP) solution by numerically minimising the negative log likelihood $\mathcal{L}(\mathbf{f}) = -N^{-1}\log p(\mathbf{f}|\mathbf{X},D,\boldsymbol{\theta})$:

$$\mathcal{L}(\mathbf{f}) = -\frac{1}{N}\sum_{i:\Delta_i=1} \log p(t_i|f_i) - \frac{1}{N}\sum_{i:\Delta_i=0} \log S(t_i|f_i) + \frac{1}{2N}(\mathbf{f}-\boldsymbol{\eta})\cdot\mathbf{K}^{-1}(\mathbf{f}-\boldsymbol{\eta}) + \frac{1}{2N}\log|\mathbf{K}|. \tag{8}$$

Numerical optimisation is performed using a gradient based optimiser in Matlab. Partial derivatives are given in the Supporting Information. Hyperparameters are determined by optimising the Laplace approximation of the marginal likelihood $\int \mathrm{d}\mathbf{f}'\, p(D|\mathbf{f}',\boldsymbol{\theta})p(\mathbf{f}'|\mathbf{X},\boldsymbol{\theta})$. We do this by firstly expanding $\mathcal{L}(\mathbf{f})$ to second order around the minimum $\hat{\mathbf{f}}$ using a Taylor expansion $\mathcal{L}(\mathbf{f}) \approx \mathcal{L}(\hat{\mathbf{f}}) + \frac{1}{2}(\mathbf{f}-\hat{\mathbf{f}})\cdot\mathbf{A}(\mathbf{f}-\hat{\mathbf{f}})$. The matrix $\mathbf{A}$ contains second order partial derivatives and is defined by $\mathbf{A}_{ij} = \partial^2/\partial w_i \partial w_j \mathcal{L}(\mathbf{f})|_{\hat{\mathbf{f}}}$. We now write the marginal likelihood as

$$\begin{aligned} p(D|\boldsymbol{\theta}) &\approx \int \mathrm{d}\mathbf{w}\, e^{-N\mathcal{L}(\hat{\mathbf{f}}) - \frac{N}{2}(\mathbf{f}-\hat{\mathbf{f}})\cdot\mathbf{A}(\mathbf{f}-\hat{\mathbf{f}})} \\ &= p(D|\hat{\mathbf{f}},\boldsymbol{\theta})p(\hat{\mathbf{f}}|\boldsymbol{\theta})(2\pi)^{N/2}|(N\mathbf{A})^{-1}(\boldsymbol{\theta})|^{1/2}. \end{aligned} \tag{9}$$

We take the negative log of this

$$\mathcal{L}_{hyp}(\boldsymbol{\theta}) = \mathcal{L}(\hat{\mathbf{f}}) - \frac{1}{2}\log 2\pi + \frac{1}{2N}\log|\mathbf{W}+\mathbf{K}^{-1}| \tag{10}$$

where the diagonal matrix is defined by $\mathbf{W}_{ii} = -\partial^2/\partial f_i^2 \log p(D|\mathbf{f},\boldsymbol{\theta})$ (given in the Supporting Information) and numerically minimise with respect to $\boldsymbol{\theta}$. Note that each evaluation of the negative hyperparameter log likelihood requires determining $\hat{\mathbf{f}}$.

*2.1.1. Predictions, hazard rates and survival curves* Having trained a GP regression model (by inferring the function values $\mathbf{f}$ and hyperparameters $\boldsymbol{\theta}$) we may wish to predict the event time $\tau^*$ for a new individual with covariates $\mathbf{x}^*$. The predictive distribution for a test output $\mathbf{f}^*$ corresponding to a test input $\mathbf{x}^*$ is Gaussian with mean and variance

$$\hat{\mu} = \mathbf{k}^*\cdot\mathbf{K}^{-1}\hat{\mathbf{f}} \tag{11}$$

$$\hat{\kappa} = k(\mathbf{x}^*,\mathbf{x}^*) - \mathbf{k}^*\cdot(\mathbf{K}+\mathbf{W}^{-1})^{-1}\mathbf{k}^* \tag{12}$$

where $[\mathbf{k}^*]_i = k(\mathbf{x}^*,\mathbf{x}_i)$. These expression are similar to the usual GP predictive mean and variance except in this case we include additional variance due to the uncertainty in $\hat{\mathbf{f}}$ (see Section 3.4.2 of [1]). The corresponding density for (the noisy prediction) $t^*$ is $\mathcal{G}(\hat{\mu},\hat{\kappa}+\beta^2)$. Finally, we need to transform back to the original time variable $\tau^*$:

$$p(\tau^*|\mathbf{x}^*,\mathbf{X},D) = \frac{e^{-\frac{1}{2(\hat{\kappa}+\beta^2)}(\log(e^{\tau^*/\gamma}-1)-\hat{\mu})^2}}{(2\pi(\hat{\kappa}+\beta^2))^{1/2}} \frac{e^{\tau^*/\gamma}}{\gamma(e^{\tau^*/\gamma}-1)}. \tag{13}$$

Once the predictive event time density has been obtained we can compute the primary hazard rate and survival function if desired. It may also be desirable to make a specific prediction of when the event will occur. This can be done by numerically computing the mean of the event time density:

$$\langle\tau^*\rangle = \int_0^{\infty} \mathrm{d}s\, sp(s|\mathbf{x}^*,\mathbf{X},D). \tag{14}$$

The variance $\langle(\tau^*)^2\rangle - \langle\tau^*\rangle^2$ can also be computed as gives us a measure of uncertainty regarding our prediction.

### 2.2. GP regression with a single risk and independent interval censoring

Existing methods for interval censored survival data are typically based on parametric or semi-parametric models. The advantage of parametric models is that expressions for the survival function can be obtained in closed form and hence the exact likelihood can be constructed for right, left or interval censored observations. See [22] for a discussion and comparison of several parametric models. Weibull accelerated failure time models are considered in [23, 24, 25]. A family of parametric models that can handle time dependent covariates is presented in [26]. Most semi-parametric models are adaptations of Cox's model. The partial likelihood argument cannot be used so usually the full likelihood is numerically optimised with respect to parameters. Some representative examples can be found in [27, 28, 29]. Another strategy is to impute the event times [30] by taking the midpoint or the end of the interval for instance [31], and then applying standard methods to the imputed event times.

Our GP model can readily be extended to accommodate interval censored observations. Now $\Delta = 1$ corresponds to an interval censored observation and $\Delta = 0$ a right censored one. We observe upper and lower times that define an interval[†] $(t_i^l, t_i^u)$ and we have $P(t_i^l, t_i^u, \Delta_i = 1|f_i) = S(t_i^l|f_i) - S(t_i^u|f_i)$. Taking the negative log of the posterior (7) and ignoring terms independent of $\mathbf{f}$ we get

$$\mathcal{L}(\mathbf{f}) = -\frac{1}{N}\sum_{i:\Delta_i=1}\log[S(t_i^l|f_i) - S(t_i^u|f_i)] - \frac{1}{N}\sum_{i:\Delta_i=0}\log S(t_i|f_i) - \frac{1}{N}\log p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta}). \tag{15}$$

As above, we find $\hat{\mathbf{f}}$ by numerically minimising the negative log likelihood. Hyperparameters are determined using the Laplace approximation of the marginal likelihood (10) but with a different matrix $\mathbf{W}$. Inference and predictions proceed as above.

### 2.3. Weibull proportional hazards model (WPHM)

For the purposes of comparison we will use a Weibull proportional hazards model. This model assumes a more traditional hazard rate (2) with a baseline hazard rate $\lambda_0(\tau) = (\nu/\rho)(\tau/\rho)^{\nu-1}$ where $\rho > 0$ is a scale parameter and $\nu > 0$ is a shape parameter. The cumulative base hazard rate is $\Lambda_0(\tau) = (\tau/\rho)^\nu$. We infer the optimal parameter values by minimising the negative log data likelihood

$$\mathcal{L}(\boldsymbol{\beta},\rho,\nu) = -\frac{1}{N}\sum_{i:\Delta_i=1}[\log\lambda_0(\tau_i) + \boldsymbol{\beta}\cdot\mathbf{x}_i] + \frac{1}{N}\sum_{i=1}^{N}\Lambda_0(\tau_i)e^{\boldsymbol{\beta}\cdot\mathbf{x}_i}. \tag{16}$$

Predictions for new individuals with covariates $\mathbf{x}^*$ can be made by computing the mean (and variance) of the event time density (using optimal parameters $\hat{\boldsymbol{\beta}}, \hat{\rho}, \hat{\nu}$)

$$\langle\tau\rangle = \int_0^\infty \mathrm{d}s\, s\lambda_0(s)e^{\hat{\boldsymbol{\beta}}\cdot\mathbf{x}^*}\exp(-\Lambda_0(s)e^{\hat{\boldsymbol{\beta}}\cdot\mathbf{x}^*}). \tag{17}$$

## 3. Multiple output GP regression with competing risks and independent censoring

We extend the transformation model from the case of a single risk (1) to the case of two competing risks (this can easily be generalised to more than two):

$$\phi(\tau_i^1) = f_1(\mathbf{x}_i) + \xi_i^1 \quad \text{and} \quad \phi(\tau_i^2) = f_2(\mathbf{x}_i) + \xi_i^2 \quad \text{for } i = 1,\ldots,N. \tag{18}$$

Each event time is related to the same covariates via two different functions corrupted with two different noise random variables. In the case of competing risks the event times may be correlated so we will use multiple output GP regression to capture dependency between outputs. Multiple output GP regression was first introduced to the machine learning community in [3] who built on work developed in [32] which illustrated that a Gaussian process can be obtained from a convolution of a Gaussian white noise process. We will follow their approach closely in this section. The noise-free outputs are written as $f_1(\mathbf{x}) = u_1(\mathbf{x}) + s_1(\mathbf{x})$ and $f_2(\mathbf{x}) = u_2(\mathbf{x}) + s_2(\mathbf{x})$ where $u_1$ and $u_2$ are GPs unique to each output and $s_1$ and $s_2$ are 'shared' GPs obtained by convolving the same Gaussian white noise process. Dependency between outputs can be captured via the shared components.

[†]Note that we are still working with the transformed event times $t = \phi(\tau)$ defined by (3).

The covariance between the noiseless outputs is (terms such as $\langle u_r(\mathbf{x}_i), s_q(\mathbf{x}_j)\rangle$ vanish)

$$\langle f_r(\mathbf{x}_i), f_q(\mathbf{x}_j)\rangle = \langle u_r(\mathbf{x}_i), u_q(\mathbf{x}_j)\rangle + \langle s_r(\mathbf{x}_i), s_q(\mathbf{x}_j)\rangle\,. \tag{19}$$

Following the example of [3] we have

$$\begin{aligned}
\langle u_r(\mathbf{x}_i), u_r(\mathbf{x}_j)\rangle &= \frac{\pi^{d/2}\sigma^2}{\sqrt{|\mathbf{\Sigma}_r|}} e^{-\frac{1}{4}\mathbf{d}\cdot\mathbf{\Sigma}_r\mathbf{d}} & \langle s_1(\mathbf{x}_i), s_2(\mathbf{x}_j)\rangle &= \frac{(2\pi)^{d/2}\omega^2}{\sqrt{|\mathbf{\Omega}_1+\mathbf{\Omega}_2|}} e^{-\frac{1}{2}(\mathbf{d}-\boldsymbol{\mu})\cdot\mathbf{\Gamma}(\mathbf{d}-\boldsymbol{\mu})} \\
\langle s_2(\mathbf{x}_i), s_1(\mathbf{x}_j)\rangle &= \frac{(2\pi)^{d/2}\omega^2}{\sqrt{|\mathbf{\Omega}_1+\mathbf{\Omega}_2|}} e^{-\frac{1}{2}(\mathbf{d}+\boldsymbol{\mu})\cdot\mathbf{\Gamma}(\mathbf{d}+\boldsymbol{\mu})} & \langle s_r(\mathbf{x}_i), s_r(\mathbf{x}_j)\rangle &= \frac{\pi^{d/2}\omega^2}{\sqrt{|\mathbf{\Omega}_r|}} e^{-\frac{1}{4}\mathbf{d}\cdot\mathbf{\Omega}_r\mathbf{d}}
\end{aligned} \tag{20}$$

where the covariance matrices are diagonal with $\mathbf{\Sigma}_r = \Sigma_r \mathbf{I}_{d\times d}$ and $\mathbf{\Omega}_r = \Omega_r \mathbf{I}_{d\times d}$ and $\mathbf{\Gamma} = \mathbf{\Omega}_1(\mathbf{\Omega}_1+\mathbf{\Omega}_2)^{-1}\mathbf{\Omega}_2$. We assume that the characteristic length scales in covariate space are the same $\Sigma_1 = \Sigma_2 = \Omega_1 = \Omega_2 = l^{-2}$. The vector $\boldsymbol{\mu}$ allows one output to be a translation of the other. We assume $[\mu]_\nu = \mu$ for $\nu = 1,\ldots,d$ with $\mu$ constant. Finally, we shall assume that the noise levels are the same for both events $\beta_1 = \beta_2 = \beta$. In the simplest case we have a six-dimensional vector of hyperparameters $\boldsymbol{\theta} = (\eta, \mu, \beta, \sigma, \omega, l)$ where $\eta \in \mathbb{R}$ (from the GP mean), $\mu \in \mathbb{R}, \beta \geq 0, \sigma \geq 0, \omega \geq 0$ and $l \geq 0$. These simplifications are by no means necessary and may not be appropriate for certain datasets. They do however make inference of hyperparameters considerably easier since the search space will in general contain local minima so lowering the dimension of the search space will have significant computational advantages.

Inserting these into (19) allows us to construct a covariance matrix which we can use to define a GP prior over $\mathbf{f} = [\mathbf{f}_1, \mathbf{f}_2] \in \mathbb{R}^{2N}$:

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{G}\left(\boldsymbol{\eta}, \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix}\right) \tag{21}$$

where $[\mathbf{K}_{rq}]_{ij} = \langle f_r(\mathbf{x}_i), f_q(\mathbf{x}_j)\rangle$ and $[\boldsymbol{\eta}]_\nu = \eta$, with $\eta \in \mathbb{R}$, is the GP mean. The block matrices have an intuitive interpretation. $\mathbf{K}_{11}$ and $\mathbf{K}_{22}$ control the covariance structure of the independent parts of each output whereas the off-diagonal blocks control the covariance between outputs.

Returning to (18) we will now assume the GP prior (21) for the function values $\mathbf{f} = [\mathbf{f}_1, \mathbf{f}_2]$. Again we let $t_1 = \phi(\tau_1)$ and $t_2 = \phi(\tau_2)$. The indicator variable can take values $\Delta_i = 0, 1, 2$ to indicate censoring, event type 1 or event type 2 respectively. With a Gaussian distribution for the noise we obtain $t_r \sim \mathcal{G}(f_r, \beta_r^2)$ for $r = 1, 2$. Assuming that right censoring is independent the joint event time density is conditionally independent given the noise-free function values

$$p(t_i^1, t_i^2|f_i^1, f_i^2) = p(t_i^1|f_i^1)p(t_i^2|f_i^2). \tag{22}$$

The conditional independence leaves us with a rather convenient event time density. All of the complicated business of correlations between risks and similarities between individuals is captured by the GP prior leaving a simple product of univariate Gaussian densities. We will discus thus more in Section 3.3.

### *3.1. Inference of noise-free function values and hyperparameters*

We can write the data likelihood as a product of Gaussian density terms and cumulative Gaussian terms

$$p(D|\mathbf{f}_1, \mathbf{f}_2) = \prod_{r=1}^{2}\left\{\prod_{i=1}^{N}[p(t_i|f_i^r)]^{\delta_{\Delta_i,r}}[S(t_i|f_i^r)]^{1-\delta_{\Delta_i,r}}\right\}. \tag{23}$$

As before, we use Bayes' theorem (7) to calculate the posterior over the function values and then use this to obtain the negative log likelihood

$$\begin{aligned}
\mathcal{L}(\mathbf{f}) = &-\frac{1}{N}\sum_{i:\Delta_i\neq 1}\log S(t_i|f_i^1) - \frac{1}{N}\sum_{i:\Delta_i\neq 2}\log S(t_i|f_i^2) - \frac{1}{N}\sum_{i:\Delta_i=1}\log p(t_i|f_i^1) \\
&- \frac{1}{N}\sum_{i:\Delta_i=2}\log p(t_i|f_i^2) + \frac{1}{2N}(\mathbf{f}-\boldsymbol{\eta})\cdot\mathbf{K}^{-1}(\mathbf{f}-\boldsymbol{\eta}) + \log 2\pi + \frac{1}{2N}\log|\mathbf{K}|.
\end{aligned} \tag{24}$$

Hyperparameters are obtained by minimising the negative log of the Laplace approximation of the marginal likelihood. Details are given in the Supporting Information.

### *3.2. Making predictions*

The predictive distribution for the output $f_*^r$ corresponding to a new input $\mathbf{x}^*$ is Gaussian with mean and variance

$$\hat{\mu}_r = \mathbf{k}_r^*\cdot\mathbf{K}^{-1}\mathbf{f} \tag{25}$$

$$\hat{\kappa}_r = k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}_r^*\cdot(\mathbf{K}+\mathbf{W}^{-1})^{-1}\mathbf{k}_r^* \tag{26}$$

where $\mathbf{f} = [\mathbf{f}_1, \mathbf{f}_2]$ and $\mathbf{k}_r^* = [\mathbf{k}_{r1}^*, \mathbf{k}_{r2}^*]$ with $[\mathbf{k}_{rq}^*]_i = \langle f^r(\mathbf{x}^*), f^q(\mathbf{x}_i)\rangle$ given by (19). The matrix $\mathbf{K}$ is the covariance matrix in (21) formed out of four block matrices. Finally, $k(\mathbf{x}_*, \mathbf{x}_*) = \langle f^r(\mathbf{x}_*), f^r(\mathbf{x}_*)\rangle = \pi^{d/2}\sigma^2/\sqrt{|\mathbf{\Sigma}_r|} + \pi^{d/2}\omega^2/\sqrt{|\mathbf{\Omega}_r|}$. The predictive density over the original event time variable is given by (13). From this the mean and variance can be numerically computed, similarly to Section 2.1.1. Once the predictive event time density has been obtained one can readily derive hazard rates or survival curves for each risk if desired.

### *3.3. 'Disabling' a risk*

A perennial question in survival analysis is how to estimate the survival probabilities in the absence of one or more risks. It is not primarily a statistical questions since 'disabling' or eliminating one or more risks will in general alter the remaining risks because the risks will in general share biological pathways or rely on the same biological systems. The quantities we infer from data correspond to a world where all of the risks are operating so we must assume that that these quantities are relevant to the hypothetical world where one or more risk have been somehow 'disabled'. Suppose now we have a total of $R$ risks. By 'disabling' all risks except risk $r$ we mean replacing

$$p_i(\tau_0, \ldots, \tau_R) = \tilde{p}_i^r(\tau_r) \lim_{\zeta\to\infty} \prod_{q\neq r} \delta(\tau_q - \zeta) \tag{27}$$

where $\tilde{p}_i^r(\tau) = \int_0^\infty (\prod_{q\neq r} \mathrm{d}s_q) p(s_0, \ldots, s_R)$ is the marginal density of event time $r$. We use tildes to denote quantities after the risks have been disabled. The survival function is

$$\tilde{S}_i^r(\tau) = \int_{\tau_r}^\infty \mathrm{d}s\, \tilde{p}_i^r(s). \tag{28}$$

Since there is only one risk the cause specific hazard rate is:

$$\tilde{\pi}_i^r(\tau) = \tilde{p}_i^r(\tau)/\tilde{S}_i^r(\tau). \tag{29}$$

From this it follows that

$$\tilde{S}_i^r(\tau) = e^{-\int_0^\tau \mathrm{d}s\, \tilde{\pi}_i^r(s)}. \tag{30}$$

Note that $\tilde{\pi}_i^q(\tau) = 0$ and $\tilde{S}_i^q(\tau) = 1$ for all $q \neq r$. The cumulative incidence function for risk $r$ becomes

$$\tilde{C}_i^r(\tau) = \int_0^\tau \mathrm{d}s\, \tilde{\pi}_i^r(s) e^{-\int_0^s \mathrm{d}s' \tilde{\pi}_i^r(s')} = 1 - \tilde{S}_i^r(\tau). \tag{31}$$

The interpretation of the marginal survival probabilities depends on whether the risks are independent or not. We examine both cases separately.

*Dependent risks:* In this case $\tilde{\pi}_i^r(\tau) \neq \pi_i^r(\tau)$. This can be seen by comparing the expression for $\pi_i^r(\tau)$

$$\pi_i^r(\tau) = \frac{\left(\prod_{q\neq r} \int_\tau^\infty \mathrm{d}s_q\right) p_i(s_0, \ldots, s_{r-1}, \tau, s_{r+1}, \ldots, s_R)}{S_i(\tau)} \tag{32}$$

to the expression for $\tilde{\pi}_i^r(\tau)$ which is given by (29). This is to be expected since switching off the other risks will change the probability to survive until a certain time and hence the hazard rate due to risk $r$ will also change. In this case the quantity $S_i^r(\tau) = \exp(-\int_0^\tau \mathrm{d}s\, \pi_i^r(s))$ cannot be interpreted as a marginal survival probability in the hypothetical world where all other risks are switched off. Consequently, $C_i^r(\tau) = 1 - S_i^r(\tau)$ does not have a valid interpretation as a cumulative probability distribution either.

*Independent risks:* In the case of independent risks the survival function can be written as $S_i(\tau) = S_i^1(\tau)\cdots S_i^R(\tau)$ where the marginal survival functions are defined as $S_i^r(\tau) = \int_\tau^\infty \mathrm{d}s\, p_i^r(s)$ for $r = 1, \ldots, R$. Since the risks are independent it immediately follows that $\tilde{p}_i^r(\tau) = p_i^r(\tau)$. From (28) and (29) it follows that $\tilde{S}_i^r(\tau) = S_i^r(\tau)$ and $\tilde{\pi}_i^r(\tau) = \pi_i^r(\tau)$. In this case the quantity $S_i^r(\tau) = \exp(-\int_0^\tau \mathrm{d}s\, \pi_i^r(s))$ is equal to (30) and hence it can be interpreted as a marginal survival probability in the hypothetical world where all other risks are switched off.

*The GP model:* In our case the conditional independence of the event times given the noise-free function means that we can always interpret $S_i^r(\tau) = \int_\tau^\infty \mathrm{d}s\, p_i^r(s|f_i^r)$ as a marginal survival probability. This true regardless of whether the underlying functions are independent or otherwise (which in the language of our model means this is true for any value of $\omega$).

## 4. Results

Here we present result from simulation studies and experimental data. We begin by explaining how the simulated data are generated. We then apply our GP regression method to simulated data with a single risk and independent right censoring. We also test the performance of the model on interval censored data. Then we apply the model to gene expression data from a study of lymphoma patients [35]. Finally, we generate simulated competing risks data.

### *4.1. Generation of simulated data with a specified event time density*

Generation of simulated data is straightforward in the case of the GP regression model. $N$ covariate vectors are randomly generated from a uniform distribution on a finite region of the covariate space where $N$ is the number of samples we wish to generate. The corresponding kernel matrix $\mathbf{K}$ is constructed, and event times are sampled from the GP prior (4) which in practice means drawing a random vector $\mathbf{f}$ from an $N$-dimensional multivariate Gaussian density, and then adding Gaussian noise to the components of $\mathbf{f}$. Finally independent right censoring is simulated by randomly selecting a subset of the individuals and generating a random number from a uniform distribution defined on the interval $[0, \tau_i)$ which is then recorded as the time of censoring. Competing risks data are generated in the the same way but with the multiple output GP prior (21).

### *4.2. Non-monotonic simulated survival data with a single risk*

Shown in Figure 1 are results from a simulated dataset that consists of $N = 25$ individuals with a single covariate $x$. There are 13 censored individuals and 12 who have experienced the primary risk. An end of trail cutoff at 6 years has been imposed and several individuals have been censored due to this (see Figure 1 (a)).

In Figure 1 (b) we have plotted the predicted mean event time using the Weibull proportional hazards model. The WPHM is poorly suited to these data as it assumes a monotonically increasing or decreasing relationship between event times and covariates. The results from our model are shown in Figure 1 (c). The model infers the underlying function and retrieves the hyperparameters reasonably well. The inferred function gives an estimate of when event times will occur. Note that the model has extrapolated the underlying function beyond the end of trial cutoff. This can be seen in the region $x \in (-3, -2)$, and the uncertainty is also greatest in this region. In Figure 1 (d) we convert these data into interval censored data by generating a random one year interval for all of the non-censored individuals. These intervals are represented by the 'error bars' in the plot. The GP regression model is capable of recovering the underlying function.

We also implemented the GP hazard rate model from [19]. Additions results are available in the Supporting Information and show that the GP hazard rate model is also capable of inferring non-linear relationships and offers comparable performance to our GP model. A disadvantage with the hazard rate model is the difficulty in interpreting the hyperparameters. This is because the function inferred in that case describes the relationship between the covariates and the hazard rates. Hyperparameters such as the noise level and overall variance have a less intuitive interpretation.

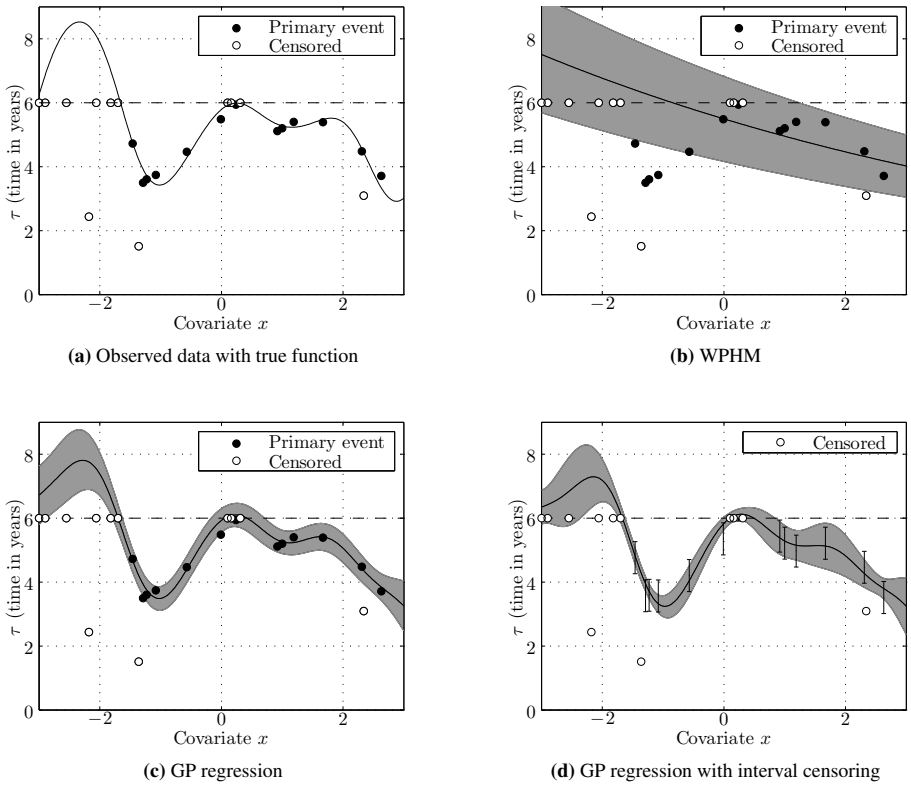
### *4.3. Experimental gene expression data*

We applied our method to the gene expression data from the Rosenwald 2002 study of lymphoma patients [35]. These data consist of $N = 240$ patients each with $d = 7399$ gene expression measurements. In the original analysis the patients had been split into a training group of 160 and a validation group of 80 individuals. These data were studied in [13] to test a transformation model with non-linear covariate effects and it was reported that some of the gene expression levels had a non-linear relationship with the time-to-event. We examined one of these genes, with UNIQID $= 33014$, with our GP method and also found a non-linear function $f(\mathbf{x})$. This function was inferred using the 160 training individuals and can be seen in Figure 2 (a). If we compare this to the top right panel of Figure 2 in [13] we can see that both functions are very similar (once we ignore the fact that, by definition, they differ in sign).

To further quantify the difference between our GP method and the WPHM we computed the mean square error (MSE) between the predicted time-to-event and the reported time-to-event in both the training and validation sets for both models. The results are displayed in Table 1. It is clear that the GP method offers vastly superior performance compared the WPHM. We can also see that the WPHM validation error is considerably larger than the training error. This is a hallmark of overfitting where the model fails to generalise well to unseen data. GP regression on the other hand does not suffer from this problem on this dataset.

### *4.4. Comparison of GP models with dependent and independent competing risks*

In order to test the performance of GP regression in the presence of competing risks we generated survival data with dependent risks and compare two GP models, one which allows for dependency between risks and one with independent risks. Recall that by fixing the value of $\omega = 0$ we force the two risks to be independent. The results are shown in Figure 3.




**(a)** Observed data with true function

**(b)** WPHM

**(c)** GP regression

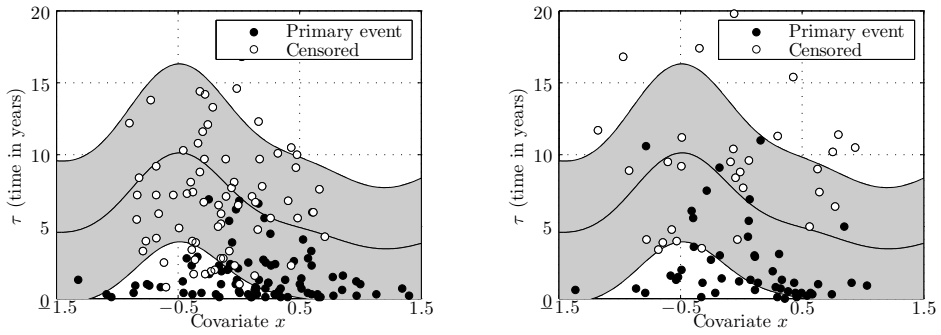
**(d)** GP regression with interval censoring

**Figure 1.** Results from a simulated dataset with $d = 1$ generated with a squared exponential kernel with hyperparameters set to $(\eta, \beta, \sigma, l) = (5, 0.2, 3, 0.7)$. There are $N = 25$ individuals, 13 of which are censored. The end of trail at 6 years is represented by the dashed line. Figure (a) shows the observed data with the 'true' function. Figure (b) is a plot of the predicted event time using the Weibull proportional hazards model. The grey region represents plus and minus one standard deviation from the mean prediction. We found $(\beta, \rho, \nu) = (0.49, 6.0, 4.7)$. The WPHM fails to infer the correct function since it assumes a monotonic relationship between covariates and event times. In (c) the mean prediction using our model is shown (i.e. we have plotted (13) as a function of $x$). Optimal hyperparameters were found to be $(\eta, \beta, \sigma, l) = (5.82, 0.32, 2.59, 0.64)$. Note the increased uncertainty at $x \in (-3, -2)$ where only censored observations were made. In (d) the non-censored observations were converted to interval censored observations by generating a random one year interval which are represented by the error bars. The inferred hyperparameters are $(\eta, \beta, \sigma, l) = (5.67, 0.14, 3.34, 0.57)$. The GP model is clearly capable of recovering non-monotonic relationships.

| | GP regression | WPHM |
|---|---|---|
| Training MSE (years$^2$) | 22.86 | 774.96 |
| Validation MSE (years$^2$) | 22.38 | 2514.7 |

**Table 1.** Comparison of mean square error (MSE) between the predicted and reported event times in the validation set using gene number 33014 from the Rosenwald lymphoma dataset [35]. Our GP regression method offers superior performance to the WPHM because it has detected a non-linear relationship between the event times and that gene expression level. The WPHM also overfits the validation data since the MSE is considerably larger than the training MSE. The GP model does not suffer from this problem in this case.

In Figure 3 (a) and Figure 3 (b) are results from a GP model where $\omega$ is inferred from the data. Hyperparameters were found to be $(\eta, \mu, \beta, \sigma, \omega, l) = (4.59, 0.41, 0.33, 0.20, 1.63, 1.01)$. The higher value of $\omega$ indicates that the model is assuming strong dependence between risks. In (c) and (d) are results from a second GP model with $\omega = 0$. Remaining hyperparameters were found to be $(\eta, \mu, \beta, \sigma, l) = (13.03, -0.60, 1.02, 1.90, 1.02)$. Note that the value of $\sigma$ is now higher as the unique part of each risk must explain all of the output variance. The advantage of allowing dependent risks becomes apparent when we examine the inferred risk 2 function towards the left of (b) and (d). In the independent model the uncertainty associated with the underlying function is much greater since knowledge of risk 1 is unavailable. In the dependent model a more accurate recovery of the risk 2 function is obtained and the uncertainty is smaller since information from risk 1 events can be utilised more effectively. In this dataset the generating hyperparameters were $(\eta, \mu, \beta, \sigma, \omega, l) = (5, 0.5, 0.5, 0.5, 2.0, 1.0)$ so the dependent model performs much better, particularly towards the left of the $x$-axis despite the complete lack of risk 2 observations. Of course, with real data we will not have the luxury of




**(a)** GP regression training data

**(b)** GP regression validation data

**Figure 2.** Univariate analysis of gene number 33014 from the Rosenwald lymphoma dataset [35]. In (a) is the function inferred on the training set (of 160 patients) using our GP regression method which clearly shows a non-linear relationship between the expression levels and event times. In (b) is the same function superimposed on the validation set. Visually, the inferred function is a plausible fit.

knowing whether an assumption of dependence is correct or not but this example nevertheless illustrates the potential usefulness of our approach.

### *4.5. Application to multi-dimensional covariates*

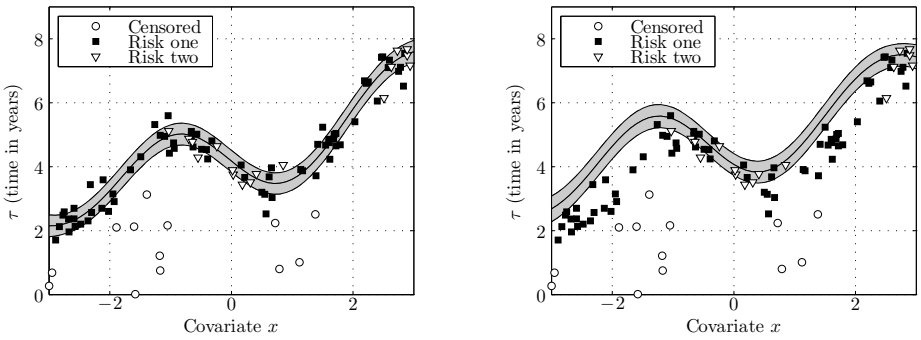
It is straightforward to deal with more than one covariate. Using the ARD hyperparameters outlined in Section 2.1 we can determine which covariates have the biggest impact on survival outcomes. In the Supporting Information we give an example of competing risks data with two covariates. Inferred ARD hyperparameters are $(l_1, l_2) = (0.52, 1.47)$ indicating that the first covariate is more important.

## 5. Discussion and conclusion

We have pursued an alternative route to many existing survival analysis methods — which assume a parametric or semi-parametric hazard rate — and focus on directly characterising the relationship between covariates and event times in a flexible and non-parametric manner. GP regression provides a powerful and elegant means to achieve this. All relevant quantities are inferred in Bayesian manner and we can obtain probabilistic predictions, survival probabilities and hazard rates. It is straightforward to incorporate censored or truncated observations (or combination thereof). We found that the GP hazard rate model used in [19] also performed well on non-monotonic data but the hyperparameters are not easy to interpret. This is because the inferred function represents the relationship between the covariates and hazard rate whereas in our case the inferred function is conceptually more straightforward and easier to interpret.

An interesting question to consider is how to interpret the underlying function we infer. In standard GP regression the function values would be considered 'noise-free' outputs which are then corrupted by Gaussian observational noise. The corresponding interpretation in our case would be that the functions represent a 'noise-free' event time. However, Gaussian noise is not appropriate in that case since it is generally not plausible to claim events could be randomly reported before they actually occur. A more appropriate choice would be noise with a semi infinite support on $(0, \infty]$ that would represent a delay between the event occurring and the time it is diagnosed or recorded. In that case the underlying function could be interpreted as a noise-free event time. An alternative interpretation of the noise, however, is that it represents a disparity or mismatch between the assumed model and the actual model of the data. In that case it is acceptable to regard the function values as noise-free event times. In future work it may be interesting to explore alternative noise distributions.

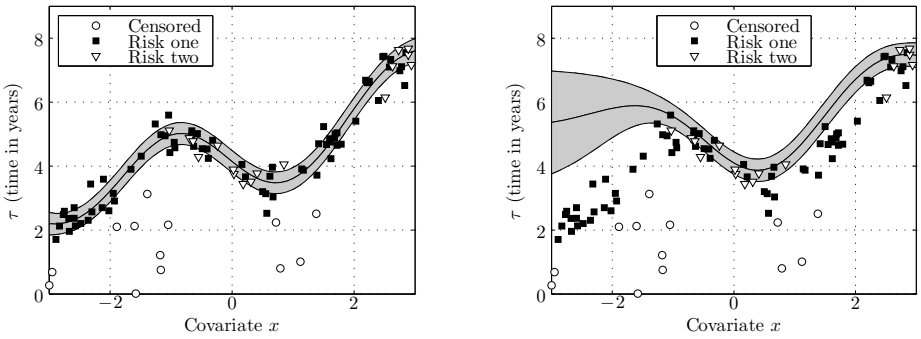
Multiple output GP regression provided a natural route to incorporate competing risks data. Working with the event time density — also called the latent event time approach — has been criticised for a number of reasons. The most serious objection is that the joint event time density is unobservable and cannot be inferred from the observed failure and censoring times (due to the identifiability problem). While one may assume that the event times are independent or perhaps assume some parametric density to model dependencies one cannot conclude on the basis of the observed data alone whether or not the event times are independent. Some authors such as [36, Section 3.3] have claimed that the latent failure times lack plausibility or are too hypothetical in nature since we are positing the existence of quantities which in reality can never be measured (see [33, Section 8.2] for further in depth discussion).




**(a)** Dependent risk 1




**(b)** Dependent risk 2




**(c)** Independent risk 1




**(d)** Independent risk 2

**Figure 3.** Results from simulated data with two competing risks. In (a) and (b) are inferred functions from both risks using a multiple output GP model with dependent risks allowed. Values of $\sigma = 0.2$ and $\omega = 1.63$ were found which indicate strongly dependent risks. In (c) and (d) the inferred risks are forced to be independent by setting $\omega = 0$. A value of $\sigma = 1.9$ is found which is sensible since the unique part of each risk must account for all of the variance. The advantage of assuming dependency can be seen around $x \in (-3, -1)$ in (b) and (d). There are no risk two observations in this region and in the independent model there is much greater uncertainty since information form risk one cannot be utilised.

We argue instead that the latent failure time approach does provide a useful conceptual framework. There are two aspects in particular where such an approach is useful. The first is in making predictions for new patients (who are still alive) since the time until the different events is highly relevant. The second is in estimating what happens when one or more risks are disabled. In both cases the marginal survival functions are the relevant quantities. In our GP formulation both predictions and marginal survival probabilities are straightforward to compute due to the conditional independence of the event times. If we want to model dependent risks (despite not being able to test our assumption due to identifiability issues) then modelling the joint event time density is a convenient starting point. The fact that the data we observe in reality do not allow a direct view of this joint density does not mean that it is not a useful concept.

We have illustrated that GP regression provides a flexible method of analysing survival data. We impose few structural assumptions due to the non-parametric nature of GP regression and avoiding specification of the hazard rate. Future research could involve applying sparse GP regression techniques to our model in order to achieve greater computational efficiency.

## Acknowledgement

This work was funded under the European Commission FP7 Imagint Project, EC Grant Agreement number 259881.

## References

1. Rasmussen CE, Williams CKI. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.
2. Cox DR. Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society: Series B (Methodological)* 1972; **34**(2):187–220.
3. Boyle P, Frean M. Dependent Gaussian processes. *Advances in neural information processing systems* 2005; **17**:217–224.
4. Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences* 1975; **72**(1):20–22.

5. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 1999; **94**(446):496–509.
6. Hougaard P. *Analysis of Multivariate Survival Data*. Springer-Verlag New York, Inc., 2000.
7. Vaida F, Xu R. Proportional hazards model with random effects. *Statistics in medicine* 2000; **19**(24):3309–3324.
8. Andersen PK, Perme MP. Pseudo-observations in survival analysis. *Statistical methods in medical research* 2010; **19**(1):71–99.
9. Lambert PC, Dickman PW, Nelson CP, Royston P. Estimating the crude probability of death due to cancer and other causes using relative survival models. *Statistics in medicine* 2010; **29**(7-8):885–895.
10. Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data. *Biometrika* 1995; **82**(4):835–845.
11. Fine JP, Ying Z, Wei LG. On the linear transformation model for censored data. *Biometrika* 1998; **85**(4):980–986.
12. Chen K, Jin Z, Ying Z. Semiparametric analysis of transformation models with censored data. *Biometrika* 2002; **89**(3):659–668.
13. Lu W, Li L. Boosting method for nonlinear transformation models with censored survival data. *Biostatistics* 2008; **9**(4):658–667.
14. Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 1979; **16**(3):439–454.
15. Fahrmeir L, Kneib T. *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. Oxford University Press, 2011.
16. Martino S, Akerkar R, Rue H. Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics* 2011; **38**:514 – 528.
17. Vanhatalo J, Riihimäki J, Hartikainen J, Jylänki P, Tolvanen V, Vehtari A. GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research* April 2013; **14**(1):1175–1179.
18. Savitsky T, Vannucci M, Sha N. Variable selection for nonparametric Gaussian process priors: models and computational strategies. *Statistical Science* 2011; **26**(1):130–149.
19. Joensuu H, Vehtari A, Riihimäki J, Nishida T, Steigen SE, Brabec P, Plank L, Nilsson B, Cirilli C, Braconi C, *et al.*. Risk of recurrence of gastrointestinal stromal tumour after surgery: an analysis of pooled population-based cohorts. *The Lancet Oncology* 2012; **13**(3):265–274.
20. Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data, Second Edition*. Springer Science+Buisness Media, LLC, 2003.
21. Snelson E, Rasmussen CE, Ghahramani Z. Warped gaussian processes. *Advances in neural information processing systems* 2004; **16**:337–344.
22. Lindsey JK. A study of interval censoring in parametric regression models. *Lifetime Data Analysis* 1998; **4**(4):329–354.
23. Odell PM, Anderson KM, D'Agostino RB. Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics* 1992; :951–959.
24. Rabinowitz D, Tsiatis A, Aragon J. Regression with interval-censored data. *Biometrika* 1995; **82**(3):501–513.
25. Komárek A, Lesaffre E. The regression analysis of correlated interval-censored data illustration using accelerated failure time models with flexible distributional assumptions. *Statistical Modelling* 2009; **9**(4):299–319.
26. Sparling YH, Younes N, Lachin JM, Bautista OM. Parametric survival models for interval-censored data with time-dependent covariates. *Biostatistics* 2006; **7**(4):599–614.
27. Sinha D, Chen MH, Ghosh SK. Bayesian analysis and model selection for interval-censored survival data. *Biometrics* 1999; **55**(2):585–590.
28. Satten GA. Rank-based inference in the proportional hazards model for interval censored data. *Biometrika* 1996; **83**(2):355–370.
29. Goetghebeur E, Ryan L. Semiparametric regression analysis of interval-censored data. *Biometrics* 2000; **56**(4):1139–1144.
30. Law CG, Brookmeyer R. Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in medicine* 1992; **11**(12):1569–1578.
31. Pan W. A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* 2000; **56**(1):199–203.
32. Higdon D. Space and space-time modeling using process convolutions. *Quantitative methods for current environmental issues*. Springer, 2002; 37–56.
33. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2002.
34. Bernoulli D. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour le prévenair. *Historie avec le Mémoires, Académie Royal des Sciences* 1760; .
35. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltnane JM. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine* 2002; **346**(25):1937–1947.
36. Beyersmann J, Allignol A, Schumacher M. *Competing Risks and Multistate Models with R*. Springer, 2012.