James McVittie*, David Wolfson, David Stephens, Vittorio Addona and David Buckeridge

Parametric models for combined failure time data from an incident cohort study and a prevalent cohort study with follow-up

https://doi.org/10.1515/ijb-2020-0042
Received April 1, 2020; accepted September 29, 2020; published online October 12, 2020

Abstract: A classical problem in survival analysis is to estimate the failure time distribution from right-censored observations obtained from an incident cohort study. Frequently, however, failure time data comprise two independent samples, one from an incident cohort study and the other from a prevalent cohort study with follow-up, which is known to produce length-biased observed failure times. There are drawbacks to each of these two types of study when viewed separately. We address two main questions here: (i) Can our statistical inference be enhanced by combining data from an incident cohort study with data from a prevalent cohort study with follow-up? (ii) What statistical methods are appropriate for these combined data? The theory we develop to address these questions is based on a parametrically defined failure time distribution and is supported by simulations. We apply our methods to estimate the duration of hospital stays.

Keywords: combined cohort; maximum likelihood estimation; survival analysis.

1 Introduction

In a medical study, researchers may wish to estimate the distribution of the duration of a disease or medical status. The data that are available depend on the study design. For example, the French Pulmonary Arterial Hypertension Network prospectively followed a cohort for three years for the occurrence of pulmonary arterial hypertension (PAH) [1]. The times between diagnosed PAH and death comprised the time-to-event data. Subjects who entered the study already diagnosed with PAH formed a prevalent cohort whereas those who had onset of PAH during the study period formed an incident cohort. Thus, the study design resulted in observed time-to-event data of two types. Similarly, the Nun Study of Aging and Alzheimer's Disease was a prospective observational study in which the enrolled subjects were classified either as incident or prevalent cases [2]. Outside the field of medical research, examples of combined incident and prevalent cohort data can be found in the areas of finance, sports analysis and public policy [3–5].

Statistical procedures for data arising exclusively from either an incident or prevalent cohort have been thoroughly examined in the survival analysis literature. When failure time data are subject to random right-censoring, the Kaplan-Meier estimator can be used to consistently estimate the unknown survivor function non-parametrically [6, 7]. Alternatively, the parametric maximum likelihood procedures outlined by Kalbfleisch and Prentice may also be used [8]. For data that are generally left-truncated and right-censored, the survivor function may be estimated non-parametrically using an altered form of the Kaplan-Meier estimator sometimes called the Tsai, Jewell and Wang (TJW) estimator [9–11]. However, if the initial dates

***Corresponding author: James McVittie**, McGill University, Mathematics and Statistics, 805 Sherbrooke Street West, Montreal, Quebec Canada, E-mail: james.mcvittie@mail.mcgill.ca. https://orcid.org/0000-0001-6039-1312
David Wolfson and David Stephens, McGill University, Mathematics and Statistics, 805 Sherbrooke Street West, Montreal, Quebec Canada, E-mail: david.wolfson@mcgill.ca (D. Wolfson)
Vittorio Addona, Macalester College, Mathematics, Statistics and Computer Science, St.Paul, Minnesota, United States
David Buckeridge, McGill University, Epidemiology, Biostatistics and Occupational Health, Montreal, Quebec Canada

of the failure times are assumed to arise independently from a stationary Poisson process resulting in uniform truncation times, sharper inference may be made [12]. Under this assumption of stationarity, Asgharian et al. derived the non-parametric maximum likelihood estimator (NPMLE) of the survivor function, and established its asymptotic properties [12]. The asymptotic properties of the TJW estimator, appropriate under arbitrary left-truncation, but less efficient than under stationarity, are given in [9]. An alternative option is to assume a fully parametric model for the survivor function, while allowing the truncation distribution to be arbitrary or, if justifiable, to be uniform. The asymptotic properties of the MLEs in either of these two settings follow from standard likelihood theory, with modifications for length-bias and censoring [13].

Frequently however, failure time data comprise two independent samples, one from an incident cohort study and the other from a prevalent cohort study with follow-up. There are drawbacks to each of these two types of study when viewed separately. Briefly, pure incident cohort studies require lengthy follow-up, first to capture a sufficient number of incident events and thereafter, to capture a sufficient number of uncensored failure times. Often cost and logistical constraints preclude extensive follow-up and consequently, the NPMLE of the survivor function is left undefined for a large part of its support. Prevalent cohort studies with follow-up suffer less from these drawbacks since failure intervals are by definition, intercepted "midstream" at the start of follow-up. Moreover, even with restricted follow-up, increasing the sample size of the initial cohort will lead to improved coverage of the support of the targeted survivor function; there is improved coverage because there is no constraint on the initiation dates of those entering the initial prevalent cohort. On the other hand, since the subjects who comprise the prevalent cohort are determined by screening a larger cohort cross-sectionally, their observed failure times are subject to left-truncation and biased. They are therefore, not representative of the underlying survival distribution. Although the observed failure times from a general prevalent cohort study with follow-up are biased, we reserve the expression *length-biased* for the particular setting in which the underlying incidence process is a stationary Poisson point process [12]. Under general left truncation, the TJW estimator, can sometimes yield visibly poor estimates of the survivor function, when used for data collected in a pure prevalent cohort study with follow-up [14, 15]

Exploiting the respective advantages of these two types of study, Wolfson et al. show that combining these data can have considerable benefit in the arbitrary truncation distribution model [15]. Importantly, the TJW estimator is the NPMLE and its asymptotic properties may be established in this combined setting [16]. However, when a uniform truncation distribution may be assumed (that is, "under stationarity") in the combined cohort setting, the NPMLE is not simply the NPMLE obtained from a pure prevalent cohort study with follow-up by setting some of the truncation times equal to zero. These zero-truncation times are not consistent with their assumed uniformity. A non-parametric estimator of the survivor function may nevertheless, be obtained [17, 18]. Unfortunately, there is a major drawback to non-parametric estimation under stationarity in this setting; the asymptotic properties of the NPMLE under random informative censoring are unknown and remain an open problem (see [19], Problem 6.4).

In this article, under stationarity, we therefore propose the use of parametric models for the survivor function with combined data. Although the use of parametric models means a loss of model robustness, we show in this article that this drawback is offset by the availability of distributional properties for our parametric estimators. This viewpoint is supported to some extent by Miller who compares the performance of common failure time parametric survival models to the Kaplan-Meier estimator [20]. We establish consistency and asymptotic Normality of the MLEs. We note that the estimators are not functions of identically distributed random variables since one set is length-biased (from the prevalent cohort) and the other is not (from the incident cohort). Further, we do not impose any structure for the onset process of the incident cases. Imposing a structure on the onset process for the incident cases may be a restriction which we wish to avoid in a meta-analysis. A further complication is that the failure times from the prevalent cohort are informatively censored while those from the incident cohort are non-informatively censored. Consequently, derivation of the asymptotics requires some care.

Several authors have considered scenarios that allow for a combination of incident and prevalent cohort failure time data, under various stationarity assumptions. In the field of geology, Laslett proposed a procedure for estimating the bivariate distribution function of lengths and angles of different types of cracks in an observed rock face [21]. Subsequently, Wijers and van der Laan considered Laslett's estimator in the one-dimensional case and derived its associated asymptotic properties under the assumption of an underlying stationary Poisson onset process for *all initiating events*, both inside and outside a window of observation [22, 23]. In particular, the incident cases that arise in the observation window are assumed to be generated by the same homogeneous Poisson process as to the left of the window. We do not impose this restriction. Importantly, the setting of Wijers and van der Laan permits only administrative censoring. This precludes the possibility of (random) censoring due to the loss of follow-up, which is a hallmark of most medical cohort studies. Vardi proposed an EM algorithm for non-parametric estimation of the survivor function for combined data allowing for random censoring [17]. He was unable to assert that his estimator is the NPMLE nor was he able to establish the distributional properties of his estimator (see [19] reference cited above). Saarela et al. considered a conditional likelihood method for making inferences about the incidence rate using combined cohort data [24]. However, their goal was entirely different from ours. They used simulations to compare their methods to analyses based on prevalent cases only.

The remainder of the article is laid out as follows. In Section 2, we define the notation for combined cohort failure time data. We give the joint likelihood function in Section 3 and state the main theorem on the consistency and asymptotic Normality of the MLE under certain regularity conditions (a detailed proof is given in the Supplementary materials). Through simulations we examine how the performance of the combined cohort MLE varies when the proportion of incident and prevalent cohort subsample sizes change while the grand sample size remains fixed, as well as when the chosen parametric model is misspecified. In Section 5, we apply our methods to estimate the durations of stays in a Montreal area hospital. Section 6 contains some concluding remarks.

2 Notation

To construct the combined likelihood function, we begin by defining the data that arise from the contributing incident and prevalent cohorts separately. We assume that all failure intervals of interest begin with initiation times (or dates), which for simplicity of exposition we shall call onset times (dates). Failure intervals will then be taken to be "disease" durations.

Let T denote the underlying failure time random variable with parametric density function $f_U(\cdot; \boldsymbol{\theta})$ and parametric survivor function $S_U(\cdot; \boldsymbol{\theta})$. An *incident cohort* comprises a cohort of subjects who are determined to be disease-free at some time origin, and who are followed for a fixed period of time. In this time period, some of the subjects (called incident cases) will experience disease onset and by the end of the study, some of these incident cases will either yield fully observed or randomly right-censored failure times. We shall assume that there is no cohort effect so that the dates of disease onset play no role other than in the determination of the fully observed or right-censored disease durations; thus, in the incident cohort we shall allow the incidence process to be arbitrary. Let C be the underlying incident cohort censoring random variable with non-parametric probability density function and survivor function, $f_C(\cdot)$ and $S_C(\cdot)$, respectively. Let n onsets occur in the study period and for $i \in \{1, 2, \ldots, n\}$, let the observed data consist of the pairs $(X_i, \delta_i) = (\min(T_i, C_i), \delta_i)$ where $\delta_i = 1$ if $T_i \leq C_i$ and 0 otherwise for $i \in \{1, 2, \ldots, n\}$.

For the prevalent cohort, let Z_j denote the onset date of subject j for $j \in \{1, 2, \ldots, k\}$. We note, in advance, that only a subset of these k onset dates will be observed, being the onset dates of those who comprise the prevalent cohort. Without loss of generality, we assume the start date of follow-up of those with prevalent disease is a fixed constant R. We call this *prevalence day*. This setup is easily extended to one that allows for staggered entry of the prevalent cases. We also assume that $\{Z_j, j = 1, 2, \ldots, k\}$ arise

from a stationary Poisson process. We define the truncation times $\widetilde{A}_j = R - Z_j$ for $j \in \{1, 2, \ldots, k\}$. The truncation times are therefore independently and uniformly distributed on every fixed interval $[a,R)$. Let $T_1^*, T_2^*, \ldots, T_k^*$ be the i.i.d. failure times of all those with onset dates prior to R. We assume that $T_i^* \sim f_U(\cdot; \boldsymbol{\theta})$, $\mathbb{E}(T_i^*) = \mu(\boldsymbol{\theta})$, and that subject i is recruited into the prevalent cohort if $T_i^* \geq \widetilde{A}_i$. We denote the left-truncation time of subject i, who is recruited into the prevalent cohort, by A_i and the residual life time of this subject by B_i for $i \in \{1, 2, \ldots, m\}$ where $m \leq k$. The A_is and B_is are equivalent, respectively, to the backward and forward recurrence times of renewal theory. Note that the A_is are not uniformly distributed as they form a selected subset of the uniformly distributed $\widetilde{A}_i$s. We further assume that each B_i, $i \in \{1, 2, \ldots, m\}$ is subject to potential random right-censoring by the random variable C_i^* with non-parametric density function and survivor function, $f_{C^*}(\cdot)$ and $S_{C^*}(\cdot)$, respectively. Thus, the observed prevalent cohort is comprised of m i.i.d. triples of random variables $(X_j^*, A_j, \delta_j^*) = (\min(B_j, C_j^*), A_j, \delta_j^*)$ where $\delta_j^* = 1$ if $B_j \leq C_j^*$ and 0 otherwise where the observed failure/censoring times are given by $Y_j = X_j^* + A_j$ for $j \in \{1, 2, \ldots, m\}$.

We denote the total sample size of the incident and prevalent cohorts by $l = n+m$. Let γ_j for $j \in \{1, 2, \ldots, l\}$ be the deterministic indicator function denoting whether the observed jth failure/censoring time belongs to the incident or prevalent cohort subsample. The combined cohort is thus comprised of l independent but not identically distributed quadruples of observations $(X_j\gamma_j + X_j^*(1-\gamma_j), \delta_j\gamma_j + \delta_j^*(1-\gamma_j), A_j(1-\gamma_j), \gamma_j)$ where $\gamma_j = 1$ if observation j is from the incident cohort and 0 otherwise, for $j \in \{1, 2, \ldots, l\}$. For a graphical representation of the observed cohort data, see Figure 1.

3 Estimation

For n i.i.d. observations from an incident cohort alone, under the assumption of non-informative random right-censoring, the likelihood function for $\boldsymbol{\theta}$ is given by

$$\mathscr{L}_I(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} f_U^{\delta_i}(x_i; \boldsymbol{\theta}) S_U^{1-\delta_i}(x_i; \boldsymbol{\theta}) \tag{1}$$

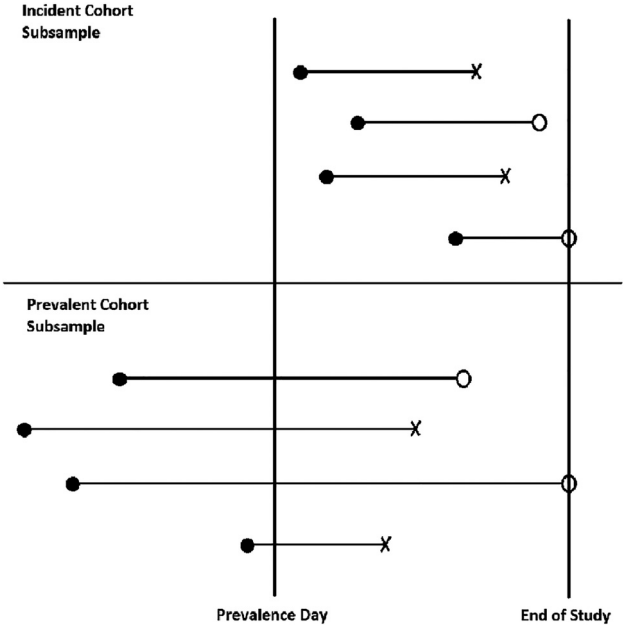

Figure 1: A graphical representation of a sample of combined incident and prevalent cohort right-censored failure time data. The filled circles represent the onset dates, the open circles represent the calendar dates of censoring and the crosses represent the calendar dates of failure. The incident cohort subsample consists of failure/censoring times with onset after prevalence day. The prevalent cohort subsample consists of failure/censoring times for which onset occurs prior to prevalence day where the associated (potentially unobserved) failure time surpasses prevalence day.

To ensure Eq. (1) yields an MLE of $\boldsymbol{\theta}$ contained in the parameter space Θ, we assume that not all n observations are censored. Studies with short follow-up may yield a high proportion of censored incident cases, which can be problematic for inference based on purely incident cohort data. However, under stationarity, the censoring mechanism is informative for a cohort of purely prevalent cohort data and the censored observations provide direct information on the failure time distribution beyond the information that failure would have occurred after them. Therefore, allowing for the addition of prevalent cases in the observed sample helps alleviate this problem. Under stationarity, for a pure prevalent cohort, the likelihood for $\boldsymbol{\theta}$ is given by

$$\mathscr{L}_P(\boldsymbol{\theta}) \propto \prod_{j=1}^{m} \frac{f_U^{\delta_j^*}(y_j;\boldsymbol{\theta}) S_U^{1-\delta_j^*}(y_j;\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \tag{2}$$

where $\mu(\boldsymbol{\theta}) = \int_0^\infty x f_U(x;\boldsymbol{\theta}) dx$ [13].

It is worth noting that the likelihood given in Eq. (2) does not require knowledge of the individual backward/forward recurrence times unlike the case of general left truncation [10], and estimation can be based solely on their sum. Under the assumption of between-subject independence in the combined cohort, the joint likelihood function, $\mathscr{L}_C$, can be expressed as the product of the likelihoods given in Eq. (1) and Eq. (2), yielding:

$$\begin{aligned} \mathscr{L}_C(\boldsymbol{\theta}) &= \mathscr{L}_I(\boldsymbol{\theta}) \times \mathscr{L}_P(\boldsymbol{\theta}) \\ &\propto \prod_{i=1}^{n} f_U^{\delta_i}(x_i;\boldsymbol{\theta}) S_U^{1-\delta_i}(x_i;\boldsymbol{\theta}) \prod_{j=1}^{m} \frac{f_U^{\delta_j^*}(y_j;\boldsymbol{\theta}) S_U^{1-\delta_j^*}(y_j;\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \end{aligned} \tag{3}$$

We denote the MLEs, obtained through maximization of Eqs. (1)–(3) using the incident, prevalent and combined cohort data, respectively, by $\widehat{\boldsymbol{\theta}}_I$, $\widehat{\boldsymbol{\theta}}_P$ and $\widehat{\boldsymbol{\theta}}_C$. As the sample data in the incident and prevalent cohorts arise from different sampling schemes, it follows immediately that the three proposed estimators are distinct. For pure incident and pure (identically distributed) prevalent cohort failure time data, it has been shown that the respective MLEs for $\boldsymbol{\theta}$ are both consistent and asymptotically Normally distributed [8, 13, 25]. However, in the combined cohort, the data do not arise from a single sampling scheme and are not identically distributed. We extend the "pure cohort asymptotic properties" of the parametric MLE to the combined cohort case through Theorem 1.

Theorem 1. *Let the underlying absolutely continuous failure time distribution function be given by* $F(\cdot\,;\boldsymbol{\theta}_0)$ *where* $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}^k$. *Let* $\widehat{\boldsymbol{\theta}}_C$ *denote the MLE of* $\boldsymbol{\theta}_0$ *obtained by maximization of Equation (3). Let* $\overline{\Gamma}(\boldsymbol{\theta}_0)$ *be some positive definite matrix. Then, as* $n+m\to\infty$

(1) $\widehat{\boldsymbol{\theta}}_C \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0$
(2) $\sqrt{n+m}(\widehat{\boldsymbol{\theta}}_C - \boldsymbol{\theta}_0) \xrightarrow{\mathscr{D}} \mathscr{N}(0, \overline{\Gamma}^{-1}(\boldsymbol{\theta}_0))$

Proof. Refer to the Supplementary materials. ∎

Remark: Following [26] and appealing to the law of large numbers,

$$\begin{aligned} \overline{\Gamma}(\boldsymbol{\theta}_0) = &-\alpha\mathbb{E}\left(\frac{d^2}{d\boldsymbol{\theta}_0^2}\left[\delta_i \log(f_U(X_i;\boldsymbol{\theta}_0)) + (1-\delta_i)\log(S_U(X_i;\boldsymbol{\theta}_0))\right]\right) \\ &-(1-\alpha)\mathbb{E}\left(\frac{d^2}{d\boldsymbol{\theta}_0^2}\left[\delta_j^* \log(f_U(Y_j;\boldsymbol{\theta}_0)) + \left(1-\delta_j^*\right)\log(S_U(Y_j;\boldsymbol{\theta}_0))\right]\right) \\ &+(1-\alpha)\left(\frac{d^2}{d\boldsymbol{\theta}_0^2}\log(\mu(\boldsymbol{\theta}_0))\right). \end{aligned} \tag{4}$$

where α is the limiting proportion of the number of incident cases to the total size of the combined cohort. An empirical estimator for the asymptotic covariance matrix (i.e. the observed Fisher information) is then given by

$$\begin{aligned}\widehat{\ddot{\ell}(\boldsymbol{\theta})} = &-\widehat{\alpha}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d^2}{d\boldsymbol{\theta}^2}\left[\delta_i \log\left(f_U(X_i;\boldsymbol{\theta})\right) + (1-\delta_i)\log\left(S_U(X_i;\boldsymbol{\theta})\right)\right]\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_C}\right)\\ &-(1-\widehat{\alpha})\frac{1}{m}\sum_{j=1}^{m}\left(\frac{d^2}{d\boldsymbol{\theta}^2}\left[\delta_j^* \log\left(f_U(Y_j;\boldsymbol{\theta})\right) + \left(1-\delta_j^*\right)\log\left(S_U(Y_j;\boldsymbol{\theta})\right)\right]\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_C}\right)\\ &+(1-\widehat{\alpha})\left(\frac{d^2}{d\boldsymbol{\theta}^2}\left[\log\left(\mu(\boldsymbol{\theta})\right)\right]\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_C}\right).\end{aligned} \tag{5}$$

where $\widehat{\alpha} = \dfrac{n}{n+m}$.

4 Simulations

We used simulated data sets to evaluate the performance of our parametric MLE, controlling the sample sizes and parameter values. In each simulation, we sampled n observations from an incident cohort and m observations from a prevalent cohort allowing n and m to vary. For the incident cohort, we sampled n pairs of Weibull distributed failure times and either constant or Exponentially distributed censoring times to correspond to subjects that are either followed for a fixed period of time after enrollment into the study or subjects that are lost to follow-up after enrollment, respectively. In the simulation studies discussed below, we specify the type of censoring that was assumed for the data generating procedure. For each pair, we recorded the minimum of the sampled failure and censoring times and whether the time was observed as a failure. For the prevalent cohort, we sampled a single onset time from a Uniform distribution with support in the interval (0,100) and then sampled a Weibull failure time which was added to the sampled onset time. If the resulting sum was greater than 100, both the sampled onset time and failure time were retained, otherwise, both were discarded. Weibull parameter values were chosen such that the implicit right-truncation of the failure times at 100 was negligible. This procedure was repeated until m pairs of (onset, failure time) data were obtained. For each sampled pair, we censored the forward recurrence time indepenently by either a fixed constant (corresponding to a fixed follow-up period) or by a random Exponentially distributed censoring time (corresponding to potential loss to follow-up). We recorded the triples made up of the onset time, the minimum of the forward censoring and forward failure time and whether the observation was a failure time. From the observed triples, the failure/censoring lengths were calculated by summing the backward recurrence times (i.e. 100– sampled onset times) and the forward failure/censoring times. We obtained a simulated combined cohort by concatenating the incident and prevalent cohort data sets as well as setting an additional variable to indicate whether the datum entry was an incident or prevalent cohort observation. These simulations were used to highlight, empirically, the following three assertions about the combined cohort parametric MLE:

(1) In a meta-analysis, perhaps obviously, when individual subject level data are available from two independent cohort studies of different types, the combined cohort parametric MLE will have a smaller standard error than the individual cohort parametric MLEs.
(2) In a combined cohort study with fixed total sample size and short follow-up, resulting in one cohort being heavily censored and the other being lightly to moderately censored, the standard error of the parametric MLE using data from both types of cohort will be smaller than the standard error of the same estimator when applied to data retrieved from only a single cohort of the same sample size.
(3) The combined cohort estimator may be robust against misspecification of the parametric model.

We consider the simulation results pertaining to each of the above statements in order.

Since there is no analytical method for comparing the relative magnitudes of the asymptotic covariance matrices of the individual and combined cohort parametric estimators, we compared the efficiency of the MLEs empirically using simulated individual cohort failure time data with sample sizes of 250 and 500 over 1000 simulation runs. In the combined cohort case, for each sample size, we used all available incident and

prevalent cohort data (i.e. 250/250 or 500/500). We used different Exponential censoring distributions to vary the censoring proportion between 30, 50 and 70%. We estimated the Weibull distribution parameters for each of the 1000 simulation runs and computed the sample covariance matrix of the estimates. We then computed the determinant of the sample covariance matrix to obtain the generalized variance of the parameter pair estimators [27]. We allowed the failure time parameters to vary to allow for increasing or decreasing hazard functions. We list the ratios of the generalized variances of the combined cohort estimators to the generalized variances of the individual cohort estimators in Table 1. Based on the ratios in Table 1, combining data from both the incident and prevalent cohorts yields a clear improvement in the magnitude of the generalized variance. Since in the incident cohort we make the standard assumption of non-informative censoring, we find that as the censoring percentage increases, the ratio of generalized variances of the combined cohort parametric MLE to the incident cohort parametric MLE decreases. In contrast, because censoring is informative in the prevalent cohort, we find that as the censoring percentage increases, the ratio of the generalized variances of the combined cohort parametric MLE to the prevalent cohort parametric MLE increases. These results show that the combined cohort parametric estimator inherits the non-informative or informative censoring properties of the incident or prevalent cohort cases, respectively. Similar results are presented for the case when the censoring percentages are allowed to vary between cohorts (see Tables 1 and 2 in the Supplementary materials).

When it is feasible to include cases from both prevalent and incident cohort studies with fixed follow-up periods, consideration must be given to the optimal proportions of each cohort type. We set a fixed grand sample size of 500 observations and varied the prevalent/incident subsample sizes in increments of 50 observations each. We considered the setting in which all enrolled subjects had the same follow-up period measured either from prevalence day or from the time of enrollment for the prevalent or incident cases, respectively. This assumption yielded failure time data that were only administratively censored by the end date of the follow-up period. The parameters of the failure time distribution were set to allow for either increasing or decreasing hazard functions, respectively. Under the assumption of an increasing hazard function, approximately 70% of the incident cases were censored. In contrast, the prevalent cases were only lightly to moderately censored. However, under the assumption of a decreasing hazard function, approximately 70% of the prevalent cases were censored with incident cases being lightly to moderately censored. We computed the generalized variances of the parameter estimates over 1000 simulation runs and plotted the ratios of the generalized variances of the MLE obtained from combined cohorts of size 500 relative to the generalized variances obtained from a pure incident cohort of size 500 in Figure 2. From the convexity of the ratio plots, we find that when the incident cohort is heavily censored (increasing hazard setting) or when the prevalent cohort is heavily censored (decreasing hazard setting) there appears to be an optimal proportion of incident to prevalent cohort cases. For example, in the increasing hazard setting, the optimal proportion is roughly 100 incident cases to 400 prevalent cases. When the follow-up periods for the individual cohorts are

Table 1: Ratios of the generalized variance for the maximum likelihood parameter estimates of the combined cohort parametric estimators relative to the individual cohort parametric estimators over 1000 simulation runs for varying sample sizes. Failure times were generated from a Weibull distribution (increasing/decreasing hazard) with random censoring times generated from an Exponential distribution.

			Censoring percentage		
Failure time distribution	**Sample sizes**	**Cohort ratio type**	**30%**	**50%**	**70%**
Weibull (2,2)	500/250	Combined/Incident	0.627	0.516	0.382
(Increasing hazard)		Combined/Prevalent	0.901	0.915	1.00
	1000/500	Combined/Incident	0.694	0.568	0.378
		Combined/Prevalent	0.853	0.842	0.895
Weibull (0.5, 1)	500/250	Combined/Incident	0.217	0.123	0.0536
(Decreasing hazard)		Combined/Prevalent	0.482	0.482	0.525
	1000/500	Combined/Incident	0.237	0.135	0.0606
		Combined/Prevalent	0.534	0.567	0.631

Table 2: Mean supnorm distances of the estimated survival curves from the true survival curves for combined cohort data using parametric maximum likelihood estimates over 1000 simulation runs. The individual cohort sample sizes were 500 each where the underlying failure times were i.i.d. according to a Weibull distribution with an Exponential censoring distribution. The three combined cohort models assumed either Weibull (Comb.1), Gamma (Comb.2) or log-Normal (Comb.3) failure time distributions.

	Censoring percentage		
Estimator type	**30%**	**50%**	**70%**
Weibull shape 2.0 scale 2.0 (increasing hazard)			
Comb.1	0.0104	0.0112	0.0125
Comb.2	0.0355	0.0370	0.0408
Comb.3	0.0820	0.0863	0.0963
Weibull shape 0.5 scale 1.0 (decreasing hazard)			
Comb.1	0.0103	0.0108	0.0118
Comb.2	0.0803	0.0869	0.0993
Comb.3	0.145	0.156	0.179
Weibull shape 1.0 scale 2.0 (constant hazard)			
Comb.1	0.0101	0.0107	0.0118
Comb.2	0.0101	0.0106	0.0114
Comb.3	0.104	0.110	0.126

different, resulting in equal censoring percentages, the optimal proportion of incident to prevalent cases appears to be roughly one half (for further details, see Figure 1 of the Supplementary materials). Similarly shaped ratio curves were obtained when the censoring times were randomly generated and not fixed constants. In general, the convexity of the ratio curves show that there are improvements to the overall parametric estimation procedure, aside from the obvious increase in total sample size (as in Table 1), when combining independent samples of incident and prevalent cohort failure time data.

To assess the impact of a misspecified parametric model, we generated combined cohort samples of size 1000 (500/500 prevalent/incident cases) for which the underlying failure times were distributed according to a Weibull distribution with either increasing or decreasing hazard. We fit the combined cohort parametric MLE assuming that the failure times arose from either Weibull, Gamma and log-Normal distributions. For each of these parametric models, we computed the absolute maximum distance between the estimated survivor function and true survivor function for which we averaged the computed distances over 1000 simulation runs, respectively. The simulation results in Table 2 suggest that the MLE in the combined cohort accommodates a

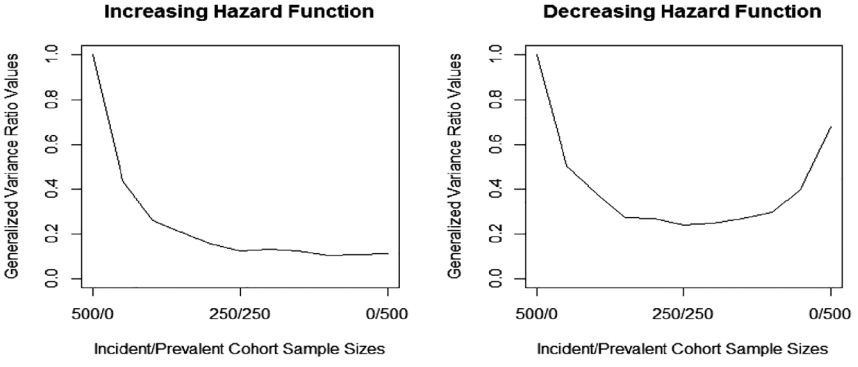

Figure 2: Ratios of the generalized variances of the maximum likelihood parameter estimates for combined cohort data relative to pure incident cohort data over 1000 simulation runs for varying individual cohort sample sizes. Failure times were generated from a Weibull distribution (increasing/decreasing hazard) with administrative incident/forward censoring times for the individual cohorts, respectively.

misspecified Gamma model quite well but not so well for a log-Normal model. In practice, as will be seen in Section 5, the parametric model should be selected with care.

5 Application

Hospital stay durations may be used as a measure of a hospital's efficiency in treating patients after accounting for 'case-mix' or variation in types and severity of illnesses. Hence, they could be used to direct the future management decisions of various hospital services by hospital administrators or policy makers [28]. We drew on duration-of-stay records from a Montreal area hospital that provided data under the Population Health Records platform project at McGill University [29]. This platform links data from administrative sources, clinical records as well as responses from surveys and then provides a system to access this data in an attempt to monitor the health of a specific population. Our data consisted of a subset of duration-of-stay records that had been collected over a period of approximately 17 years. For reasons of confidentiality, the hospital cannot be identified. For brevity, we shall refer to the hospital as the "PopHR hospital". The true admission/discharge dates were anonymized on a day-length integer scale where even the start date of the 17 year observation window was not divulged. Using a small subset of this data, our goal was to estimate the distribution of the duration of stays in the PopHR hospital. An individual's duration was measured from the date of admission to the date of discharge or death. Admissions to the hospital based on scheduled surgeries, childbirth or between hospital/ward transfers were not included in the data. Admission and discharge dates that were less than 24 h apart were also not included. We considered two observation windows of approximately 15 days in length, measured from days 800–815 (165 incident/69 prevalent cases) and 1165–1180 (151 incident/84 prevalent cases), respectively. In the earlier window, approximately 40 and 44% of the incident and prevalent cohort subject failure times were censored, respectively. Similarly, in the later window, approximately 38 and 34% of the incident and prevalent cohort subject failure times were censored, respectively.

Using the earlier window of observations as an independent training data set, we fit Weibull, Gamma and log-Normal parametric models. As we observed the admission dates for the entire 17 year observation window, we were able to check for uniformity in the onset dates using graphical methods. We found no reason to doubt the stationarity assumption of the onset dates. We remark that it is even possible to check for stationarity using prevalent cohort data, where the underlying onset process is not fully observed (see [30, 31]). Using the separate incident and prevalent cohort data, we compared the parametric survival function estimates based on the Weibull, Gamma and log-Normal distributions, to their respective non-parametric estimates by calculating the supnorm distances between them. We selected the log-Normal parametric model as it yielded small supnorm distances for the separate cohorts. Using the log-Normal distribution, we then found separate cohort estimates from the data observed in the second observation window. The (parametrically and non-parametrically) estimated survivor functions from the individual cohorts are displayed, along with the parametrically estimated survivor function obtained by combining the separate cohort data, in Figure 3. Since the observation window was restricted to 15 days, the Kaplan-Meier estimate in the left panel of Figure 3 is not defined past 15 days. In contrast, the NPMLE of the survival function using only the prevalent cohort data is defined past the 15 day mark as the data consists of longer time durations which were cross-sectionally sampled. Under the log-Normal parametric model, we found that the median time from admission to discharge was approximately 4–5 days using the incident cohort data, and approximately 5–7 days using either the prevalent or combined cohort data. The parametrically estimated survivor function using the combined cohort data appeared to incorporate the features of the individual cohort estimates by fitting closely to the estimated incident cohort survivor curve for shorter times (<8 days) and then fitting closely to the estimated prevalent cohort survivor curve for longer times (>20 days). The combined cohort estimated survivor curve tends to always be between the individual cohort estimated survivor curves.

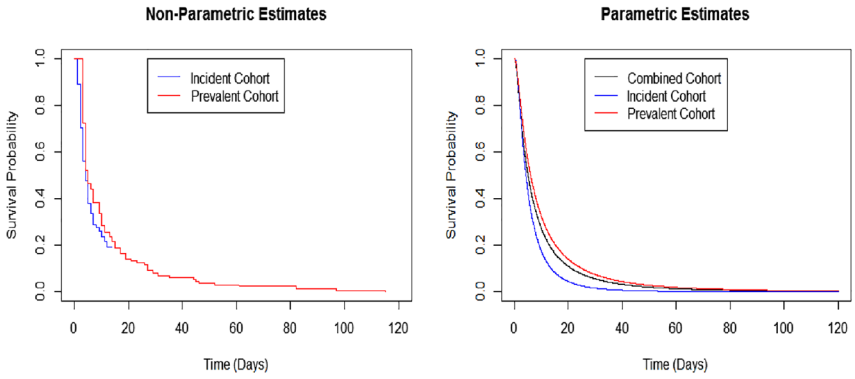

Figure 3: Estimated survival functions of hospital stay durations measured from admission to discharge for a PopHR hospital. The left panel displays the individual cohort non-parametric estimates (i.e. Kaplan-Meier and length-biased right-censored NPMLE). The right-panel displays the estimated survivor curves using an underlying log-Normal parametric distribution.

6 Discussion

We use a parametric model in the combined survival data setting of this article because the asymptotic properties of the NPMLE of the survivor function remain unestablished at this time. Moreover, we believe that a careful choice of parametric model can provide a good practical alternative to non-parametric estimation. Combining incident and prevalent cohort data can have benefits in at least four different ways: (i) Even relatively few incident cases can considerably enhance the inference when added to a prevalent cohort that has been followed up (see Table 1). (ii) Conversely, adding a few cases from a prevalent cohort study with follow-up can considerably enhance the inference from a pure incident cohort, particularly when study follow-up is short (see Figure 2). (iii) In recent years, many funding agencies and journals have required researchers to make their subject-level data widely available. Consequently, individual participant data meta-analyses that are able to use full study data are becoming more common [32]. Such meta analyses of survival data could be based on the union of data from incident cohort studies and data from prevalent cohort studies with follow-up. (iv) In a single study, although the original intent may not have been to combine the two types of data, it is clear that increasing the sample size by combining these data (if available) should increase the efficiency of the parameter estimators. For a study in which both types of data were collected and where no single unified analysis was carried out, see [33].

We were able to show empirically that, under certain parameter and censoring combinations, the ratios of the generalized variances of the combined cohort parametric estimator to the generalized variance derived from a pure incident cohort, was convex. This suggests that the optimal proportion of prevalent-to-incident cases occurs at the minimum. However, this ratio depends on the very parameters one is attempting to estimate. Therefore, one would need rough parameter estimates when designing a future study with intent to use both prevalent and incident cohorts.

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: The first author was supported by a Natural Sciences and Engineering Research Council of Canada PGSD-3 award. David Stephens acknowledges the support of a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).
Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

1. Humbert M, Sitbon O, Yaïci A, Montani D, O'Callaghan DS, Jaïs X, et al. On behalf of the French Pulmonary Arterial Hypertension Network. Survival in incident and prevalent cohorts of patients with pulmonary arterial hypertension. Eur Respir J 2010;36: 549–55.

2. Lee CH, Ning J, Kryscio RJ, Shen Y. Analysis of combined incident and prevalent cohort data under a proportional mean residual life model. Stat Med 2019;38:2103–14.
3. Daepp MIG, Hamilton MJ, West GB, Bettencourt LMA. The mortality of companies. J R Soc Interface 2015;12. https://doi.org/10.1098/rsif.2015.0120.
4. Groothuis PA, Hill JR. Pay discrimination, exit discrimination or both? Another look at an old issue using NBA data. J Sports Econ 2011;14:171–85.
5. Welch S.M. Nonparametric estimates of the duration of welfare spells. Econ Lett 1998;60:217–21.
6. Andersen PK, Borgan Ø, Gill RD, Keiding N. Statistical Models Based on Counting Processes. New York: Springer-Verlag; 1993.
7. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–81.
8. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data, 2nd ed. New York: Wiley; 1980.
9. Tsai W-Y, Jewell NP, Wang M-C. A note on the product-limit estimator under right censoring and left truncation. Biometrika 1987;74:883–6.
10. Wang M-C. Nonparametric estimation from cross-sectional survival data. J Am Stat Assoc 1991;86:130–43.
11. Zhou Y. A note on the TJW product-limit estimator for truncated and censored data. Stat Probab Lett 1996;26:381–7.
12. Asgharian M, M'Lan CE, Wolfson DB. Length-biased sampling with right censoring: an unconditional approach. J Am Stat Assoc 2002;97:201–9.
13. Bergeron P-J, Asgharian M, Wolfson DB. Covariate bias induced by length-biased sampling of failure times. J Am Stat Assoc 2008;103:737–42.
14. Pan W, Chappell R. A nonparametric estimator of survival functions for arbitrarily truncated and censored data. Lifetime Data Anal 1998;4:187–202.
15. Wolfson DB, Best AF, Addona V, Wolfson J, Gadalla SM. Benefits of combining prevalent and incident cohorts: an application to myotonic dystrophy. Stat Methods Med Res 2019;28:3333–45.
16. McVittie JH, Wolfson DB, Stephens DA. A note on the applicability of the standard non-parametric maximum likelihood estimator for combined incident and prevalent cohort data. Stat 2020;9. https://doi.org/10.1002/sta4.280.
17. Vardi Y. Nonparametric estimation in the presence of length bias. Ann Stat 1982;10:616–20.
18. Vardi Y. Empirical distributions in selection bias models. Ann Stat 1985;13:178–203.
19. Gill RD, Vardi Y, Wellner JA. Large sample theory of empirical distributions in biased sampling models. Ann Stat 1988;16:1069–112.
20. Miller RG, Jr. What price Kaplan-Meier?. Biometrics 1983;39:1077–81.
21. Laslett GM. The survival curve under monotone density constraints with application to two-dimensional line segment processes. Biometrika 1982;69:153–60.
22. van der Laan MJ. Efficiency of the NPMLE in the line-segment problem. Scand J Stat 1996;23:527–50.
23. Wijers BJ. Consistent non-parametric estimation for a one-dimensional line segment process observed in an interval. Scand J Stat 1995;22:335–60.
24. Saarela O, Kulathinal S, Karvanen J. Joint analysis of prevalence and incidence data using conditional likelihood. Biostatistics 2009;10:575–87.
25. Ibragimov IA, Has'minskii RZ. Statistical Estimation: Asymptotic Theory. Springer-Verlag; 1981.
26. Hoadley B. Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. Ann Math Stat 1991;42:1977–91.
27. Wilks SS. Multidimensional statistical scatter. In: Olkin I, Ghurye S, Hoeffding W, Madow W, Mann H, editors Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling Stanford University Press; 1960. pp. 486–503.
28. Verma A, Rochefort C, Powell G, Buckeridge D. Hospital readmissions and the day of the week. J Health Serv Res Pol 2018;23:21–7.
29. Shaban-Nejad A, Lavinge M, Okhmatovskaia A, Buckeridge DL. PopHR: a knowledge-based platform to support integration, analysis, and visualization of population health data. Ann N Y Acad Sci 2017;1387:44–53.
30. Addona V, Atherton J, Wolfson DB. Testing the assumptions for the analysis of survival data arising from a prevalent cohort study with follow-up. Int J Biostat 2012;8. https://doi.org/10.1515/1557-4679.1419.
31. Addona V, Wolfson DB. A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up. Lifetime Data Anal 2006;12:267–84.
32. Tierney JF, Pignon J-P, Gueffyier F, Clarke M, Askie L, Vale CL, et al. On behalf of the Cochrane IPD Meta-analysis Methods Group. How individual participant data meta-analyses have influenced trial design, conduct, and analysis. J Clin Epidemiol 2015;68:1325–35.
33. Wolfson C, Wolfson DB, Asgharian M, M'Lan CE, Østbye T, Rockwood K, et al. For the Clinical Progression of Dementia Study Group. A reevaluation of the duration of survival after the onset of dementia. N Engl J Med 2001;344:1111–16.

Supplementary material: The online version of this article offers supplementary material (https://doi.org/10.1515/ijb-2020-0042).