A latent class model for competing risks

M. Rowley[a,b], H. Garmo[c], M. Van Hemelrijck[c], W. Wulaningsih[c], B. Grundmark[d,e], B. Zethelius[f,e], N. Hammar[g,h], G. Walldius[i], M. Inoue[j], L. Holmberg[c] and A.C.C. Coolen[a]

a Institute for Mathematical and Molecular Biomedicine, King's College London, London, U.K.
b Saddle Point Science, London, U.K.
c Cancer Epidemiology Group, King's College London, Guy's Hospital, London, U.K.
d Department of Surgical Sciences, Uppsala University, Uppsala, Sweden.
e Medical Products Agency, Uppsala, Sweden.
f Department of Public Health and Caring Sciences/Geriatrics, Uppsala University, Sweden.
g Department of Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Sweden.
h AstraZeneca Sverige, Södertalje, Sweden.
i Department of Cardiovascular Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Sweden.
j Department of Electrical Engineering and Bioscience, Waseda University, Tokyo, Japan.

Correspondence to: Mark Rowley, Institute for Mathematical and Molecular Biomedicine, King's College London, Hodgkin Building, London SE1 1UL, U.K. E-mail: mark.rowley@kcl.ac.uk

Contract/grant sponsor: Prostate Cancer UK, European Union FP-7 Programme (IMAGINT), and the Ana Leaf Foundation

Abstract

Survival data analysis becomes complex when the proportional hazards assumption is violated at population level, or when crude hazard rates are no longer estimators of marginal ones. We develop a Bayesian survival analysis method to deal with these situations, based on assuming that the complexities are induced by latent cohort or disease heterogeneity that is not captured by covariates, and that proportional hazards hold at the level of individuals. This leads to a description from which risk-specific marginal hazard rates and survival functions are fully accessible, 'decontaminated' of the effects of informative censoring, and which includes Cox, random effects and latent class models as special cases. Simulated data confirm that our approach can map a cohort's substructure, and remove heterogeneity-induced informative censoring effects. Application to data from the ULSAM cohort leads to plausible alternative explanations for previous counter-intuitive inferences on prostate cancer. The importance of managing cardiovascular disease as a comorbidity in women diagnosed with breast cancer is suggested on application to data from the AMORIS study.

Keywords: survival analysis; heterogeneity; informative censoring; competing risks

1 Introduction

The analysis of survival data is often complicated by latent cohort or disease heterogeneity and informative censoring arising from competing risks. In its simplest form, such heterogeneity could reflect unobserved covariates, but a cohort could also exhibit variations in risk associations or in the temporal features of base hazard rates. Association heterogeneity tends to cause a proportional hazards assumption to be violated at cohort level. Moreover, if latent heterogeneity affects several risks in a correlated manner, it may cause informative censoring. A cohort is subject to informative censoring if the event times of the primary and non-primary (competing) risks are statistically dependent. Unfortunately, one cannot infer the presence or absence of event-time correlations from survival data alone [1, 2]. Unaccounted for risk correlations can lead to incorrect inferences or interpretations [3, 4, 5, 6, 7], and the importance of having reliable epidemiological tools for isolating effects from interrelated comorbid diseases is increasingly recognised [8]. In discussing informative censoring in survival data, we shall refer to the *crude* cause-specific hazard rates and survival

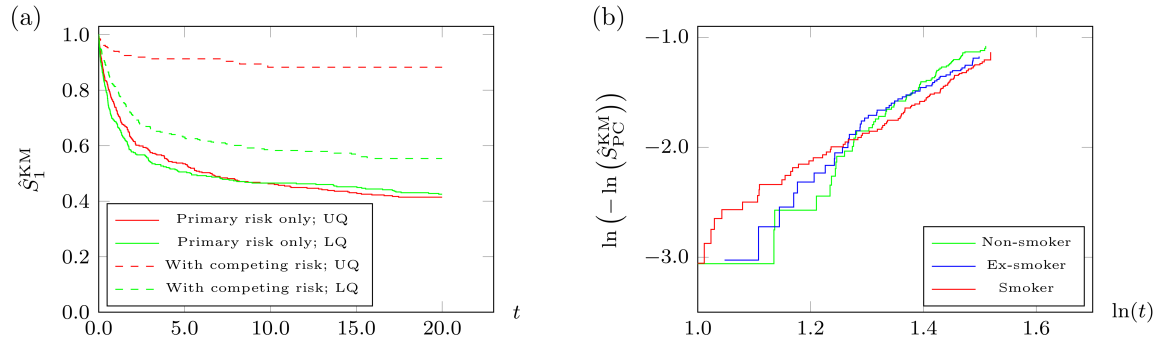

Figure 1: *Effects and signatures of heterogeneity and informative censoring:* Panel (a) illustrates the danger of misinterpreting Kaplan-Meier (KM) estimators in the presence of competing risks. The KM-curves for the lower and upper quartiles (LQ and UQ) of a covariate associated with the primary risk are presented for two simulated cohorts, sharing *identical* primary risk characteristics. In the first cohort (solid lines) no informative censoring is present. The second cohort (dashed lines) is subjected to a competing risk whose event times correlate with those of the primary risk. In the absence of informative censoring, the LQ versus UQ KM-curves are similar. In the presence of informative censoring, the LQ and UQ KM curves suggest completely different primary risk survivals. Characteristics for these cohorts are based on those of *Cohort B*, described in Section 3, where risk $r = 2$ simulates the competing risk. The crossing log-log survival curves in panel (b) indicate a violation of the proportional hazards assumption in the different smoking strata in the ULSAM cohort [35, 36]. The non-proportional hazards for prostate cancer (r =PC) smoking status in combination with stratum dependent competing risks, is suggestive of heterogeneity; the KM-estimates indicate that the survival of smokers is greater than that of ex- and non-smokers (an alternative view is suggested from our analysis; see Section 4).

functions, in which the influence of competing risks is present, and their *marginal* counterparts, which have been decontaminated of the effects of competing risks. Inference from survival data of the marginal risk characteristics, which describe the hypothetical situation where all non-primary risks were disabled, requires that the competing risk problem is addressed [9].

Conventional survival analysis tools, such as Kaplan-Meier estimators [10] and Cox regression [11], cannot distinguish between different mechanisms of informative censoring. True protectivity and aetiology effects describe direct associations of covariates with the primary risk that would have persisted even in the absence of competing risks, whereas false protectivity and aetiology effects would have disappeared. The interpretation of Kaplan-Meier curves or cause-specific hazard ratios in Cox regression can therefore be dangerous when there are many censoring events. This is demonstrated in Figure 1(a), which shows that, due to informative censoring by a competing risk, the covariate-conditioned estimators for the primary risk differ significantly between two simulated heterogeneous cohorts, even if these cohorts have identical primary risk characteristics. In both cohorts, the relevant covariate was associated with increased hazard against the primary risk for one half of the cohort, and with reduced hazard for the other half; informative censoring was simulated in one cohort through the introduction of a competing risk, with the covariate being associated with increased hazard against the competing risk for those cohort members for which the covariate was associated with increased hazard against the primary risk. However, Kaplan-Meier estimators can offer clues as to the presence of heterogeneity and informative censoring in a cohort. Inspecting covariate-stratified risk-specific Kaplan-Meier estimators can confirm violation of the proportional hazards assumption, as is evident for the cohort in Figure 1(b).

Many approaches have been used to model residual cohort heterogeneity, usually building on Cox-type cause-specific hazard rates, such as random effects models e.g. [12, 4, 13, 14, 15, 16, 17, 18, 19, 20, 21] and latent class models e.g. [22, 23, 24, 25, 26]. However, most random effects models quantify only the hazard rate of the primary risk. They can capture some consequences of cohort heterogeneity, but without also modelling non-primary risks it is impossible to deal with the competing risk problem. Another attempt to address competing risks is made in [27] where the authors parametrise the covariate-conditioned cumulative incidence function, an estimator of the cause-specific cumulative probability. The approach is conceptually similar to that of [11]; both model the primary risk profile in the presence of all competing risks. Cumulative incidence functions appear more intuitive than hazard rates as they estimate the cause-specific cumulative event probability, irrespective of the presence of correlated event times. However, expressing the likelihood

function counterpart of the cumulative incidence function is more cumbersome than expressing it for the hazard rates in [11]. While [27] thus quantifies cause-specific cumulative probabilities, the competing risk problem is still not addressed. Further developments involve e.g. alternative parametrisations [28, 29], application to the cumulative incidence of non-primary risks [30], and the inclusion of frailty factors [31]. Other authors have focused on identifying which mathematical constraints need to be imposed on multi-risk survival analysis models in order to circumvent the identifiability problem of [1], and infer the joint event-time distribution unambiguously from survival data e.g. [32, 33, 34].

Informative censoring can either appear at cohort-level as a consequence of residual (disease- or patient-) heterogeneity not captured by the recorded covariates, or the different risks are dependent already at the level of individual patients. We refer to the former as heterogeneity-induced informative censoring, and to the latter as individual-level risk dependence; one cannot distinguish between these on the basis of survival data alone. In this paper, we assume informative censoring to be of the heterogeneity-induced type; this is an *unverifiable assumption*. However, it is much weaker than assuming risk independence, and still found to impose sufficient constraints to enable us to derive exact formulae for cause-specific marginal hazard rates and survival functions. In this paper, we develop a generic model, that builds upon the random effects and latent class approaches, from which relative frailties, covariate associations, and base hazard rates for *each* latent class and for *all* risks can be estimated. Our choice of *simultaneously* modelling *all* risks exploits the cohort information more fully than traditional analyses focused only on the primary risk, and our analysis employs Bayesian model selection for optimal cohort substructure and parameter estimation.

In Section 2 we first classify the distinct levels of 'risk complexity' from the competing risk perspective in a cohort, and define the concept of heterogeneity-induced informative censoring, before presenting the mathematical development behind our latent class approach to survival analysis. In Section 3 the effectiveness of our analysis to characterise heterogeneous cohorts is demonstrated, via application to simulated survival data. The results obtained are compared with those obtained from Kaplan-Meier estimators and Cox regression. In Sections 4 and 5 we apply our method to real prostate and breast cancer survival data from the ULSAM longitudinal cohort [35, 36] and the AMORIS cohort [37, 38], respectively. The application to the ULSAM data leads to appealing and transparent new explanations for previously counter-intuitive inferences. Age-related survival differences between women diagnosed with breast cancer were found on application to data from the AMORIS cohort. In Section 6 we summarise our findings.

2 Heterogeneity-induced informative censoring

Understanding the relationships between heterogeneity-induced informative censoring, appearing at the cohort level, and risk dependence at individual level, is critical to any analysis that seeks to address the issue of residual cohort heterogeneity. This is formalised in Section 2.1. In Section 2.2 we discuss the extent to which different risk (in)dependence assumptions limit survival analysis. Our present approach assumes that risk event-time correlations are caused by residual cohort heterogeneity. While this assumption is weaker than assuming risk independence, it still leads to intuitive and transparent parametrisations of hazard rates, and allows us to overcome informative censoring and make quantities decontaminated of its effects accessible. Formulae for marginal cause-specific survival functions and hazard rates are derived in Section 2.3. In Section 2.4 the we outline our latent class approach to modelling heterogeneous cohorts (with further details and identities given in Appendix A). Details of our Bayesian approach for optimal cohort parameter estimation are given in Section 2.5, and in Section 2.6 the various outputs of our analysis and their benefits are discussed.

2.1 Connection between cohort-level and individual-level descriptions

The standard mathematical relations of multi-risk survival analysis can be derived directly from the underlying joint event-time distributions. Below, we formalise the relationships between the cohort as a whole and its individual members for a number of quantities of interest; the index i is used to denote quantities specific to individual i, with no such index being present for cohort-level quantities. We imagine a cohort of N individuals, who are each subject to R true risks, labelled by $r = 1 \dots R$, and an end-of-trial censoring event denoted by $r = 0$. In the interest of readability, products and summations which run over the true risks and the end-of-trial risk shall be written as $\prod_r$ and $\sum_r$ respectively, such that $r = 0, \dots, R$; should a particular risk r' be excluded then the product shall be written as $\prod_{r \neq r'}$. Summations over the cohort run over all individuals, and shall be written as $\sum_i$, where $i = 1, \dots, N$.

The function $p(t_0, \dots, t_R)$ shall be used to describe, for the cohort as a whole, the (unknown) probability

density of the joint event times $(t_0, \ldots, t_R)$, where $t_r \geq 0$ is the time at which risk r triggers an event. The personalised event-time distribution of any individual i in this cohort is denoted by $p_i(t_0, \ldots, t_R)$. The cohort-level risk event-time density is, by definition, a direct average of the personalised risk event-time densities, such that $p(t_0, \ldots, t_R) = N^{-1} \sum_i p_i(t_0, \ldots, t_R)$. The covariate-conditioned cohort-level risk event-time distribution describes the risk event-time statistics of the sub-cohort of those individuals satisfying $\mathbf{z}_i = \mathbf{z}$, and is the average of their personalised event-time distributions, such that $p(t_0, \ldots, t_R|\mathbf{z}) = n_{\mathbf{z}}^{-1} \sum_{i,\ \mathbf{z}_i=\mathbf{z}} p_i(t_0, \ldots, t_R)$, where $n_{\mathbf{z}} = \sum_{i,\ \mathbf{z}_i=\mathbf{z}} 1$.

The crude cause-specific hazard rates follow from the joint event-time distribution. At cohort level, the covariate-conditioned cause-specific hazard rates are given by,

$$h_r(t|\mathbf{z}) = \frac{1}{S(t|\mathbf{z})} \int_0^\infty \ldots \int_0^\infty \mathrm{d}t_0 \ldots \mathrm{d}t_R \; p(t_0, \ldots, t_R|\mathbf{z}) \delta(t - t_r) \prod_{r' \neq r} \theta(t_{r'} - t), \tag{1}$$

where $S(t|\mathbf{z})$ represents the covariate-conditioned survival function, the delta-distribution $\delta(x)$ is defined by the identity $\int_{-\infty}^{\infty} \mathrm{d}x \; \delta(x) f(x) = f(0)$, and the step function is defined as $\theta(x > 0) = 1$ and $\theta(x < 0) = 0$. The cohort-level covariate-conditioned survival function, $S(t|\mathbf{z})$, is given by,

$$S(t|\mathbf{z}) = \mathrm{e}^{-\sum_{r'} \int_0^t \mathrm{d}s \; h_{r'}(s|\mathbf{z})}. \tag{2}$$

The covariate-conditioned joint risk event-time probability density, $p(t, r|\mathbf{z})$, is given by,

$$p(t, r|\mathbf{z}) = h_r(t|\mathbf{z}) \mathrm{e}^{-\sum_{r'} \int_0^t \mathrm{d}s \; h_{r'}(s|\mathbf{z})}. \tag{3}$$

Analogous relations define the cause-specific hazard rate $h_r^i(t)$, the survival function $S_i(t)$, and the joint risk event-time probability density $p_i(t, r)$, for each individual i. These relations are of identical form to those above, though individual-level quantities replace cohort-level quantities and explicit covariate-conditioning is unnecessary.

The relationship between quantities at cohort-level and at individual-level is simple for those which depend linearly on the event-time distribution. For instance, the cohort-level survival function is the average of the individual survival functions, $S(t|\mathbf{z}) = n_{\mathbf{z}}^{-1} \sum_{i,\ \mathbf{z}_i=\mathbf{z}} S_i(t)$, and the cohort-level risk event-time probability density is given by $p(t, r|\mathbf{z}) = n_{\mathbf{z}}^{-1} \sum_{i,\ \mathbf{z}_i=\mathbf{z}} p_i(t, r)$, with $n_{\mathbf{z}} = \sum_{i,\ \mathbf{z}_i=\mathbf{z}} 1$. In contrast, cohort-level expressions for quantities that depend on the risk event-time distribution in a more complicated way, such as crude cause-specific hazard rates, are *not* direct averages over their individual-level counterparts. The cohort-level cause-specific hazard rates, for instance, are (Appendix A.1),

$$h_r(t|\mathbf{z}) \quad = \quad \frac{\sum_{i,\mathbf{z}_i=\mathbf{z}} h_r^i(t) \mathrm{e}^{-\sum_{r'} \int_0^t \mathrm{d}s \; h_{r'}^i(s)}}{\sum_{i,\mathbf{z}_i=\mathbf{z}} \mathrm{e}^{-\sum_{r'} \int_0^t \mathrm{d}s \; h_{r'}^i(s)}}. \tag{4}$$

The cause-specific cumulative incidence function, $F_r(t) = \int_0^t \mathrm{d}t' \; S(t') h_r(t')$, describes the probability that event r has been *observed* at any time prior to time t. Although $F_r(t)$ refers to risk r specifically, it can be heavily influenced by other risks. Without additional assumptions, it is not possible to distinguish between the cumulative incidence being small because event r is intrinsically unlikely, or because it tends to be preceded by other (competing) events $r' \neq r$.

2.2 Risk complexity in heterogeneous cohorts

In this section, we explore the consequences of risk event-time correlations and cohort-level heterogeneity. Cohorts are generally expected and allowed to be heterogeneous in terms of covariates. Here we refer to heterogeneity in terms of the relationship between covariates and risks. A homogeneous cohort is one in which the relationship between an individual's covariates and their risk is uniform throughout the cohort. Here the personalised event-time distribution $p_i(t_0, \ldots, t_R)$ can depend on i only through an individual's covariates $\mathbf{z}_i$; there can be no informative censoring, and the crude and marginal cause-specific hazard rates and survival functions are fully identical. In heterogeneous cohorts, in contrast, the individuals have further relevant features which are not captured by the available covariates. Now a gradual 'filtering' will be observed: high-risk individuals will drop out early, causing time dependencies at cohort-level that have no counterpart at individual level. In such cohorts, it is quite possible to have uncorrelated individual-level risks, $p_i(t_0, \ldots, t_R) = \prod_r p_i(t_r)$, but to have correlated covariate-conditioned cohort-level risks, $p(t_0, \ldots, t_R|\mathbf{z}) \neq \prod_r p(t_r|\mathbf{z})$.

Table 1: Risk complexity to capture cohort heterogeneity, risk event-time correlations, and competing risks

Risk complexity		Individual	Cohort
I	Homogenous cohort No competing risks at individual level No competing risks at cohort level	$p_i(t_0,\ldots,t_R)=\prod_r p(t_r\vert\mathbf{z}_i)$	$p(t_0,\ldots,t_R\vert\mathbf{z})=\prod_r p(t_r\vert\mathbf{z})$
II	Heterogeneous cohort No competing risks at individual level No competing risks at cohort level	$p_i(t_0,\ldots,t_R)=\prod_r p_i(t_r)$	$p(t_0,\ldots,t_R\vert\mathbf{z})=\prod_r p(t_r\vert\mathbf{z})$
III	Heterogeneity-induced competing risks No competing risks at individual level Cohort level competing risks	$p_i(t_0,\ldots,t_R)=\prod_r p_i(t_r)$	$p(t_0,\ldots,t_R\vert\mathbf{z})\neq\prod_r p(t_r\vert\mathbf{z})$
IV	Heterogeneous cohort Individual level competing risks Cohort level competing risks	$p_i(t_0,\ldots,t_R)\neq\prod_r p_i(t_r)$	$p(t_0,\ldots,t_R\vert\mathbf{z})\neq\prod_r p(t_r\vert\mathbf{z})$

Risk event-time correlations in a cohort can be generated at different levels. This leads to natural hierarchy of cohorts in terms of risk complexity, as summarised in Table 1, with implications for the applicability or interpretation of survival analysis methods. Assuming statistically independent risk event-times at cohort-level is a commonly adopted, and practical, approach in studies where the interpretation of cause-specific hazards extracted from Cox regression is desirable e.g. when aiming to design clinical interventions from knowledge of the effects of modifiable covariates on a particular disease. This assumption is valid only in cohorts with risk complexity levels I and II. At level II there is still no competing risk problem, but heterogeneity may demand parametrisations of crude cohort-level primary hazard rates that are more complex than those of the Cox model. This is the rationale behind random effects models, and behind the latent class models of [24]. However, most of these approaches still only model the primary risk, and therefore cannot handle cohorts beyond level II.

In this paper we focus on developing survival analysis tools to investigate cohorts with risk complexity level III. The event-times of all risks are assumed to be statistically independent for each individual, but residual heterogeneity leads to risk correlations at cohort-level. Here, the correlations between cohort-level event-times have their origin strictly in *correlations between disease susceptibilities and covariate associations of individuals*. For example, someone with a high hazard rate for event A may also be likely to have a high hazard rate for event B, for reasons not explained by their covariates. At this level of risk complexity, heterogeneity-induced competing risks phenomena will be observed, as the risk correlations cause informative censoring. Cohorts of risk complexity level IV are the most complex, with the event-times of different risks being correlated at both individual and cohort level. Modelling tools for such cohorts are not explored in this paper.

2.3 Separating direct from indirect associations and quantifying informative censoring

The assumption of heterogeneity-induced competing risks allows us to investigate *analytically* the effects of informative censoring. With this assumption, the risk event-time marginal distributions are accessible, and it therefore becomes possible to develop expressions for the marginal cause-specific hazard rates and survival functions, in addition to their crude counterparts. The cause-specific event-time probability at the individual-level, given the assumption of risk independence, is given by $p_i(t_r)=h^i_r(t)\mathrm{e}^{-\int_0^t \mathrm{d}s\ h^i_r(s)}$. The cohort-level covariate-conditioned risk event-time marginals are therefore given by $p(t_r|\mathbf{z})=\sum_{i,\mathbf{z}_i=\mathbf{z}} h^i_r(t)\mathrm{e}^{-\int_0^t \mathrm{d}s\ h^i_r(s)}/\sum_{i,\mathbf{z}_i=\mathbf{z}}1$, and can be used to develop expressions for the marginal survival functions and hazard rates[1].

Given the assumption of heterogeneity-induced competing risks, the crude cause-specific hazard rates $h_r(t|\mathbf{z})$ and their marginal counterparts $\tilde{h}_r(t|\mathbf{z})$ are given by

$$h_r(t|\mathbf{z})=\frac{\sum_{i,\mathbf{z}_i=\mathbf{z}} h^i_r(t)\mathrm{e}^{-\sum_{r'}\int_0^t \mathrm{d}s\ h^i_{r'}(s)}}{\sum_{i,\mathbf{z}_i=\mathbf{z}}\mathrm{e}^{-\sum_{r'}\int_0^t \mathrm{d}s\ h^i_{r'}(s)}},\qquad \tilde{h}_r(t|\mathbf{z})=\frac{\sum_{i,\mathbf{z}_i=\mathbf{z}} h^i_r(t)\mathrm{e}^{-\int_0^t \mathrm{d}s\ h^i_r(s)}}{\sum_{i,\mathbf{z}_i=\mathbf{z}}\mathrm{e}^{-\int_0^t \mathrm{d}s\ h^i_r(s)}}. \tag{5}$$

[1]The marginal survival functions and hazard rates follow from $\tilde{S}_r(t|\mathbf{z})=\int_t^\infty \mathrm{d}t_r\ p(t_r|\mathbf{z})$ and $\tilde{h}_r(t|\mathbf{z})=-\frac{\mathrm{d}}{\mathrm{d}t}\log\tilde{S}_r(t|\mathbf{z})$.

Table 2: Personalised cause-specific hazard rates, for each of the R active risks, to capture latent cohort heterogeneity

$M = 1$	Heterogeneous frailties Homogeneous associations Homogeneous base hazard rates	$h_r^i(t) = \lambda_r(t)\mathrm{e}^{\beta_r^{\ell 0}+\sum_\mu \beta_r^\mu z_i^\mu}$
$M = 2$	Heterogeneous frailties Heterogeneous associations Homogeneous base hazard rates	$h_r^i(t) = \lambda_r(t)\mathrm{e}^{\beta_r^{\ell 0}+\sum_\mu \beta_r^{\ell\mu} z_i^\mu}$
$M = 3$	Heterogeneous frailties Heterogeneous associations Heterogeneous base hazard rates	$h_r^i(t) = \lambda_r^\ell(t)\mathrm{e}^{\beta_r^{\ell 0}+\sum_\mu \beta_r^{\ell\mu} z_i^\mu}$

The crude and marginal cause-specific hazard rates will generally have different values. In the marginal cause-specific hazard rate, $\tilde{h}_r(t|\mathbf{z})$, the probability that individual i survives until time t is given by $\exp[-\int_0^t \mathrm{d}s\ h_r^i(s)]$. The probability of survival until time t for individual i in the crude cause-specific hazard rate, $h_r(t|\mathbf{z})$, in contrast, depends on *all* risks. The crude cause-specific survival function $S_r(t|\mathbf{z})$, and its marginal counterpart $\tilde{S}_r(t|\mathbf{z})$, are given by

$$S_r(t|\mathbf{z}) = \mathrm{e}^{-\int_0^t \mathrm{d}s\ h_r(s|\mathbf{z})}, \qquad \tilde{S}_r(t|\mathbf{z}) = \frac{\sum_{i,\mathbf{z}_i=\mathbf{z}} \mathrm{e}^{-\int_0^t \mathrm{d}s\ h_r^i(s)}}{\sum_{i,\mathbf{z}_i=\mathbf{z}} 1}. \tag{6}$$

The marginal cause-specific survival function, $\tilde{S}_r(t|\mathbf{z})$, depends only on the risk r, whereas its crude counterpart depends on *all* risks through the crude hazard rate $h_r(t|\mathbf{z})$. We conclude that the assumption that competing risks, if present, are induced by residual cohort or disease heterogeneity, leads to relatively simple explicit formulae for the marginal cause-specific quantities of interest. It remains to identify the *minimal* level of description required for evaluating these formulae, and to determine how the required information can be estimated from survival data.

2.4 Modelling the heterogeneous cohort

In this section we develop our latent class framework. We assume a heterogeneous cohort to be comprised of L initially unknown sub-cohorts, or latent classes, labelled by $\ell = 1, \ldots, L$. Each class obeys the proportional hazards assumption, but the cohort collectively need not. Each individual from the cohort is assumed to belong to exactly one latent class. The relative influence of each class depends on the fraction of individuals belonging to that class. At the heart of our model are the personalised cause-specific hazard rates, $h_r^i(t)$, being of the Cox form and obeying the assumption of proportional hazards. These personalised hazard rates shall depend on cause-specific and possibly class-specific frailty, association, and base hazard rate parameters. Inevitably, modelling *all* risks and their correlations leads to models with more parameters than those which model *only* the primary risk. To avoid overfitting, three variants of the personalised cause-specific hazard rates are introduced, with differing degrees of heterogeneity and complexity, as summarised in Table 2.

The frailty parameters $\beta_r^{\ell 0}$ capture effects that cannot be attributed to the included covariates, within class ℓ for risk r. The association parameters $\beta_r^{\ell\mu}$ quantify how strongly each of the $\mu = 1, \ldots, P$ covariates influence the personalised hazard rate for risk r and class ℓ. The functions $\lambda_r^\ell(t)$ describe the cause-specific personalised base hazard rates, for risk r, for each individual in class ℓ. They are parametrized through a spline construction (Appendix A.6); the number K of spline time points required increases with the irregularity of the base hazard rates. Upon representing the set of those individuals belonging to class ℓ by I_ℓ, all quantities of the form $\sum_{i,\mathbf{z}_i=\mathbf{z}} f(h_r^i(t))/\sum_{i,\mathbf{z}_i=\mathbf{z}} 1$, i.e. all covariate-conditioned averages over expressions involving the personalised cause-specific hazard rates, can then be written as $\sum_\ell w_\ell(\mathbf{z}) f(h_r^\ell(t|\mathbf{z}))$, where the class membership fractions, $w_\ell(\mathbf{z})$, give the probability that a randomly drawn individual with covariates $\mathbf{z}$ will belong to class ℓ. In this paper, these fractions are chosen to be independent of the covariates; this amounts to assuming that all classes have identically distributed covariates. It is, in principle, easy to extend our model to incorporate covariate-conditioned class weightings, as in e.g. [24], though this would introduce additional model parameters.

The various crude and marginal quantities which describe a cohort are given below in terms of our latent class parametrisation for the fully heterogeneous variant of the personalised cause-specific hazard rate ($M = 3$);

corresponding expressions are easily obtained for the simpler hazard rate variants on substitution of class-independent association(s) ($M = 1$) and class-independent base hazard rate(s) ($M = 2$), as appropriate. We adopt the following compact notation. The effect of the frailty and associations, for each true risk r and class ℓ, are summarised by the product $\boldsymbol{\beta}_r^\ell \cdot \mathbf{z} = \beta_r^{\ell 0} + \sum_\mu \beta_r^{\ell\mu} z^\mu$. The end-of-trial risk ($r = 0$) is assumed to be independent of the covariates, and accordingly the class-independent associations for the censoring risk are defined to be zero, $\boldsymbol{\beta}_0 = 0$. The integrated base hazard rates are denoted by $\Lambda_r^\ell(t) = \int_0^t \mathrm{d}s\ \lambda_r^\ell(s)$. The parametrised crude and marginal hazard rates, for risks $r = 1, \ldots, R$, are now given by,

$$h_r(t|\mathbf{z}) = \frac{\sum_\ell w_\ell\ \lambda_r^\ell(t) \mathrm{e}^{\boldsymbol{\beta}_r^\ell \cdot \mathbf{z} - \sum_{r'=1}^R \exp(\boldsymbol{\beta}_{r'}^\ell \cdot \mathbf{z}) \Lambda_{r'}^\ell(t)}}{\sum_\ell w_\ell\ \mathrm{e}^{-\sum_{r'=1}^R \exp(\boldsymbol{\beta}_{r'}^\ell \cdot \mathbf{z}) \Lambda_{r'}^\ell(t)}}, \qquad \tilde{h}_r(t|\mathbf{z}) = \frac{\sum_\ell w_\ell\ \lambda_r^\ell(t) \mathrm{e}^{\boldsymbol{\beta}_r^\ell \cdot \mathbf{z} - \exp(\hat{\boldsymbol{\beta}}_r^\ell \cdot \mathbf{z}) \Lambda_r^\ell(t)}}{\sum_\ell w_\ell\ \mathrm{e}^{-\exp(\boldsymbol{\beta}_r^\ell \cdot \mathbf{z}) \Lambda_r^\ell(t)}}. \tag{7}$$

The corresponding crude and marginal cause-specific survival functions are given by,

$$S_r(t|\mathbf{z}) = \exp\left(-\int_0^t \mathrm{d}s\ h_r(s|\mathbf{z})\right), \qquad \tilde{S}_r(t|\mathbf{z}) = \sum_\ell w_\ell\ \mathrm{e}^{-\exp(\boldsymbol{\beta}_r^\ell \cdot \mathbf{z}) \Lambda_r^\ell(t)}. \tag{8}$$

The crude cause-specific hazard rate is influenced by *all* risks, hence the same is true for the crude cause-specific survival function. In Appendix A.2 the crude and marginal survival are shown to be identical in the case where there is one risk only, i.e. in the absence of any potential informative censoring. Our latent class model leads in the same manner to an intuitive and easily interpreted formulation of the cause-specific cumulative incidence function, in which the role of all model parameters is completely transparent:

$$F_r(t|\mathbf{z}) = \int_0^t \mathrm{d}t'\ \mathrm{e}^{-\Lambda_0(t')} \sum_\ell w_\ell\ \lambda_r^\ell(t') \mathrm{e}^{\boldsymbol{\beta}_r^\ell \cdot \mathbf{z} - \sum_{r'=1}^R \exp(\boldsymbol{\beta}_{r'}^\ell \cdot \mathbf{z}) \Lambda_{r'}^\ell(t')}. \tag{9}$$

Finally, our approach also offers *retrospective* determination of an individual's class membership probability, given their covariates $\mathbf{z}$ and their survival information (t, r). Following Bayesian arguments (as detailed in Appendix A.3), the probability that an individual belongs to class ℓ, given their covariates and survival information, is found to be

$$p(\ell|t, r, \mathbf{z}) = \frac{w_\ell\ \lambda_r^\ell(t) \mathrm{e}^{\boldsymbol{\beta}_r^\ell \cdot \mathbf{z} - \sum_{r'=1}^R \exp\left(\boldsymbol{\beta}_{r'}^\ell \cdot \mathbf{z}\right) \Lambda_{r'}^\ell(t)}}{\sum_{\ell'=1}^L w_{\ell'}\ \lambda_r^{\ell'}(t) \mathrm{e}^{\boldsymbol{\beta}_r^{\ell'} \cdot \mathbf{z} - \sum_{r'=1}^R \exp\left(\boldsymbol{\beta}_{r'}^{\ell'} \cdot \mathbf{z}\right) \Lambda_{r'}^{\ell'}(t)}}. \tag{10}$$

We define the retrospective class weight f_ℓ of a class as the fraction of the cohort for which, according to (10), ℓ is the most probable class, i.e. the fraction of individuals i for which $\mathrm{argmax}_{\ell'=1\ldots L}\, p(\ell'|t_i, r_i, \mathbf{z}_i) = \ell$. Note that one will generally find that $(f_1, \ldots f_L) \neq (w_1, \ldots, w_L)$. The search for informative new covariates could be aided by this retrospective class assignment, and thereby increase our ability to predict personalised risk in heterogeneous cohorts. Such new covariates are expected to be features that patients who belong to the same class, as suggested by (10), have in common.

2.5 Characterisation of cohort heterogeneity

The most appropriate choices for the personalised hazard rate model, the number of latent classes, and the complexity of the base hazard rates are all unknown at the outset of an analysis, so their optimal values and the corresponding optimal class weightings, frailties, associations, and base hazard rates all need to be estimated. For this we use Bayesian inference.

The term 'model' shall be used to describe the combination of a particular number of latent classes, L, a particular form of the base hazard rates having K spline time points, and a particular parametrisation of the personalised hazard rates, M, and shall be denoted by $\mathcal{H}_{KLM}$. The parameter vector, $\boldsymbol{\theta}_{\mathcal{H}_{KLM}}$, denotes the model parameters (i.e. the class weightings, the frailty, association, and base hazard rate parameters) of model $\mathcal{H}_{KLM}$. The optimal description of a cohort, $\boldsymbol{\theta}^\star_{\mathcal{H}^\star_{KLM}}$, is identified by finding the most probable parameter values, $\boldsymbol{\theta}^\star_{\mathcal{H}_{KLM}}$, of the most probable model, $\mathcal{H}^\star_{KLM}$. In practice, the optimal parameter values are *first* determined for each model from an ensemble of candidate models *before* the model supported by the greatest evidence is identified. The optimal parameter values are those which maximise the posterior distribution, $p(\boldsymbol{\theta}_{\mathcal{H}_{KLM}}|D)$, which gives the probability of the model parameters, $\boldsymbol{\theta}_{\mathcal{H}_{KLM}}$, conditioned on a cohort's survival data, D. The probability density, $p(t, r|\mathbf{z})$, to find the earliest event occurring at time t, and

corresponding to risk r, provides the link between our latent class model parameters and the survival data. In terms of our latent class parametrisation, the probability density for an individual with covariate vector $\mathbf{z}$ to report (t, r) is given by,

$$p(t, r|\mathbf{z}) = \mathrm{e}^{-\Lambda_0(t)} \sum_\ell w_\ell \; \lambda_r^\ell(t) \mathrm{e}^{\boldsymbol{\beta}_r^\ell \cdot \mathbf{z} - \sum_{r'=1}^R \exp(\boldsymbol{\beta}_{r'}^\ell \cdot \mathbf{z}) \Lambda_{r'}^\ell(t)}. \tag{11}$$

Here the parameters are those of the fully-heterogeneous personalised hazard rate ($M = 3$) defined in Table 2. The data likelihood, $p(D|\boldsymbol{\theta}_{\mathcal{H}_{KLM}}) = \prod_i p(t_i, r_i|\mathbf{z}_i, \boldsymbol{\theta}_{\mathcal{H}_{KLM}})$, gives the joint probability of the event-time t_i and risk r_i for every individual in the cohort, conditioned on their covariates, $\mathbf{z}_i$, and the model parameters, $\boldsymbol{\theta}_{\mathcal{H}_{KLM}}$, for the model $\mathcal{H}_{KLM}$.

The posterior can be written as $p(\boldsymbol{\theta}_{\mathcal{H}_{KLM}}|D) = Z_{\mathcal{H}_{KLM}}^{-1} \exp(\mathcal{L}(\boldsymbol{\theta}_{\mathcal{H}_{KLM}}, D))$, where the log-likelihood is defined as $\mathcal{L}(\boldsymbol{\theta}_{\mathcal{H}_{KLM}}, D) = \log(p(D|\boldsymbol{\theta}_{\mathcal{H}_{KLM}}) p(\boldsymbol{\theta}_{\mathcal{H}_{KLM}}))$, the normalisation constant is denoted by $Z_{\mathcal{H}_{KLM}}$, and the prior distribution over the model parameters is given by $p(\boldsymbol{\theta}_{\mathcal{H}_{KLM}})$ (Appendix A.4). The contribution to the log-likelihood from the data-likelihood, for our fully heterogeneous personalised hazard rate ($M = 3$) latent class model, is given by,

$$\log p(D|\boldsymbol{\theta}_{\mathcal{H}_{KLM}}) = -\sum_i \Lambda_0(t_i) + \sum_i \log \Big[\sum_\ell w_\ell \; \lambda_{r_i}^\ell(t_i) \mathrm{e}^{\boldsymbol{\beta}_{r_i}^\ell \cdot \mathbf{z}_i - \sum_{r=1}^R \Lambda_r^\ell(t_i) \exp(\boldsymbol{\beta}_r^\ell \cdot \mathbf{z}_i)} \Big]. \tag{12}$$

The optimal model, $\mathcal{H}_{KLM}^\star$ is that which is supported by the greatest 'evidence', $p(\mathcal{H}_{KLM}|D)$. In this study all models are given identical prior probabilities, $p(\mathcal{H}_{KLM})$. The model evidence describes the likelihood of the model $\mathcal{H}_{KLM}$ conditioned on the cohort's survival data, D, and is proportional to the posterior normalisation constant, $Z_{\mathcal{H}_{KLM}} = \int \mathrm{d}\boldsymbol{\theta}_{\mathcal{H}_{KLM}} \; p(D|\boldsymbol{\theta}_{\mathcal{H}_{KLM}}, \mathcal{H}_{KLM}) p(\boldsymbol{\theta}_{\mathcal{H}_{KLM}}|\mathcal{H}_{KLM})$, for that model. Here we determined the evidence for each model from its optimal parameter values, $\boldsymbol{\theta}_{\mathcal{H}_{KLM}}^\star$, through the use of Gaussian approximations to the posteriors (Appendix A.5).

2.6 Practical tools for survival analysis

The analysis protocol described above has been implemented in our software package, *ALPACA* (Advanced Latent Class Prediction And Competing Risk Analysis), using the C programming language. The frailty, association, and base hazard rate parameters are located, for each model, by maximum a posteriori (MAP) estimation, using a stochastic refinement of the downhill simplex method [39]. Numerical estimation of the curvature of the posterior distribution around the location of maximum probability enables both the error bars for the parameter estimates and the model evidence to be determined. As the search for the optimal parameter values is achieved using a stochastic optimization algorithm, this procedure is typically performed multiple times and the best overall estimation is selected.

A cohort's survival data is pre-processed before application of our analysis algorithms. This involves linear rescaling of the *raw* covariate values such that the *transformed* counterparts are described by zero average and unit variance distributions, equivalent to the definition of Z-scores. The parameter estimates obtained using our analysis are translated to the more familiar language of hazard ratios (HR), 95% confidence intervals (CI), and p-values. The hazard ratio, $\mathrm{HR}_r^{\ell\mu}$, the lower and upper bounds, ${}^-\mathrm{HR}_r^{\ell\mu}$ and ${}^+\mathrm{HR}_r^{\ell\mu}$ respectively, of the 95% confidence interval, $\mathrm{CI}_r^{\ell\mu} = [{}^-\mathrm{HR}_r^{\ell\mu}, {}^+\mathrm{HR}_r^{\ell\mu}]$, and the p-value associated with normalised covariate μ, can be computed, for every event type r and each class ℓ, according to

$$\mathrm{HR}_r^{\ell\mu} = \mathrm{e}^{2\beta_r^{\ell\mu}}, \qquad {}^\pm\mathrm{HR}_r^{\ell\mu} = \exp(2(\beta_r^{\ell\mu} \pm 1.96\sigma_r^{\ell\mu})), \qquad p_r^{\ell\mu} = 1 - \mathrm{erf}(|\beta_r^{\ell\mu}|/\sqrt{2}\sigma_r^{\ell\mu}). \tag{13}$$

Here $\beta_r^{\ell\mu}$ and $\sigma_r^{\ell\mu}$ are the estimated association parameter and its uncertainty for covariate μ, $|\beta_r^{\ell\mu}|$ denotes the magnitude of the estimated association, and the error integral is given by $\mathrm{erf}(t) = (2/\sqrt{\pi}) \int_0^t \mathrm{d}x \; \mathrm{e}^{-x^2}$. As a consequence of our data pre-processing, the magnitudes of the effects for different associations can be directly compared.

Crude and marginal survival curves, cumulative incidence curves, and retrospective class allocations are known once the model parameters have been estimated. Any differences between crude and marginal (covariate-conditioned) survival curves are indicative of the extent to which competing risks are present in a cohort subject to heterogeneity-induced informative censoring. False protectivity effects should be suspected if the crude survival function exceeds the marginal survival function; the opposite would suggest the influence of false exposure. Class-specific marginal survival may offer valuable insight into the expected progression of

individuals belonging to different latent classes. Expressing the marginal cause-specific survival (8), $\tilde{S}_r(t|\mathbf{z})$, as the weighted sum of cause- and class-specific survival, $\tilde{S}_r^\ell(t|\mathbf{z})$, as given by,

$$\tilde{S}_r^\ell(t|\mathbf{z}) = \mathrm{e}^{-\exp(\boldsymbol{\beta}_r^\ell \cdot \mathbf{z})\Lambda_r^\ell(t)}, \tag{14}$$

allows the class-specific marginal survival for each of the latent classes to be compared. Similarly, the cause-specific cumulative incidence function (9) can be expressed as a weighted sum of its class-specific components, $F_r^\ell(t|\mathbf{z})$, given by,

$$F_r^\ell(t|\mathbf{z}) = \int_0^t \mathrm{d}t' \; \mathrm{e}^{-\hat{\Lambda}_0(t')} \; \hat{\lambda}_r^\ell(t') \mathrm{e}^{\hat{\boldsymbol{\beta}}_r^\ell \cdot \mathbf{z} - \sum_{r'=1}^R \exp(\hat{\boldsymbol{\beta}}_{r'}^\ell \cdot \mathbf{z}) \hat{\Lambda}_{r'}^\ell(t')}, \tag{15}$$

such that $F_r(t|\mathbf{z}) = \sum_\ell w_\ell F_r^\ell(t|\mathbf{z})$. The class- and cause-specific cumulative incidence provides information about the relative occurrence of events for each latent class and each cause at any time.

Retrospective class assignment, by identification of the most probable latent class to which an individual belongs using (10), can offer additional insight into differences between latent classes. Retrospectively allocated class-conditioned time-to-event distributions may offer clues as to the expected survival time for members of the different classes. The detection of differences between covariate distributions for those individuals retrospectively assigned to different latent classes offers a potentially powerful means of identifying novel informative covariates.

3 Application to simulated survival data

In this section we present the results of applying our algorithms to simulated data (see Appendix B), modelling a variety of conditions. The effectiveness of our algorithm to characterise a heterogeneous cohort with three latent classes, two of which differ only in their base hazard rates, is demonstrated in Section 3.1. In Section 3.2 we test the ability of our method to accurately characterise a cohort in the presence of heterogeneity-induced informative censoring. The model space searched for the analysis of each simulated cohort covered *all* combinations of latent classes between one and four (i.e. $L = 1 - 4$), base hazard rate complexities between one and eight (i.e. $K = 1 - 8$), and all variants of the personalised hazard rate (i.e. $M = 1, 2, 3$). Parameter estimates were obtained for each model at least five times; the model with the greatest overall evidence was selected for cohort estimation.

Since the class membership of every individual is known when using simulated data, the accuracy of the retrospective class allocation algorithm (10) can be measured as the fraction of correctly assigned individuals. The performance of (10) will, of course, depend on the quantitative features of a cohort; it is reasonable to expect that retrospective class assignment should perform most effectively when the differences between latent classes are pronounced, and that the algorithm may struggle when the latent classes do not differ significantly. It is important to note that, even if all parameters were known exactly, due to the stochasticity of event times, class allocation will never be perfectly accurate[2].

3.1 Revealing heterogeneity and cohort sub-structure

Our analysis is first applied to simulated data, modelling a heterogeneous cohort having three latent classes ($L = 3$), but free of informative censoring. The characteristics of *Cohort A* are given in Table 3; the three latent classes are of equal size, each individual i has three covariates (z_i^1, z_i^2, z_i^3) and is subject to either one real risk or end-of-trial censoring at time $t = 20$. The personalised cause-specific hazard rates of individuals in *Cohort A* have class-dependent associations and base hazard rates (i.e. $M = 3$). The associations of individuals belonging to classes $\ell = 1$ and $\ell = 2$ are identical; these classes differ only in that the base hazard rate of class $\ell = 1$ is time-independent whereas that of class $\ell = 2$ increases exponentially with time. The base hazard rate of class $\ell = 3$ is time-independent, though it differs from that of class $\ell = 1$.

The estimation of such a cohort is challenging even in the absence of informative censoring; to successfully characterise *Cohort A*, any method must be able to identify the three latent classes, distinguishing between two classes which differ only in their base hazard rates. The effectiveness of our analysis to characterise *Cohort*

[2] A fundamental limit on the accuracy of retrospective allocation can be determined using the retrospective allocation and event time probability relations under the assumption of an infinitely large cohort from which the parameters have been extracted perfectly. For a cohort subject to one risk, having two latent classes of equal size, a class-independent constant base hazard rate and only one covariate, the best possible allocation quality in the case where $\beta_1^{11} = -\beta_1^{21} = 2$ is about 83%.

Table 3: Modelling a heterogeneous cohort: The parameters used to generate simulated data modelling a cohort having three latent classes, one real risk, and with end-of-trial censoring at time $t = 20$.

Cohort A: ($L = 3, M = 3$)
Heterogeneity and class-dependent base hazard rates

	Class, $\ell = 1$	Class, $\ell = 2$	Class, $\ell = 3$
w_ℓ	1/3	1/3	1/3
$\lambda_1^\ell(t)e^{\beta_1^{\ell 0}}$	3/10	$e^{t/4}/100$	1/10
$\beta_1^{\ell 1}$	2	2	-2
$\beta_1^{\ell 2}$	0	0	0
$\beta_1^{\ell 3}$	0	0	0

A is summarised in Figure 2, which shows the estimated associations and base hazard rates for cohort sizes of $N = 20000$, $N = 2000$, and $N = 200$. The association parameters were accurately estimated for cohort sizes of $N = 2000$ and $N = 20000$, and the estimated and true base hazard rates are sufficiently close, for each of the classes, for the model supported by the greatest evidence to be considered a good estimation of the *true* cohort structure. Although it is clear that a sample size of $N = 200$ is insufficient for *Cohort A* to be accurately estimated, it is noteworthy that, despite such a meagre sample size, our analysis correctly reports the cohort to be comprised of three latent classes, and the estimated associations are accurate (within the estimated uncertainty) for covariates 2 and 3 for all three classes, and are accurate for classes 1 and 3 for covariate 1.

The optimal models reported for data set sizes $N = 20000$ and $N = 2000$ were $\mathcal{H}^\star_{KLM} = (K=4, L=3, M=3)$ and $\mathcal{H}^\star_{KLM} = (K = 3, L = 3, M = 3)$, respectively. In total, 23 parameters (two weights, and for each of the three latent classes one frailty, three associations, and three modifiable spline points for the parametrised base hazard rate approximation) were estimated for the optimal model for the $N = 20000$ data set. As the Bayesian model selection stage of our analysis balances complexity of a particular model against the evidence for it in the survival data, it is unsurprising that models containing fewer parameters were found to be optimal for the $N = 2000$ and $N = 200$ data sets, having 20 and 16 parameters respectively. Ultimately, there is not enough information available in the $N = 200$ data set to justify the selection of a model having the required complexity to accurately describe *Cohort A* and here the optimal model $\mathcal{H}^\star_{KLM} = (K = 3, L = 3, M = 2)$ is reported, which has class-independent base hazard rates.

Our analysis correctly indicates that in *Cohort A* only covariate 1 has a statistically significant association with hazard for the primary risk. The estimates for each of the three classes for the $N = 2000$ data set are (within the estimated uncertainty) in agreement with the true values ($\ell = 1$: $\beta_1^{11} = 1.90 \pm 0.12$, HR=46, 95% CI=[29,72], $p < 10^{-7}$; $\ell = 2$: $\beta_1^{21} = 1.91 \pm 0.11$, HR=45, 95% CI=[28,70], $p < 10^{-7}$; $\ell = 3$: $\beta_1^{31} = -2.16 \pm 0.14$, HR=0.013, 95% CI=[0.008,0.023], $p < 10^{-7}$). The associations for covariates 2 and 3 were correctly estimated for all three data sets, and found to be statistically insignificant (according to p-value) in all three latent classes, and the 95% CI included the true HR of unity in all cases. As heterogeneity is not accounted for in a standard Cox regression, it should be expected that the application of such an analysis to *Cohort A* would be likely to yield incorrect estimates. Indeed, Cox regression produces an erroneous interpretation of the *Cohort A* data with $N = 2000$, indicating that covariates 1 and 2 are both associated with an increased hazard (cov.1: HR=1.33, 95% CI=[1.21,1.47], $p < 10^{-7}$; cov. 2: HR=1.12, 95% CI=[1.01,1.23], p=0.03).

The accuracy of the estimated base hazard rates can be assessed both by visual inspection, for similarity between the estimated and true rates (Figure 2), and numerically. The time-independent base hazard rate of class $\ell = 3$ is reliably estimated for both the $N = 20000$ and $N = 2000$ data sets. The base hazard rates of classes $\ell = 1$ and $\ell = 2$ are more accurately estimated for the larger data set size; the estimated base hazard rate of class $\ell = 1$ deviates from the true rate for times greater than about $t = 10$ for the $N = 2000$ analysis, whereas deviation is not observed at times earlier than about $t = 15$ for the $N = 20000$ set. The suitability of a spline-approximation to describe the exponentially increasing base hazard rate of the class $\ell = 2$ can also be gauged by comparing the known values of the parameters, λ_0 and α, obtained on fitting the generalised form of the exponentially increasing base hazard rate, $\lambda(t) = \lambda_0 e^{\alpha t}$, to the estimated base hazard rate; for the $N = 20000$ data set these parameters were determined to be $\lambda_0 = 0.012$ and $\alpha = 0.233$ and for the $N = 2000$ data set they were found to be $\lambda_0 = 0.014$ and $\alpha = 0.227$ (the true values of these parameters are $\lambda_0 = 0.01$ and $\alpha = 0.25$).

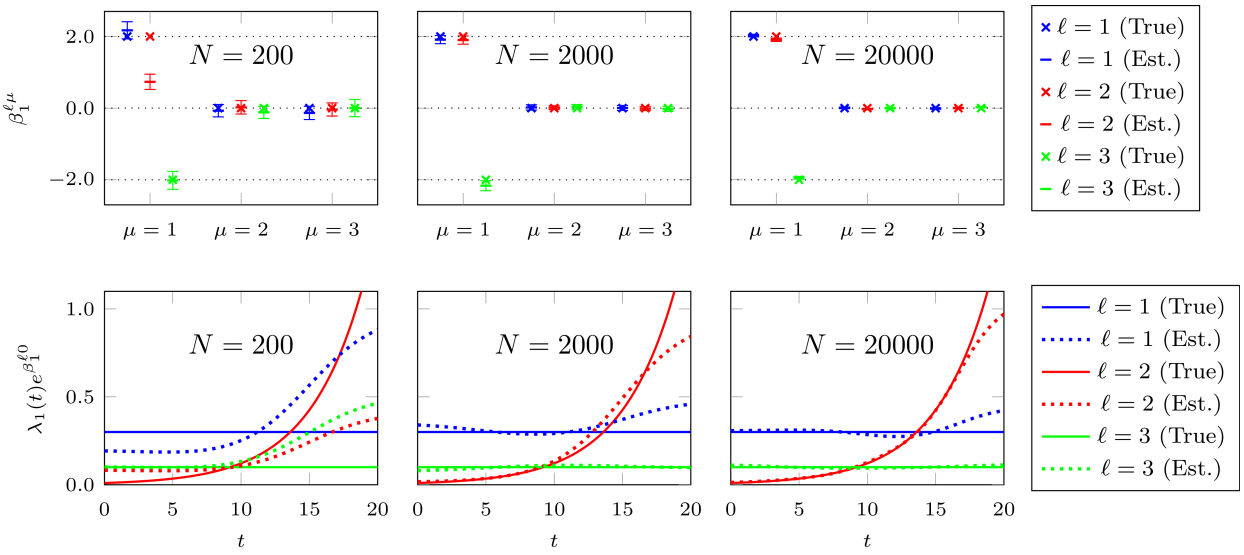

Figure 2: *Bayesian estimation of a simulated heterogeneous cohort:* The estimated associations and base hazard rates for analysis of *Cohort A* data, as specified in Table 3, are shown along with their true values for sample sizes $N = 200$, $N = 2000$, and $N = 20000$. Our method correctly identifies three latent classes, even for the meagre sample size of $N = 200$. The association parameters are accurately estimated for all covariates and all classes on analysis of *Cohort A* data of size $N = 2000$; the base hazard rates deviate only minimally from their true value at times greater than about $t = 10$. The precision of the estimated associations is superior for the data set of size $N = 20000$, and the estimated base hazard rates are true to their actual values even beyond time $t = 15$.

Random assignment of patients to the classes of *Cohort A* would give an accuracy of 33%. In contrast, on applying our class assignment algorithm to the $N = 20000$ data set, 73% of individuals were allocated correctly; 72% and 66% of those allocated to classes $\ell = 1$ and $\ell = 2$ actually belonged to those classes, and 84% of those allocated to class $\ell = 3$ actually belonged to that class. Class assignment applied to the $N = 2000$ data set yielded a similar accuracy for those assigned to classes $\ell = 1$ and $\ell = 2$, as was obtained with the $N = 20000$ data set; 78% of those allocated to class $\ell = 3$ actually belonged to that class. Despite there being insufficient information in the $N = 200$ data set for the associations and base hazard rates to be accurately estimated, overall 70% of individuals were still correctly allocated to their true class.

3.2 Effective analysis in the presence of heterogeneity-induced informative censoring

Next we examine the ability of our analysis to accurately model survival in the presence of heterogeneity-induced informative censoring. We analyse two simulated cohorts, which are identical in their primary risk characteristics; one cohort has been modelled to simulate the effects of false protectivity whilst the other includes the influence of false-aetiology. *Cohort B* and *Cohort C* both have two latent classes ($L = 2$), of equal size, and personalised cause-specific hazard rates with class-dependent associations (i.e. of type $M = 2$); both are subject to two real risks, the primary risk ($r = 1$) and a competing risk ($r = 2$), and end-of-trial censoring. The base hazard rates are time-independent but differ between the two risks. The quantitative characteristics of the simulated cohorts are given in Table 4.

The Bayesian-determined optimal model was found to be accurate for both cohorts on analysis of data with $N = 1500$ observations; the estimated class sizes, associations, and base hazard rates were in close agreement with their true values for both data sets (Table 4). The estimates for both *Cohort B* and *Cohort C* suggest correctly that covariate 1 is strongly associated with an increased hazard for the primary risk in one class, and reduced hazard in the other class. Covariates 2 and 3 were found to be statistically insignificant, according to their respective p-values, for both classes in both cohorts. The estimates for the secondary risk, responsible for informative censoring, are also consistent with the true values for both cohorts. Covariate 1 was found to be associated with an increased hazard for the secondary risk in class $\ell = 1$ only in *Cohort B*, and with a

Table 4: Modelling heterogeneity-induced informative censoring: The estimated weights, w_ℓ, and association parameters, $\beta_r^{\ell\mu}$, for the two real risks $r = 1$ and $r = 2$, from the Bayesian-determined optimal model from analysis of the *Cohort B* and *Cohort C* data sets with $N = 1500$. True parameter values are shown in brackets to the right of the estimated values.

Cohort B $(K = 1, L = 2, M = 2)$
Heterogeneity-induced false protectivity

	Class, $\ell = 1$		Class, $\ell = 2$	
w_ℓ	0.51 ± 0.02	(0.5)	0.49 ± 0.02	(0.5)
$\beta_1^{\ell 1}$	1.85 ± 0.14	(2.0)	-1.97 ± 0.10	(-2.0)
$\beta_1^{\ell 2}$	-0.05 ± 0.10	(0.0)	-0.04 ± 0.07	(0.0)
$\beta_1^{\ell 3}$	0.15 ± 0.10	(0.0)	0.01 ± 0.09	(0.0)
$\beta_2^{\ell 1}$	3.17 ± 0.10	(3.0)	-0.10 ± 0.07	(0.0)
$\beta_2^{\ell 2}$	-0.07 ± 0.08	(0.0)	0.03 ± 0.06	(0.0)
$\beta_2^{\ell 3}$	-0.06 ± 0.07	(0.0)	0.01 ± 0.06	(0.0)

Cohort C $(K = 1, L = 2, M = 2)$
Heterogeneity-induced false aetiology

	Class, $\ell = 1$		Class, $\ell = 2$	
w_ℓ	0.51 ± 0.02	(0.5)	0.49 ± 0.02	(0.5)
$\beta_1^{\ell 1}$	2.04 ± 0.12	(2.0)	-1.94 ± 0.10	(-2.0)
$\beta_1^{\ell 2}$	-0.06 ± 0.07	(0.0)	-0.01 ± 0.06	(0.0)
$\beta_1^{\ell 3}$	-0.02 ± 0.08	(0.0)	0.14 ± 0.06	(0.0)
$\beta_2^{\ell 1}$	-3.02 ± 0.09	(-3.0)	0.11 ± 0.08	(0.0)
$\beta_2^{\ell 2}$	-0.02 ± 0.06	(0.0)	0.05 ± 0.07	(0.0)
$\beta_2^{\ell 3}$	0.04 ± 0.07	(0.0)	-0.03 ± 0.06	(0.0)

reduced hazard for the secondary risk for class $\ell = 1$ only in *Cohort C*. The associations for covariates 2 and 3 were found to be statistically insignificant.

Survival estimators are shown in Figure 3 for both cohorts. False protectivity is clearly indicated in *Cohort B*, as the crude survival function for the primary risk, S_1, exceeds its marginal counterpart, $\tilde{S}_1$, throughout the trial duration. The opposite is true for *Cohort C*, exposing false aetiology effects (Figure 3 (a,e)). In the presence of informative censoring, a naive interpretation of the Kaplan-Meier estimators for the primary risk, $S_1^{\rm KM}$, conditioned on the value of covariate 1, would be highly misleading, yielding extremely poor estimates for the lower and upper quartiles of covariate 1 due to risk correlations in the cohorts (Figure 3 (b,f)). For example, in *Cohort B* the KM estimator suggests that individuals having an upper quartile value of covariate 1 are likely to have a relatively good survival against the primary risk, with about 90% surviving at the end-of-trial time, whereas almost 50% of such individuals actually survive until this time in both cohorts; in *Cohort C* the KM estimator for individuals having a lower quartile value of covariate 1 suggests no survival beyond about half of the trial duration. It is reassuring that the marginal class-specific survival functions against the primary risk, $\tilde{S}_1^1$ and $\tilde{S}_1^2$, are similar for both cohorts, given that their primary risk characteristics were indeed identical (Figure 3 (c,g)). Our analysis correctly estimated the *direct* (or 'decontaminated') associations for the primary risk as being identical, within the uncertainty of the parameter estimates, both in the presence of heterogeneity-induced false protectivity and heterogeneity-induced false aetiology. The weighted class-specific cumulative incidence are also shown (Figure 3 (d,h)).

4 Applications to prostate cancer data

Prostate cancer (PC) data are notorious for exhibiting competing risk effects [40], largely due to the fact that the disease occurs late in life when there is an increased number of non-primary events whose incidence could correlate with prostate cancer. Here we analyse data from the ULSAM cohort [35] and compare the outcomes of Cox's proportional hazards regression [11] and our present method. The ULSAM cohort has $N = 2047$ individuals of which 208 reported PC as the first event, as described previously [36], and we have included five relevant covariates of the ULSAM data [35].

Smoking is suggested to have a weak protective effect against PC risk (HR=0.85, 95% CI=[0.65,1.11], p=0.23) in the ULSAM cohort according to Cox's proportional hazards analysis. However, as the Kaplan-Meier estimator for PC risk stratified on smoking status (see Figure 1) does not meet the proportional hazards assumption, any interpretations drawn from such an analysis (on the ULSAM cohort data as a whole, at least) may be invalid. The non-proportionality of the smoking covariate-conditioned Kaplan-Meier estimator for PC risk could be due to heterogeneity in the ULSAM cohort.

Analysis of the ULSAM cohort using our method suggests that the cohort should rather be viewed as consisting of two distinct classes: one class ($\ell = 1$) with relatively frail individuals (in terms of both primary and secondary risk) which contains about 16% of the cohort according to retrospective class allocation, and another class ($\ell = 2$) with rather healthy individuals which contains the remainder of the cohort. The estimated association parameters, base hazard rates, and class-specific survival functions for PC are shown in

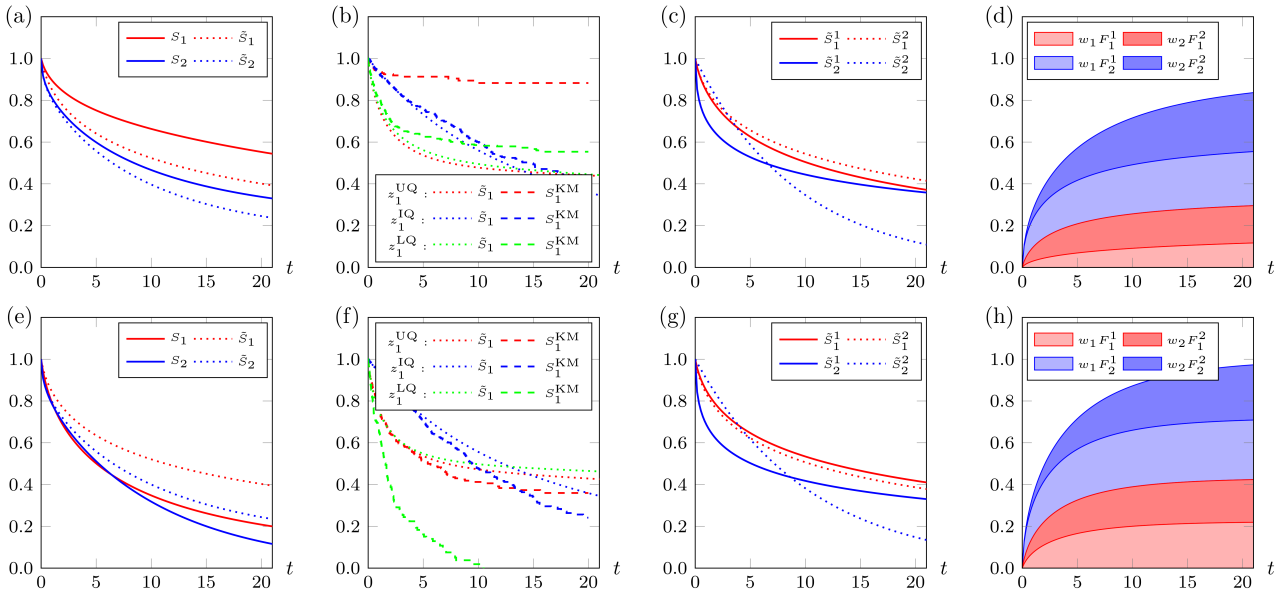

Figure 3: *Decontaminating heterogeneity-induced false protectivity and false aetiology effects from survival data:* Survival and cumulative incidence curves from the analysis of *Cohort B* and *Cohort C* (Table 4) simulated survival data, with $N = 1500$, demonstrating the effectiveness of our method to expose the influence of heterogeneity-induced informative censoring. For *Cohort B*: in (a) the crude and marginal risk-specific survival curves, $S_r(t)$ and $\tilde{S}_r(t)$, for the primary and secondary risks; in (b) the marginal survival curves, $\tilde{S}_1$, and the risk-specific Kaplan-Meier estimators, S_1^{KM}, for the lower quartile (LQ), upper quartile (UQ), and the inter-quartile range (IQ) of covariate 1; in (c) the class- and risk-specific marginal survival curves, $\tilde{S}_r^{\ell}$, for the two latent classes for both risks; in (d) the stacked weighted class-specific cumulative incidence. The same estimators are shown for *Cohort C* in (e), (f), (g) and (h) respectively.

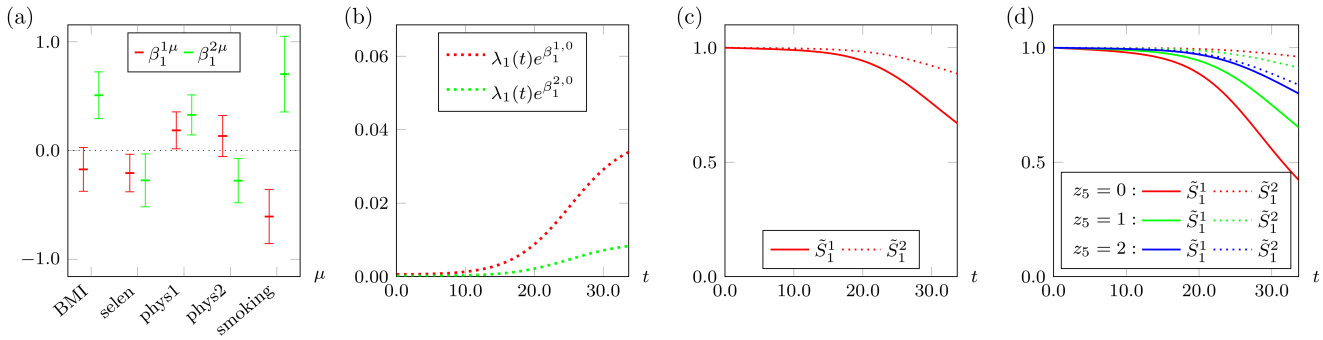

Figure 4: *Bayesian-determined heterogeneous characterisation of the ULSAM cohort:* In (a) and (b) the associations and frailty weighted-base hazard rates for PC risk, for the two latent classes of the optimal model $\mathcal{H}^{\star}_{KLM} = (K = 3, L = 2, M = 2)$ obtained on analysis of the ULSAM data. The included covariates are: body mass index (real-valued), serum Selenium level (*selen*, integer valued), leisure time physical activity (*phys1*, discrete levels 0/1/2), work physical activity (*phys2*, discrete levels 0/1/2), and smoking status (*smoking*, discrete levels 0/1/2). The estimated weighting of the classes for this model is $w_1 = 0.32 \pm 0.08$, $w_2 = 0.68 \pm 0.08$. The retrospective weighting of the classes for this model is $f_1 = 0.16$ (332 of 2047 patients), $f_2 = 0.84$ (1715 of 2047 patients). In (c) the class-and risk-specific survival functions; for PC risk it is clear that the survival of members of the frailer class ($\ell = 1$) is poorer than those of the healthier class ($\ell = 2$), significantly and increasingly so for times beyond about 20 years. In (d) the class-and risk-specific survival function conditioned on the three values of the smoking covariate ($z_5 = 0$: non-smoker, $z_5 = 1$: ex-smoker, $z_5 = 2$: smoker). For the majority of the cohort ($\ell = 2$) smoking is associated with poorer survival outcomes, but for the members of the frailer class ($\ell = 1$) smokers and ex-smokers have greater survival than non-smokers.

Fig. 4 for the most probable model $\mathcal{H}^{\star}_{KLM} = (K=3, L=2, M=2)$. The marginal survival function for PC was found to be slightly less than the crude survival function for follow-up times exceeding about 20 years; this is indicative of false protectivity effects for PC risk in the ULSAM cohort data. In the Bayesian-determined two-class description of the ULSAM cohort, BMI and smoking are recognised as serious PC risk factors for those individuals in the healthier class. In this class ($\ell = 2$), smoking is associated with elevated risk of PC (HR=4.08, 95% CI=[1.09,15.29], p=0.04), as is having a higher BMI (HR=2.77, 95% CI=[1.20,6.36], p=0.02). Conversely, in the frailer class ($\ell = 1$), smoking is associated with a decreased risk of PC (HR=0.30, 95% CI=[0.12,0.76], p=0.01). In the frail class the regression coefficients are weaker than those of the stronger class, and one expects the negative coefficients for e.g. BMI and smoking to reflect reverse causality: within this group, having a higher BMI and still being *able* to smoke may well be an indicator of *relatively* good health.

Our current explanation and interpretation of the ULSAM data is not necessarily the final one. There are always alternative ways to do the regression. The conclusion to be drawn is that the new two-class explanation of the ULSAM data is both probabilistically and intuitively more plausible than the one provided by the Cox model.

5 Applications to breast cancer data

Here we summarise the results of applying our analysis to data from the Swedish Apolipoprotein Mortality Risk Study (AMORIS), see e.g. [37, 38]. Our data set, for which competing risk analysis was presented in [41], describes $N = 1798$ women from the AMORIS population, for whom baseline serum glucose, triglyceride, and total cholesterol measurements were available within three months to three years prior to breast cancer diagnosis. The event time is the duration between diagnosis and breast cancer death (BC), cardiovascular disease death (CV), death from other causes, or departure from or the end of the study (censoring). Age, fasting status, and socio-economic status data were also included as covariates.

Our analysis suggests that the cohort is best described by three latent classes, sharing the same trend of having base hazard rates which decrease with time for BC death risk, but increase with time for CV death risk. The largest class ($\ell = 1$) contains about 66% of the cohort according to retrospective class allocation, and has a poorer survival against BC death but a greater survival against CV death than does the second class ($\ell = 2$),

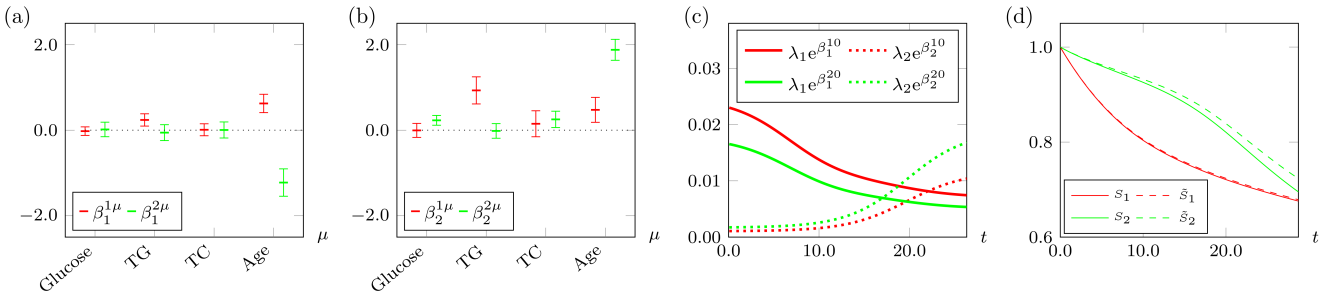

Figure 5: *Bayesian-determined heterogeneous characterisation of a cohort of women from the AMORIS population diagnosed with breast cancer:* Parameter estimates for the Bayesian-determined optimal model $\mathcal{H}^{\star}_{KLM} = (K=3, L=3, M=2)$, obtained on analysis of data for $N = 1798$ women from the AMORIS population diagnosed with breast cancer; the estimated retrospective weighting of the classes for this model is $f_1 = 0.66$ (1189 of 1798 patients), $f_2 = 0.32$ (574 of 1798 patients), $f_3 = 0.02$ (35 of 1798 patients). In (a) and (b) the estimated hazard associations, $\beta_r^{\ell\mu}$, with serum glucose, triglyceride (TG), total cholesterol (TC) and age, for BC death ($r = 1$) and CV death ($r = 2$) respectively for the dominant ($\ell = 1$: red) and next largest ($\ell = 2$: green) latent classes. In (c) the frailty-weighted base hazard rates for BC and CV death (solid and dashed lines respectively) for the two dominant latent classes; note that the base hazard rate for CV death exceeds that for BC death for follow-up times greater than about 15 years ($\ell = 1$: red, $\ell = 2$: green). In (d) the crude (solid lines) and marginal (dashed lines) cause-specific survival against BC (red lines) and CV death (green lines); observe that the crude survival against CV death, S_2, is poorer than its marginal counterpart, $\tilde{S}_2$.

which accounts for 32%. The smallest latent class ($\ell = 3$) is almost wholly comprised of relatively younger individuals, for whom death from other causes was reported. It is likely that the inclusion of this third class, containing only 35 of the 1798 individuals according to retrospective class assignment, enables more effective estimation of the other two classes, as it allows the effects of two differing groups within those reported as having death from other causes to be distinguished. As BC and CV death shall be the focus of the remainder of this section, parameter estimates and survival curves pertaining to the small third latent class ($\ell = 3$) are omitted. The estimated associations, base hazard rates, and marginal class-specific survival curves are shown in Figure 5.

The largest class ($\ell = 1$) has the greatest hazard for BC death, having a base hazard rate which exceeds that of the second class ($\ell = 2$) over the entire time interval (Figure 5 (c)); the opposite is true for the hazard for CV death, with the second class ($\ell = 2$) having the greatest hazard for CV death. The base hazard rate for BC death is most significant at diagnosis ($t = 0$) and diminishes with time whereas the risk of CV death increases with time. The base hazard for CV death exceeds the risk of BC death after about 20 years from diagnosis for those in the first class ($\ell = 1$) and after about 16 years for those in the second class ($\ell = 2$). This is generally consistent with previous findings regarding the importance of considering comorbidities for women with breast cancer [42]. Accordingly, survival against BC death is greater for the less frail second class ($\ell = 2$) than for the first class ($\ell = 1$), and survival against CV death is greater in the first class ($\ell = 1$) than in the second class ($\ell = 2$). In the relatively frail first class ($\ell = 1$), triglyceride levels were found to be associated with an increased hazard for both CV death (CV, $\ell = 1$: HR=6.4, 95% CI=[1.9,22.3], $p=0.003$) and for BC death (BC, $\ell = 1$: HR=1.6, 95% CI=[0.9,2.8], $p=0.09$). Standard Cox analysis suggests a weaker association with an increased hazard against CV death (CV, Cox: HR=1.6, 95% CI=[1.2,2.2], $p=0.003$) than is suggested by our analysis, a probable consequence of those members of the cohort for which triglyceride levels are not associated with increased hazard ($\ell = 2$) diluting the effects of those for which triglyceride levels are associated with increased hazard ($\ell = 1$). The same is also true for BC death, for which standard Cox analysis indicates that triglyceride levels are associated with a modestly increased hazard for BC death (BC, Cox: HR=1.2, 95% CI [1.0,1.5], $p=0.076$). Serum glucose was found to be associated with an increased hazard for CV death (CV, $\ell = 2$: HR=1.6, 95% CI=[1.0,2.5], $p=0.04$) for those in the relatively less frail second class ($\ell = 2$), as was total cholesterol (CV, $\ell = 2$: HR=1.7, 95% CI=[0.8,3.5], $p=0.18$).

In contrast to standard Cox analysis, which suggests that age is associated with a reduced hazard for BC death for the cohort as a whole (BC, Cox: HR=0.8, 95% CI=[0.6,1.0], $p=0.05$), our present analysis indicates that for the majority of the cohort age is associated with an increased hazard for BC death (BC, $\ell = 1$:

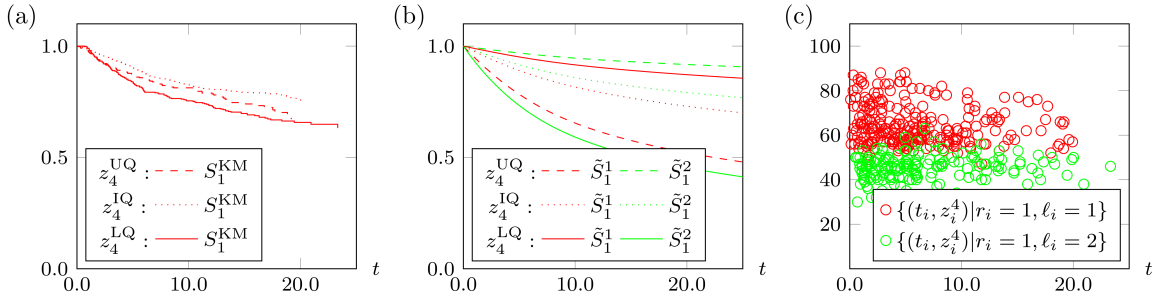

Figure 6: *Age and survival in women from the AMORIS population diagnosed with breast cancer:* In (a) the age-conditioned Kaplan-Meier estimator for BC death; survival of both the upper and lower quartiles of age the proportional hazards is poorer than that for the inter-quartile range for age (i.e. the proportional hazards assumption is not met) and suggests heterogeneity in the cohort. In (b) the class-specific marginal survival curves, $\tilde{S}_r^\ell$ conditioned on the age covariate (z_4), for BC death ($r = 1$). The substantial difference in survival against BC death between those in the lower quartile (LQ) and upper quartile (UQ) for age is indicative of the strength of the association of age with hazard for BC death (see Figure 5). In the more frail first class ($\ell = 1$) survival against BC death decreases with increasing age, while the opposite is true of the second class ($\ell = 2$). However, survival of the second class ($\ell = 2$) is actually poorer for younger individuals than it is for older members of the more frail first class ($\ell = 1$). In (c) the time-to-event and age for each individual having succumbed to BC death is shown, showing a clear stratification, according to age, of the latent classes to which individuals are retrospectively assigned. Almost all individuals younger than about 50 years old are assigned to the second latent class ($\ell = 2$), whereas those older than about 50 years are assigned to the first class ($\ell = 1$).

HR=3.5, 95% CI=[1.5,8.1], p=004) and is associated with a reduced hazard for BC death for about only one third of the cohort (BC, $\ell = 2$: HR=0.09, 95% CI=[0.02,0.30], $p \approx 10^{-4}$). The age of an individual was also found to be associated with their hazard for CV death, most significantly being strongly associated with an increased hazard for CV death in the less frail second class (CV, $\ell = 2$: HR=42, 95% CI=[16,112], $p < 10^{-7}$). The age-conditioned risk-specific Kaplan-Meier estimator and class-specific decontaminated survival for BC death are shown in Figure 6, along with the age, time-to-event, and retrospectively assigned class, for those individuals recorded as succumbing to BC death. The stratification of the two retrospectively allocated classes for those individuals succumbing to BC death, occurring clearly between ages about 50 and 55 years, suggest that menopausal status may be informative as to expected survival against BC death, as shown in Figure 6(c).

Survival against BC death for the more frail first class ($\ell = 1$) is greatest for the younger members of the cohort, with over 80% survival for those in the LQ for age as opposed to less than 50% survival for those in the UQ for age. In the second class ($\ell = 2$) survival against BC death is markedly poorer for younger members of the cohort as a consequence of age being strongly associated with reduced hazard for BC death for this class.

6 Discussion

In this study, we have introduced a Bayesian latent class approach for survival analysis, designed to deal with the challenges of residual cohort or disease heterogeneity (i.e. heterogeneity not captured by the available covariates) and informative censoring by competing risks. Any regression method that, in addition to estimating crude hazard rates and survival functions, aims to estimate also the *marginal* hazard rates and survival probabilities, which would appear to be a prerequisite for distinguishing between real and false covariate protectivity or aetiology effects, must inevitably make assumptions that are not verifiable from the survival data alone [1]. Here, the introduction and assumed validity of heterogeneity-induced informative censoring, where risks are independent only at the level of individuals, in combination with assuming proportional hazards to hold at the level of individuals, has enabled us to identify the marginal hazard rates and survival functions. Our (unverifiable) assumption that any risk event time correlations are caused by residual cohort heterogeneity would appear to be a natural one, and is certainly weaker than assuming uncorrelated risks.

The software package, *ALPACA*, which implements the formalism introduced herein provides practical tools for survival analysis which can be applied to heterogeneous cohorts and in the presence of heterogeneity-induced informative censoring. Cohort parameter estimation is achieved through the combination of i) a latent class

model which captures, for *all* risks and for *each* class, the base hazard rate, covariate associations and relative frailty, and ii) Bayesian model selection to determine the optimal number of classes, the optimal parametrisation of each base hazard rate, and the extent of heterogeneity in the cohort. Once the optimal quantitative description of the cohort has been estimated, crude and marginal cause-specific survival curves can be compared to gauge the extent of any heterogeneity-induced informative censoring, and differences between class-specific survival curves can be examined. Exploration of correlations between covariate values and retrospectively assigned class membership can offer additional insight into a cohort, and may aid the search for new informative biomarkers. Applied to simulated data, our analysis was shown to effectively characterise heterogeneous cohorts, successfully removing heterogeneity-induced false protectivity and false aetiology effects, and even discriminating between classes that differ only in their base hazard rates. Retrospective class allocation was demonstrated to have an impressive accuracy, even for survival data modelling a cohort containing 200 individuals.

On application to survival data from the ULSAM cohort, with prostate cancer as the primary risk, our analysis leads to plausible alternative explanations for previous counter-intuitive inferences (such as a weak protective effect on PC of smoking), in terms of distinct sub-groups of patients with distinct risk factors and overall frailties. In the ULSAM cohort, it was also shown that the men's metabolic status introduce competing risk problems, although using traditional methods, several different analyses had to be done to reveal the underlying risk pattern [40], which with our proposed method could be done coherently in one analysis.

Applied to survival data for women diagnosed with breast cancer from the AMORIS cohort, our analysis suggests that for breast cancer death risk the base hazard rate decreases with time but increases with time for cardiovascular death risk, and the age-class membership correlations may suggest differences in association patterns and survival against breast cancer between pre- and post-menopausal women.

References

[1] Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences of the United States of America* 1975; **72**(1):20-22.

[2] Gail M. A review and critique of some models used in competing risk analysis. *Biometrics* 1975; **31**(1):209-222. DOI: 10.2307/2529721

[3] Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. *International Journal of Epidemiology* 2012; **41**(3):861-870. DOI: 10.1093/ije/dyr213

[4] Di Serio C. The protective impact of a covariate on competing failures with an example from a bone marrow transplantation study. *Lifetime Data Analysis* 1997; **3**(2):99-122. DOI: 10.1023/A:1009672300875

[5] Scharfstein DO, Robins JM. Estimation of the failure time distribution in the presence of informative censoring. *Biometrika* 2002; **89**(3):617-634.

[6] Dignam JJ, Zhang Q, Kocherginsky M. The use and interpretation of competing risks regression models. *Clinical Cancer Research* 2012; **18**(8):2301-2308. DOI: 10.1158/1078-0432.CCR-11-2097

[7] Thompson CA, Zhang Z-F, Arah OA. Competing risk bias to explain the inverse relationship between smoking and malignant melanoma. *European Journal of Epidemiology* 2013; **28**(7):557-567. DOI: 10.1007/s10654-013-9812-0

[8] Soneji S, Beltrán-Sánchez H, Sox HC. Assessing progress in reducing the burden of cancer mortality, 1985-2005. *Journal of Clinical Oncology* 2014; **32**(5):444-448. DOI: 10.1200/JCO.2013.50.8952

[9] Klein JP. Competing risks. *Wiley Interdisciplinary Reviews: Computational Statistics* 2010; **2**(3):333-339. DOI: 10.1002/wics.83

[10] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**(282):457-481. DOI: 10.2307/2281868

[11] Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1972; **34**(2):187-220.

[12] Vaida F, Xu R. Proportional hazards model with random effects. *Statistics in Medicine* 2000; **19**(24):3309-3324. DOI: 10.1002/1097-0258(20001230)19:24<3309::AID-SIM825>3.0.CO;2-9

[13] Rosner B, Glynn RJ, Tamimi RM, Chen WY, Colditz GA, Willett WC, Hankinson SE. Breast cancer risk prediction with heterogeneous risk profiles according to breast cancer tumor markers. *American Journal of Epidemiology* 2013; **178**(2):296-308. DOI: 10.1093/aje/kws457

[14] Wienke A. *Frailty Models in Survival Analysis*. Chapman & Hall CRC Biostatistics Series: Boca Raton, 2010.

[15] Duchateau L, Janssen P. *The Frailty Model (Statistics for Biology and Health)*. Springer: New York, 2008. DOI: 10.1007/978-0-387-72835-3

[16] Lancaster T. Econometric methods for the duration of unemployment. *Econometrica* 1979; **47**(4):939-956. DOI: 10.2307/1914140

[17] Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 1979; **16**(3):439-454. DOI: 10.2307/2061224

[18] Zahl PH. Frailty modelling for the excess hazard. *Statistics in Medicine* 1997; **16**(14):1573-1585. DOI: 10.1002/(SICI)1097-0258(19970730)16:14<1573::AID-SIM585>3.0.CO;2-Q

[19] Yashin AI, Iachine IA. Genetic analysis of durations: correlated frailty model applied to survival of Danish twins. *Genetic Epidemiology* 2005; **12**(5):529-538. DOI: 10.1002/gepi.1370120510

[20] Gorfine M, Hsu L. Frailty-based competing risks model for multivariate survival data. *Biometrics* 2011; **67**(2):415426. DOI: 10.1111/j.1541-0420.2010.01470.x

[21] Keiding N, Andersen PK, Klein JP. The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine* 1997; **16**(2):215-224. DOI: 10.1002/(SICI)1097-0258(19970130)16:2<215::AID-SIM481>3.0.CO;2-J

[22] Lazarsfeld PF. The logical and mathematical foundation of latent structure analysis. *SA Stouffer at al. (Eds) Measurement and Prediction*. Princeton: Princeton University Press, 1950.

[23] Huang X, Wolfe RA. A frailty model for informative censoring. *Biometrics* 2002; **58**(3):510-520. DOI: 10.1111/j.0006-341X.2002.00510.x

[24] Muhten B, Masyn K. Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics* 2005; **30**(1):27-58. DOI: 10.1.1.333.1807

[25] Reboussin BA, Anthony JC. Latent class marginal regression models for modelling youthful drug involvement and its suspected influences. *Statistics in Medicine* 2001; **20**(4):623-639. DOI: 10.1002/sim.695

[26] Proust-Lima C, Mbéry S, Taylor JMG, Jacqmin-Gadda H. Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research* 2014; **23**(1):74-90. DOI: 10.1177/0962280212445839

[27] 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. DOI: 10.2307/2670170

[28] Fine JP. Regression modeling of competing crude failure probabilities. *Biostatistics* 2001; **2**(1):85-97. DOI: 10.1093/biostatistics/2.1.85

[29] Klein JP, Andersen PK. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* 2005; **61**(1):223-229. DOI: 10.1111/j.0006-341X.2005.031209.x

[30] Jeong JH, Fine JP. Parametric regression on cumulative incidence function. *Biostatistics* 2007; **8**(2):184-196.

[31] Katsahian S, Boudreau C. Estimating and testing for center effects in competing risks. *Statistics in Medicine* 2011; **30**(13):1608-1617. DOI: 10.1002/sim.4132

[32] Heckman JJ, Honoré BE. The identifiability of the competing risks model. *Biometrika* 1989; **76**(2):325-330. DOI: 10.1093/biomet/76.2.325

[33] Abbring JH, van den Berg GJ. The identifiability of the mixed proportional hazards competing risks model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2003; **65**(3):701-710. DOI: 10.1111/1467-9868.00410

[34] Zheng M, Klein JP. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika* 1995; **82**(1):127-138. DOI: 10.1093/biomet/82.1.127

[35] *Uppsala Longitudinal Study of Adult Men*. Department of Public Health and Caring Sciences/ Geriatrics, Uppsala Universitet. www2.pubcare.uu.se/ULSAM/.

[36] Grundmark B, Zethelius B, Garmo H, Holmberg L. Serum levels of selenium and smoking habits at age 50 influence long term prostate cancer risk; a 34 year ULSAM follow-up. *BMC Cancer* 2011; **11**:431. DOI: 10.1186/1471-2407-11-431

[37] Holme I, Aastveit AH, Hammar N, Jungner I, Walldius G. Inflammatory markers, lipoprotein components and risk of major cardiovascular events in 65,005 men and women in the Apolipoprotein MOrtality RISk study (AMORIS). *Atherosclerosis* 2010; **213**(1):299-305. DOI: 10.1016/j.atherosclerosis.2010.08.049

[38] Holme I, Aastveit AH, Jungner I, Walldius G. Relationships between lipoprotein components and risk of myocardial infarction: age, gender and short versus longer follow-up periods in the Apolipoprotein MOrtality RISk study (AMORIS). *J. Intern. Med.* 2008; **264**(1):30-38. DOI: 10.1111/j.1365-2796.2008.01925.x

[39] Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical recipes in C - the art of scientific computing*. Cambridge: Cambridge University Press, 1992.

[40] Grundmark B, Garmo H, Loda M, Busch C, Holmberg L, Zethelius B. The metabolic syndrome and the risk of prostate cancer under competing risks of death from other causes. *Cancer Epidemiol Biomarkers Prev.* 2010; **19**(8):2088-2096. DOI: 10.1158/1055-9965.EPI-10-0112

[41] Wulaningsih W, Vahdaninia M, Rowley M, Holmberg L, Garmo H, Malmstrom H, Lambe M, Hammar N, Walldius G, Jungner I, Coolen ACC, Van Hemelrijck M. Prediagnostic serum glucose and lipids in relation to survival in breast cancer patients: a competing risk analysis. *BMC Cancer* 2015; **15**:913. DOI: 10.1186/s12885-015-1928-z

[42] Patnaik JL, Byers T, Di Guiseppi C, Dabelea D, Denberg TD. Cardiovascular disease competes with breast cancer as the leading cause of death for older females diagnosed with breast cancer: a retrospective cohort study. *Breast Cancer Res.* 2011; **13**:R64. DOI: 10.1186/bcr2901

[43] MacKay DJC. *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press, 2003.

A Latent class model: additional details and identities

A.1 Connection between cohort-level and individual-level cause-specific hazard rates

The relationship between the cohort-level covariate-conditioned cause-specific hazard rate, $h_r(t|\mathbf{z})$, and the hazard rates of the individual members of the cohort, $h_r^i(t)$, is obtained by substitution of $p(t_0,\ldots,t_R|\mathbf{z}) = \sum_{i,\ \mathbf{z}_i=\mathbf{z}} p_i(t_0,\ldots,t_R) / \sum_{i,\ \mathbf{z}_i=\mathbf{z}} 1$ into (1), to give,

$$h_r(t|\mathbf{z})S(t|\mathbf{z}) = \frac{\sum_{i,\ \mathbf{z}_i=\mathbf{z}} \int_0^\infty \cdots \int_0^\infty \mathrm{d}t_0 \ldots \mathrm{d}t_R \; p_i(t_0,\ldots,t_R)\delta(t-t_r)\prod_{r'\neq r}\theta(t_{r'}-t)}{\sum_{i,\ \mathbf{z}_i=\mathbf{z}} 1} = \frac{\sum_{i,\ \mathbf{z}_i=\mathbf{z}} S_i(t)h_r^i(t)}{\sum_{i,\ \mathbf{z}_i=\mathbf{z}} 1}. \quad (16)$$

Insertion of the identity $S(t|\mathbf{z}) = \sum_{i,\ \mathbf{z}_i=\mathbf{z}} S_i(t) / \sum_{i,\ \mathbf{z}_i=\mathbf{z}} 1$ and the individualised survival function, $S_i(t) = \mathrm{e}^{-\sum_r \int_0^t \mathrm{d}s \; h_r^i(s)}$, into the above leads to the relationship expressed by (4).

A.2 Equivalence of crude and marginal survival in the absence of competing risks

We show below that, in the case where a cohort is exposed to only one risk, the crude and marginal survival functions, $S_r(t)$ and $\tilde{S}_r(t)$ respectively (6), are equivalent. As only one risk is present (i.e. $R = 1$), the crude and marginal cohort-level hazard rates, $h_1(t|\mathbf{z})$ and $\tilde{h}_1(t|\mathbf{z})$ respectively (5), are identical. As the crude and marginal survival are initially equal (i.e. $S_r(t = 0) = \tilde{S}_r(t = 0) = 1$), to prove that they are the same at any time t, in the absence of *any* competing risks, it is sufficient to show that their time derivatives are also equal, as follows,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\ln S_1(t|\mathbf{z}) - \ln \tilde{S}_1(t|\mathbf{z})\right] = -h_1(t|\mathbf{z}) + \frac{\sum_{i,\mathbf{z}_i=\mathbf{z}} h_1^i(t)\mathrm{e}^{\int_0^t \mathrm{d}s\ h_1^i(s)}}{\sum_{i,\mathbf{z}_i=\mathbf{z}} \mathrm{e}^{\int_0^t \mathrm{d}s\ h_1^i(s)}} = -h_1(t|\mathbf{z}) + h_1(t|\mathbf{z}) = 0. \tag{17}$$

A.3 Bayesian retrospective class assignment

Bayesian arguments allow us to calculate class membership probabilities *retrospectively* for any individual for whom we have their covariates $\mathbf{z}$ and survival information (t, r). The probability of an individual belonging to class ℓ, conditioned on their covariates and survival information, follows from (11), and is given by,

$$p(t, r|\mathbf{z}, \ell) = \mathrm{e}^{-\Lambda_0(t)}\ \lambda_r^\ell(t)\mathrm{e}^{\boldsymbol{\beta}_r^\ell\cdot\mathbf{z} - \sum_{r'=1}^R \exp(\boldsymbol{\beta}_{r'}^\ell\cdot\mathbf{z})\Lambda_{r'}^\ell(t)}. \tag{18}$$

Given that $p(t, r, \ell|\mathbf{z}) = P(t, r|\mathbf{z}, \ell)w_\ell$ and $p(t, r|\mathbf{z}) = \sum_{\ell'=1}^L p(t, r|\mathbf{z}, \ell')w_{\ell'}$, it follows that the probability of an individual belonging to class ℓ is given by

$$p(\ell|t, r, \mathbf{z}) = \frac{w_\ell p(t, r|\mathbf{z}, \ell)}{\sum_{\ell'=1}^L w_{\ell'} p(t, r|\mathbf{z}, \ell')}. \tag{19}$$

Substitution of $p(t, r|\mathbf{z}, \ell)$ into the above expression leads to the retrospective class membership probability, as is given for the fully heterogeneous model variant ($M = 3$) in (10).

A.4 Prior distributions for the latent class model parameters

To estimate a cohort's characteristics using Bayesian inference (Section 2.5), it is necessary that a prior distribution, $p(\boldsymbol{\theta}_{KLM})$, on the latent class model parameters, $\boldsymbol{\theta}_{KLM}$, be defined. This requires that suitable prior distributions be chosen to encode any available information regarding the weights, frailties, associations, and those parameters used in the base hazard rate approximation. In this paper we chose the maximum entropy prior on the L weight parameters, subject to the constraint $\sum_{\ell\leq L} w_\ell = 1$, which is given by $p(w_1, \ldots, w_L) = 1/Z_L$, where $Z_L = \int_0^1 \mathrm{d}w_1 \ldots \int_0^1 \mathrm{d}w_L\ \delta(\sum_\ell w_\ell - 1) = 1/(L-1)!$ (the so-called flat Dirichlet distribution). Unit-variance zero-average Gaussian priors were chosen for the frailty and association parameters, justified by our decision to pre-process the data such that all covariate distributions are normalised (by linear rescaling) to zero average and unit variance over the cohort. For all (nonnegative) base rate parameters $\xi_{kr\ell}$, see Appendix A.6, we chose identical exponential priors.

A.5 Model evidence determination using a Gaussian approximation

Estimation of a cohort's characteristics using Bayesian model selection requires that the model supported by the greatest 'evidence' be determined. The posterior distribution can be written as $p(\boldsymbol{\theta}|D) = Z_{KLM}^{-1}\exp[-S(\boldsymbol{\theta}, D)]$, where $S(\boldsymbol{\theta}, D) = -\ln[p(\boldsymbol{\theta}|\mathcal{H}_{KLM})p(D|\boldsymbol{\theta}, \mathcal{H}_{KLM})]$, and the model evidence is proportional to Z_{KLM}, as given by

$$Z_{KLM} = \int \mathrm{d}\boldsymbol{\theta}\ p(\boldsymbol{\theta}|\mathcal{H}_{KLM})p(D|\boldsymbol{\theta}, \mathcal{H}_{KLM}). \tag{20}$$

In our analysis, determination of the evidence (the volume of the posterior) is achieved via a Gaussian approximation to the posterior, equivalent to making a Taylor expansion of $S(\boldsymbol{\theta}, D)$ around its maximum, as described in e.g. [43]. The posterior is thus approximated by $p(\boldsymbol{\theta}|D, \mathcal{H}_{KLM}) = Z_{KLM}^{-1}\exp[-S(\boldsymbol{\theta}^\star) - (\boldsymbol{\theta}-\boldsymbol{\theta}^\star).A(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)/2]$, where $\boldsymbol{\theta}^\star$ is the location of its maximum and A is the Hessian matrix. The Gaussian-approximation for (20), $Z_{KLM}^\star$, is obtained by integrating over the normalised approximated posterior distribution, and is given by,

$$Z_{KLM}^\star = \mathrm{e}^{-S(\boldsymbol{\theta}^\star)}(2\pi)^{Y/2}(\det A)^{-1/2}, \tag{21}$$

where Y is the dimensionality of $\boldsymbol{\theta}$. In this work, the matrix A is estimated numerically by investigating the curvature of the posterior distribution around its maximum.

A.6 Base hazard rate parametrisation using a spline approximation

A standard interpolation method has been chosen for the parametrisation of the cause-specific, and possibly class-specific, base hazard rates, $\lambda_r^\ell(t)$. A spline construction, with $K+1$ equidistant time points covering the range of survival times in a cohort's survival data, is the basis of the parametrised base hazard rates used in our analysis. More irregular base hazard rates can be modelled by increasing the number of time points. The time points are defined by $\tilde{t}_k = t_{\text{min}} + k(t_{\text{max}} - t_{\text{min}})/K$, with $k = 0 \dots K$, where $t_{\text{min}} = \text{min}_{i\in\{1,\dots,N\}} t_i$ and $t_{\text{max}} = \text{max}_{i\in\{1,\dots,N\}} t_i$ are the minimum and maximum survival time in the cohort's survival data. A base rate parameter, $\xi_{kr\ell}$, is assigned for each time point $\tilde{t}_k$, for each risk $r = 0 \dots R$, and for each class $\ell = 1, \dots, L$s. This allows us to define the base hazard rates, $\lambda_r^\ell(t)$, for each risk and class as smooth Gaussian convolutions, with uniform variation time scale, $\sigma = \frac{1}{2}(\tilde{t}_{k+1} - \tilde{t}_k)$, as follows,

$$\lambda_r^\ell(t|\boldsymbol{\xi}) = \frac{\sum_{k=0}^K \xi_{kr\ell}\ \text{e}^{-\frac{1}{2}(t-\tilde{t}_k)^2/\sigma^2}}{\sum_{k=0}^K \text{e}^{-\frac{1}{2}(t-\tilde{t}_k)^2/\sigma^2}}, \qquad \sigma = \frac{t_{\text{max}} - t_{\text{min}}}{2K}, \qquad t \geq 0. \tag{22}$$

where $\boldsymbol{\xi}$ represents the set of all base rate parameters. The integrated rates $\Lambda_r^\ell(t|\boldsymbol{\xi}) = \int_0^t \text{d}s\ \lambda_r^\ell(s|\boldsymbol{\xi})$ are obtained numerically from (22), via 11-point Gaussian Quadrature, see e.g. [39], applied separately to the n intervals $[jt/n, (j+1)t/n]$ with $j = 0 \dots n-1$ and $n = 20$.

B Generation of simulated time-to-event data

Simulated data, describing risks that are independent at the level of individuals, and with individual cause-specific hazard rates of the form given in Table 2 was generated as described below. The latent class, $\ell \in 1, \dots, L$, to which each individual, $i = 1, \dots, N$, belongs is set at the time of generation. Covariate values for each individual, $z_i^\mu \in \mathcal{N}(0, 1)$, were generated independently from a zero average and unit variance normal distribution. A latent event time, t_i^r, was generated for each individual and for each risk, $r = 1 \dots R$, according to $t_i^r(u) = \Lambda_r^{\ell,\text{inv}}(\exp(-\boldsymbol{\beta}_r^\ell.\mathbf{z}_i).\log(1/u))$ where $\Lambda_r^{\ell,\text{inv}}(v)$ is the inverse of $\Lambda_r^\ell(t) = \int_0^t \text{d}s\ \lambda_r^\ell(s)$ and $u \in [0, 1]$ is a uniformly distributed random variable. The outcome data for individual i is that for which the latent event-time is smallest $t_i = \text{min}_{r\in\{1,\dots,R\}} t_i^r$, and $r_i = \text{argmin}_{r\in\{1,\dots,R\}} t_i^r$. Should t_i exceed the trial duration τ, then for individual i we report $(r_i, t_i) = (0, \tau)$.