BMC Medical Research Methodology

RESEARCH Open Access

Stacked survival models for residual lifetime data

James H. McVittie[1*], David B. Wolfson[1], Vittorio Addona[2] and Zhaoheng Li[2]

Abstract

When modelling the survival distribution of a disease for which the symptomatic progression of the associated condition is insidious, it is not always clear how to measure the failure/censoring times from some true date of disease onset. In a prevalent cohort study with follow-up, one approach for removing any potential influence from the uncertainty in the measurement of the true onset dates is through the utilization of only the residual lifetimes. As the residual lifetimes are measured from a well-defined screening date (prevalence day) to failure/censoring, these observed time durations are essentially error free. Using residual lifetime data, the nonparametric maximum likelihood estimator (NPMLE) may be used to estimate the underlying survival function. However, the resulting estimator can yield exceptionally wide confidence intervals. Alternatively, while parametric maximum likelihood estimation can yield narrower confidence intervals, it may not be robust to model misspecification. Using only right-censored residual lifetime data, we propose a stacking procedure to overcome the non-robustness of model misspecification; our proposed estimator comprises a linear combination of individual nonparametric/parametric survival function estimators, with optimal stacking weights obtained by minimizing a Brier Score loss function.

Keywords: Survival analysis, Residual lifetime data, Nonparametric estimation, Stacking

Introduction

The Canadian Study of Health and Aging (CSHA) was a nation-wide study whose primary goal was to determine the prevalence of dementia in five different regions in Canada [1, 2]. In 1991, at the first stage of the study (CSHA-1), approximately 10,000 individuals over the age of 65 were screened for various types of dementia. A total of 823 participants were classified at CSHA-1 as having either probable Alzheimer's disease, possible Alzheimer's disease or vascular dementia. They were followed for a subsequent five years until the second stage of the study in 1996 (CSHA-2). Death dates were recorded for those who died between 1991 and 1996 together with the censoring dates of those who were lost to follow-up or survived until 1996. The onset dates of the participants who screened positive at CSHA-1 were retrospectively reported through the recollections of their caregivers. The observed (right-censored) survival times were the durations of time from reported onset to failure/censoring. The resulting failure times were therefore considered to be left-truncated and right-censored as typically occurs in a prevalent cohort study with follow-up [3]. Suppose, further, it is assumed that the underlying process that defines all of the onset dates, including those not associated with the observed prevalent cohort, is a stationary Poisson process. Then we shall say that our inference is carried out "under the stationarity assumption" [4], an assumption that is crucial for the methods proposed in this article.

Now, due to the onset date recording protocols in the CSHA as well as the insidious symptomatic onset of dementia, the true failure/censoring times were almost surely measured with error. Under the stationarity assumption, in more general prevalent cohort studies with follow-up, uncertainty in the onset dates can be accounted for in at least two ways. First, under the assumption that the failure time distribution is defined parametrically,

*Correspondence: james.mcvittie@mail.mcgill.ca
[1]Department of Mathematics and Statistics, McGill University, Montreal, CA
Full list of author information is available at the end of the article

McVittie et al. defined an adjusted classical measurement error model for the reported onset dates, to derive maximum likelihood estimators for the unknown failure time distribution parameters [5].

A second approach, under stationarity, is to discard the information contained in the uncertain onset times by using only the residual lifetimes for estimation of the survival distribution. The residual lifetimes extend from the date of screening to the date of failure/censoring. As the residual lifetimes are not dependent on the onset dates, these durations are error free. It should be noted that without stationarity it is impossible to make inference about the failure time distribution based only on observation of the residual lifetimes. In order to review the literature on this approach, we use notation which will be more systematically introduced in Section 2. Let $S_U(\cdot)$ be the underlying (unbiased) survival function (the estimation target) and let μ be its mean. Let $f_{\text{res}}(\cdot)$ be the residual lifetime probability density function (pdf). Then, this setting can be regarded as equivalent to the scenario in which the residual lifetimes (with pdf $f_{\text{res}}(\cdot)$) of a stationary renewal process, with interarrival time survivor function $S_U(\cdot)$, are the observations [6]. Exploiting this equivalence, and a well known property of stationary renewal processes [7–9], it can be shown that,

$$\begin{aligned} f_{\text{res}}(\cdot) &= S_U(\cdot)/\mu \\ &= S_U(\cdot) f_{\text{res}}(0) \end{aligned} \tag{1}$$

It follows from (1), that if S_U is parametrically defined, its maximum likelihood estimator (MLE) can be found by finding the MLEs of the parameters that define f_{res}; the (possibly censored) residual lifetimes can be used to construct the likelihood function [7, 8]. Non-parametric estimation of S_U, however, requires much more care and has been the subject of much research both in the context of renewal processes and survival analysis. The difficulty arises from the instability in $\hat{S}_U(\cdot) = \hat{f}_{\text{res}}(\cdot)/\hat{f}_{\text{res}}(0)$, due to its dependence on the boundary-point estimator $\hat{f}_{\text{res}}(0)$. For uncensored data, and recognizing that by (1), $f_{\text{res}}(u)$ is non-increasing in u, Grenander showed that the NPMLE of $F_U(\cdot) = 1 - S_U(\cdot)$, is the least concave majorant of the empirical distribution function [10]. Woodroofe and Sun proposed a penalized maximum likelihood procedure to consistently estimate the residual lifetime density function at the boundary [11]. For right-censored residual lifetime data, the least concave majorant of the cumulative distribution function estimated using the Kaplan-Meier estimator, in place of the empirical survival function, is no longer the NPMLE [12, 13]. Denby and Vardi proposed an iterative EM based algorithm to determine the NPMLE using right-censored residual lifetime data [14]. They also proposed a "corrected" NPMLE to account for the bad behavior of the NPMLE at times close to zero. Huang and Zhang remark that in the right-censored setting, the boundary point estimator using the NPMLE is both unstable and inconsistent [15]. Although an approach combining the Denby and Vardi algorithm with the penalization procedure of Woodroofe and Sun has been alluded to in the literature, it has not been formally described and compared to other methodologies. Recently, Westling and Carone surveyed the asymptotic properties of nonparametric survival function estimators subject to monotonicity constraints [9]. Using current duration data, Keiding studied the behaviour of the corrected NPMLE and associated parametric models for modelling the time to pregnancy [7, 8]. He remarked that the confidence intervals obtained from the corrected NPMLE were wide due to the unstable boundary estimation problem at time $t = 0$ [7, 8]. Using data collected from the CSHA, this phenomenon is most evident in the nonparametric maximum likelihood survival function estimates for subjects with possible Alzheimer's disease and vascular dementia (see Fig. 1).

There is another disadvantageous feature of the NPMLE of S_U: If the study follow-up period is short, the NPMLE is unlikely to estimate S_U well, beyond the largest observation in the sample. This feature is demonstrated in the survival curve estimates of Fig. 1 as all three curves drop to near 0 at approximately 60 months. When we started on this research we anticipated using an estimator, based on the (possibly censored) residual lifetimes, that includes both the NPMLE and several suitable parametric estimators, thereby counteracting the drawbacks of each of these two types of estimator. We reasoned that in the context of standard right-censored failure time data, Wey et al. had proposed a stacking procedure which successfully combines non/semi-parametric and parametric survival function estimators into a single estimator of the underlying survival distribution [16, 17]. To our surprise, however, particularly when applied to right-censored residual lifetime data with short follow-up, we found that there was little advantage to including the NPMLE in the stacking procedure; that is, the NPMLE received very little weight.

We adapt the stacking approach of Wey et al. to estimate S_U using right-censored residual lifetime data. Our goals are to: (i) enable estimation of the survival function past the last observed failure/censoring time when follow-up is short, (ii) provide an estimation procedure which is robust to model misspecification and (iii) reduce the width of the confidence intervals that would be obtained from the NPMLE alone. In Section 2, we introduce notation for prevalent cohort studies with follow-up and specify how the procedure of Wey et al. is modified for residual lifetime data. In Section 3, we use simulated failure time data to examine the performance of the stacked estimator against estimators based on individual models. We apply

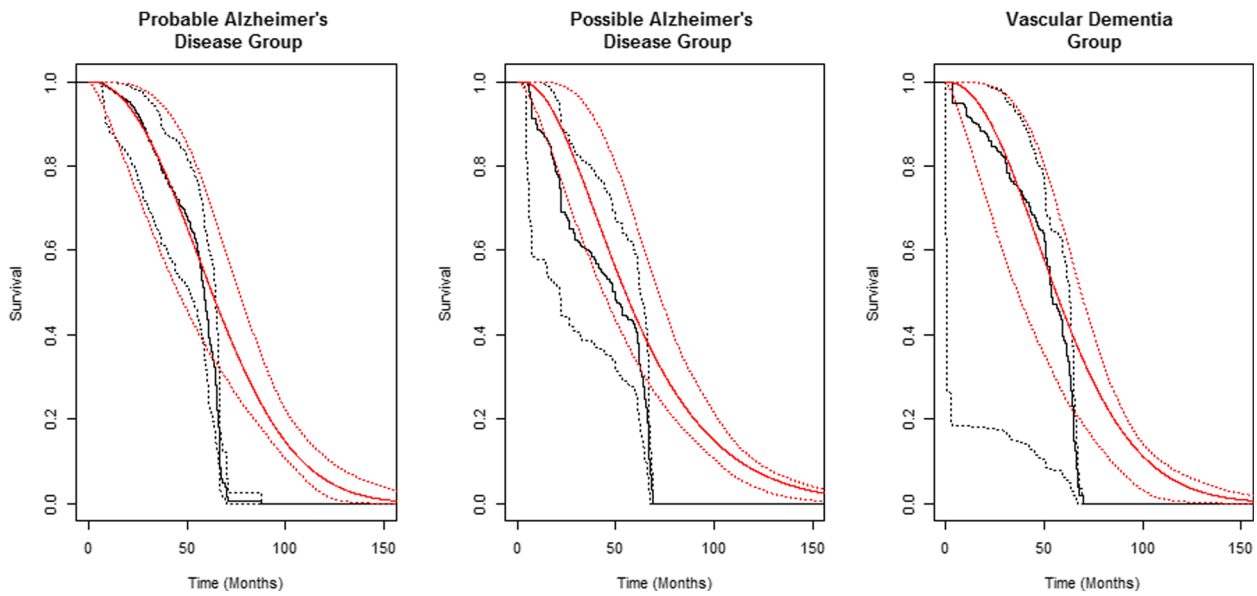

Fig. 1 Corrected non-parametric maximum likelihood survival function estimates with 95% bootstrapped pointwise confidence limits (black) for varying dementia subgroups in the Canadian Study of Health and Aging along with stacked survival function estimates with 95% bootstrapped pointwise confidence limits (red)

our stacking methodology to the CSHA data set in Section 4 and provide some concluding remarks in Section 5.

Notation and methodology

Let (O, T) denote the random variable pair consisting, respectively, of a generic onset date drawn from a stationary Poisson process and a generic failure time with survival function $S_U(\cdot)$ where O is independent of T. Let the fixed constant R denote the screening date at which the prevalent cohort is determined and which we define as "prevalence day". The prevalent cohort then consists of subjects with (onset, failure time) pairs such that $O < R$ and $O + T > R$. Let C denote the censoring time (measured from prevalence day) with cumulative distribution $G(\cdot)$ corresponding to subjects who are either lost to follow-up or have not failed by the end of the study (administratively censored). The full prevalent cohort data then comprises the triples $\{(A_i, Y_i, \delta_i) = (R - O_i, \min(T_i, R - O_i + C_i), 1_{\{T_i < R - O_i + C_i\}}) : T_i > R - O_i, i = 1, 2, ..., n\}$. As the residual lifetimes consist only of the failure/censoring times measured from prevalence day and their associated indicator functions, they are given by the pairs $\{(V_i, \delta_i) = (\min(T_i - (R - O_i), C_i), 1_{\{T_i < R - O_i + C_i\}}) : T_i > R - O_i, i = 1, 2, ..., n\}$. For a depiction of residual lifetime data, see Fig. 2.

For convenience, we repeat Eq. 1, now numbered (2):

$$f_{\text{res}}(\cdot) = \frac{S_U(\cdot)}{\mu} \tag{2}$$

where $\mu = \mathbb{E}(T)$. By evaluating the pdf f_{res}, at time $t = 0$, and utilizing the property that $S_U(0) = 1$, from Eq. 2, it follows that $\mu = \frac{1}{f_{\text{res}}(0)}$, and hence $f_{\text{res}}(\cdot) = S_U(\cdot) f_{\text{res}}(0)$. This suggests

$$\hat{S}_U(t) = \frac{\hat{f}_{\text{res}}(t)}{\hat{f}_{\text{res}}(0)} \tag{3}$$

as a plug-in estimator for S_U. When $S_U(\cdot; \boldsymbol{\theta})$ is defined parametrically, for some unknown p-dimensional parameter $\boldsymbol{\theta}$, the likelihood function is given by [7]:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(\frac{S_U(v_i; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \right)^{\delta_i} \left(\int_{v_i}^{\infty} \frac{S_U(x; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} dx \right)^{1-\delta_i} \tag{4}$$

Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$, obtained from (4). Although the parametric maximum likelihood estimator $S_U(\cdot; \hat{\boldsymbol{\theta}})$ is model-dependent and possibly biased, it has a smaller variance than does its non-parametric counterpart in (3).

One approach which combines nonparametric and parametric estimators is through the machine learning procedure known as *stacking*. A stacked survival function estimator is a weighted linear combination of sub-model survival function estimators for which the optimal weights are determined through optimization of a particular loss function. Here, we consider the approach of Wey et al. which allows for right-censoring of the data [16]. We begin by proposing $m - 1$ parametric models for $f_{\text{res}}(\cdot)$: $f_{\text{res},1}(\cdot; \boldsymbol{\theta}_1), f_{\text{res},2}(\cdot; \boldsymbol{\theta}_2), ..., f_{\text{res},m-1}(\cdot; \boldsymbol{\theta}_{m-1})$ for $m \geq 2$. Let $\hat{f}_{\text{res},1}(\cdot), \hat{f}_{\text{res},2}(\cdot), ..., \hat{f}_{\text{res},m}(\cdot) = f_{\text{res},1}(\cdot; \hat{\boldsymbol{\theta}}_1), f_{\text{res},2}(\cdot; \hat{\boldsymbol{\theta}}_2), ...,$

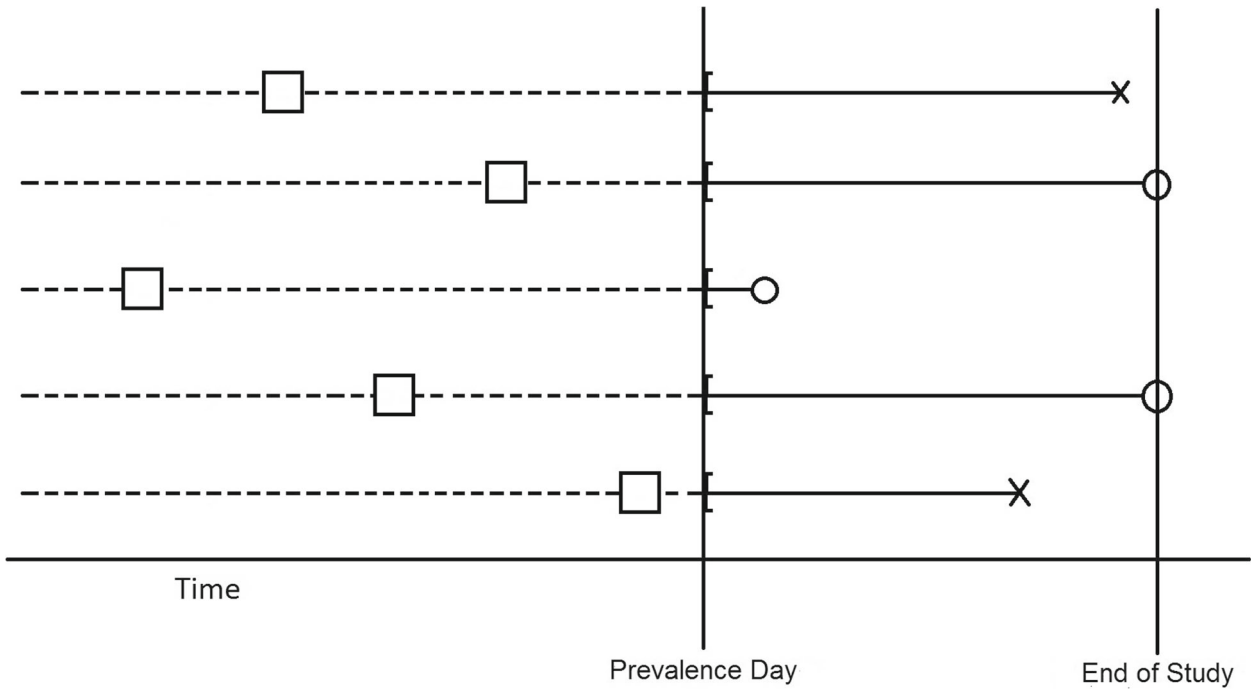

Fig. 2 A depiction of a sample of right-censored residual lifetime data. The open circles represent the calendar dates of censoring and the crosses represent the calendar dates of failure. Brackets represent the earliest confirmed time for the start of the failure/censoring time durations. Dashed lines represent uncertainty in the measurement of the underlying onset dates (represented by open squares)

$f_{\text{res},m-1}(\cdot;\hat{\boldsymbol{\theta}}_{m-1}), \hat{f}_{\text{res},m}(\cdot)$ be m estimators of $f_{\text{res}}(\cdot)$, where $\hat{f}_{\text{res},m}(\cdot)$ is the (non-increasing) corrected NPMLE defined by Denby and Vardi, and $\hat{f}_{\text{res},i}(\cdot)$ is the parametrically defined MLE for $i = 1, 2, ..., m-1$ [14]. We define a stacked density estimator of $f_{\text{res}}(\cdot)$ as:

$$\hat{f}_{\text{res,stack}}(\cdot) = \alpha_1 \hat{f}_{\text{res},1}(\cdot) + \alpha_2 \hat{f}_{\text{res},2}(\cdot) + ... + \alpha_m \hat{f}_{\text{res},m}(\cdot),$$

where $\alpha_i \in [0,1]$ and $\sum_{i=1}^{m} \alpha_i = 1$. Since each $\hat{f}_{\text{res},i}(\cdot)$ is non-increasing, $\hat{f}_{\text{res,stack}}$ is also non-increasing. By integrating the linear combination of pdf estimators, we obtain a stacked residual lifetime survival function estimator given by:

$$\hat{S}_{\text{res,stack}}(\cdot) = \alpha_1 \hat{S}_{\text{res},1}(\cdot) + \alpha_2 \hat{S}_{\text{res},2}(\cdot) + ... + \alpha_m \hat{S}_{\text{res},m}(\cdot) \quad (5)$$

The general idea is to find the optimal weights $\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_m$ by minimizing an objective function of $\hat{S}_{\text{res,stack}}(\cdot)$. Specifically, let $V_i'(t) = \min(V_i, t)$, $\delta_i'(t) = 1_{\{(\min(V_i,t)<C_i)\}}$, $Z_i(t) = 1_{\{V_i>t\}}$ and let $\hat{G}(\cdot)$ be the Kaplan-Meier estimator of the residual censoring time distribution function. Following Wey et al., we minimize the inversely weighted (by the probability of censoring) objective function, the Brier score, to determine the optimal weights, $\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_m$ [16]. To control for possible overfitting, we use cross-validation and evaluate the Brier score over a set of s specified evaluation points to obtain:

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}:\alpha_k \in [0,1]}{\arg\min} \sum_{r=1}^{s} \sum_{i=1}^{n} \frac{\delta_i'(t_r)}{\hat{G}(Z_i(V_i'(t_r)))} \times \left\{ Z_i(t_r) - \sum_{k=1}^{m} \alpha_k \hat{S}_{\text{res},k}^{(-i)}(t_r) \right\}^2 \quad (6)$$

where the superscipt $(-i)$ denotes that the estimate was determined by leaving the i$^{\text{th}}$ observation out during the estimation procedure. Due to computational constraints, we performed 5 fold cross-validation and evaluated the optimal weight parameters over nine equally spaced out points covering the support of the observed residual lifetimes, as suggested by Wey et al. [16]. Finally, exploiting Eq. 6 to obtain $\hat{\boldsymbol{\alpha}}$, and Eq. 3 to obtain $\hat{S}_{U,m}(\cdot)$ from $\hat{f}_{\text{res},m}(\cdot)$, we define the stacked survival function estimator

$$\hat{S}_{U,\text{stack}}(\cdot) = \hat{\alpha}_1 \hat{S}_{U,1}(\cdot) + \hat{\alpha}_2 \hat{S}_{U,2}(\cdot) + ... + \hat{\alpha}_m \hat{S}_{U,m}(\cdot). \quad (7)$$

The estimator presented in (7) is, admittedly, an ad-hoc proposal, but one that is necessary given our lack of access

to data arising from the underlying survival distribution, but rather from the residual lifetime distribution.

Simulations

Using (potentially) right-censored simulated residual lifetime data, we evaluated the stacking estimator. We examined the performance of the individual parametric/nonparametric estimators relative to the stacked estimator when the residual lifetime data were subject to random right-censoring as well as increasing proportions of administrative right-censoring. The goal was to assess the increasing advantage, as follow-up decreases, of using a stacking estimator with both the corrected NPMLE and parametric survivor functions in the stack. A general description of the simulations examined in this manuscript is given in Table 1.

To simulate a set of right-censored residual lifetime data, we first generated an onset date, O, from a Uniform distribution with support $(0, 50)$. We generated a failure time, T, from either a Weibull distribution with shape and scale parameters equal to 2 and 2, respectively or from a mixture model of Weibull, Log-Logistic, Log-Normal and Gamma distributions. The latter failure time distribution was used to assess the predictive performance of the stacked estimator when the underlying failure time distribution was not included in any of the parametric models included in the stack. With the addition of covariates, the methods described by Bender et al. may be used to simulate failure times from a proportional hazards model [18]. However, with our proposed methodology, we do not consider the inclusion of covariate data through a regression-type model. We sampled onset, failure time pairs (O, T), for which $T > 50 - O$, until a sample of size n was selected. The sampled residual failure times, $T_i - (50 - O_i)$ for $i = 1, 2, ..., n$ were then right-censored either by a constant C^* to correspond to administrative censoring or by the random variable C_i drawn from an Exponential distribution to allow for random censoring (i.e. loss to follow-up).

Table 1 A summary of the simulation studies examining the performance of the proposed stacked survival model estimation procedure

Simulation Number	Simulation Study Description
Simulation 1	Weibull distributed failure times with various amounts of administrative censoring (10%, 20%, 30%, 40%) acting on the residual failure time data.
	Two stacked models fitted: All submodels, All submodels except Weibull
Simulation 2	Mixture model distributed failure times with 30% random censoring.
	One stacked model fitted: All submodels

In our first set of simulations, we assumed the underlying failure times were distributed according to a Weibull distribution and the residual failure times were administratively censored by moving up end-of-study dates to result in, respectively, 10%, 20%, 30% or 40% censoring. For each censoring percentage, we fit the corrected NPMLE, Weibull, Log-Logistic, Log-Normal and Gamma models. Using all five submodels, we determined the optimal stacking weights and computed the discrete integrated squared survival errors (DISSE) for the models when fitted separately and when combined in a stacked model. The DISSE is given by:

$$\text{DISSE} = \sum_{j=1}^{k} (t_j - t_{j-1})(\hat{S}(t_j) - S_0(t_j))^2$$

where we defined a uniform mesh, $0 = t_1 < t_2 < ... < t_k = 50$, to evaluate the predictive performance of the estimated survival functions over the support of the underlying survival function. The DISSE is the discretized version of the integral given by: $\int_0^\infty \left(\hat{S}(t) - S(t)\right)^2 dt$. To evaluate this integral numerically, we proposed a uniform mesh over the majority of the support of the estimated/true survival functions ($\hat{S}$, S, respectively). The upper bound of the support was set to "50" as the underlying survival functions in both the Weibull simulations and mixture model simulations had negligible probability beyond this point. The gauge of the mesh was set to 0.1 but could be made finer for a better approximation to the integral. We also considered a second stacked model which included the corrected NPMLE and all parametric models except the Weibull (i.e., the true data generating model). We utilized samples of size 125 (i.e. 125 observed residual lifetimes) over 100 simulation runs and report the average DISSEs in Table 2 as well as the average weights for the stacked models in Table 3. The average survival function estimates and 95% pointwise confidence intervals using the NPMLE or stacked model (without the Weibull submodel) are plotted in Fig. 3. As the proportion of administrative censoring increases, we see that although the average DISSEs of the individual and stacked models all increase, the NPMLE DISSE increases at a much faster rate than those of the individual parametric models and the stacked models. This is expected as administrative censoring shortens the follow-up period and the range of the observed residual lifetime data thus affecting the nonparametric maximum likelihood estimator most severely. In Fig. 3, the corrected NPMLE is clearly biased beyond the administrative censoring times with narrow confidence interval widths in this range. Using the stacked model, we find that almost all of the weight is shifted away from the corrected NPMLE and to the correct underlying Weibull failure time distribution. When the Weibull model is excluded from the

Table 2 Average discrete integrated squared survival errors for individual and stacked models for a Weibull (2,2) failure time distribution with varying amounts of administrative censoring for samples of size 125 over 100 simulation runs

	Proportion of Administrative Censoring			
Model	10%	20%	30%	40%
NPMLE	0.09594	0.1639	0.2414	0.3645
Weibull	0.02776	0.03843	0.06635	0.09601
Log-Logistic	0.03244	0.03350	0.05696	0.08661
Log-Normal	0.03009	0.03660	0.06222	0.07615
Gamma	0.03056	0.04181	0.06025	0.08460
Stacked Model (all)	0.02877	0.04203	0.06680	0.09865
Stacked Model (w/o Weibull)	0.03049	0.04303	0.06546	0.09010

stack, the majority of the weight shifted to the Gamma model and the stacked model still yielded a smaller DISSE than the corrected NPMLE. The stacked model without the Weibull submodel appears roughly unbiased and exhibits narrower pointwise confidence intervals than the NPMLE when the censoring proportion is low.

Our second simulation considered residual failure time data that were generated from a mixture model. The failure time mixture model was comprised of four evenly weighted (25%) models consisting of the Weibull (shape, scale equal to 4, 2), Log-Logistic (shape, scale equal to 1, 2), Log-Normal (meanlog, standard deviation-log equal to -5 and 1) and Gamma (shape, scale equal to 25, 1) distributions. To generate a sampled failure time from the mixture model, first we made a single draw from a multinomial distribution with four states with equal probabilities of 0.25. The multinomial draw determined from which parametric model we sampled the failure time. Once the failure time was sampled, we sampled an onset time and then repeated the same left-truncating/right-censoring procedure as was conducted in the first set of simulations to generate residual lifetime data drawn from a prevalent cohort study with follow-up. We chose a mixture model in order to produce a survival function, with "kinks", that does not resemble the survival function of any of the standard parametric models used in survival analysis. With this mixture model, we anticipated that because of its flexibility, the NPMLE would out-perform any of the estimators based on the standard models, even the stacking model. We believe, though, that in most applications the survival function is unlikely to arise from a mixture. The residual failure times were randomly right-censored to allow for approximately 30% censoring. In this simulation scenario, there was no administrative censoring. We chose not to allow for administrative censoring in order to isolate the effect on the predictive performance of the stacked model when the underlying failure time model was not a member of the class of submodels included in the stacking procedure. We fit all five submodels individually and combined them in a stacked model. In Fig. 4, we plot the underlying mixture survival function in black with the corrected NPMLE and

Table 3 Mean weights for a stacked model including all submodels or including all submodels except Weibull. The failure time data were generated according to a Weibull (2, 2) distribution with varying amounts of administrative censoring for samples of size 125 over 100 simulation runs

	Individual submodel type of stacked estimator				
Administrative Censoring Proportion	**NPMLE**	**Weibull**	**Log-Logistic**	**Log-Normal**	**Gamma**
10%	0.01394	0.08208	0.02781	0.04167	0.09579
	0.01607	N/A	0.05645	0.08028	0.8472
20%	5.586×10^{-9}	0.9061	0.01439	0.04962	0.02992
	8.419×10^{-9}	N/A	0.04088	0.08901	0.8701
30%	3.900×10^{-9}	0.8503	0.03581	0.04999	0.06389
	5.520×10^{-9}	N/A	0.07797	0.08018	0.8418
40%	2.169×10^{-9}	0.7888	0.02656	0.01465	0.1700
	3.463×10^{-9}	N/A	0.1319	0.04835	0.8197

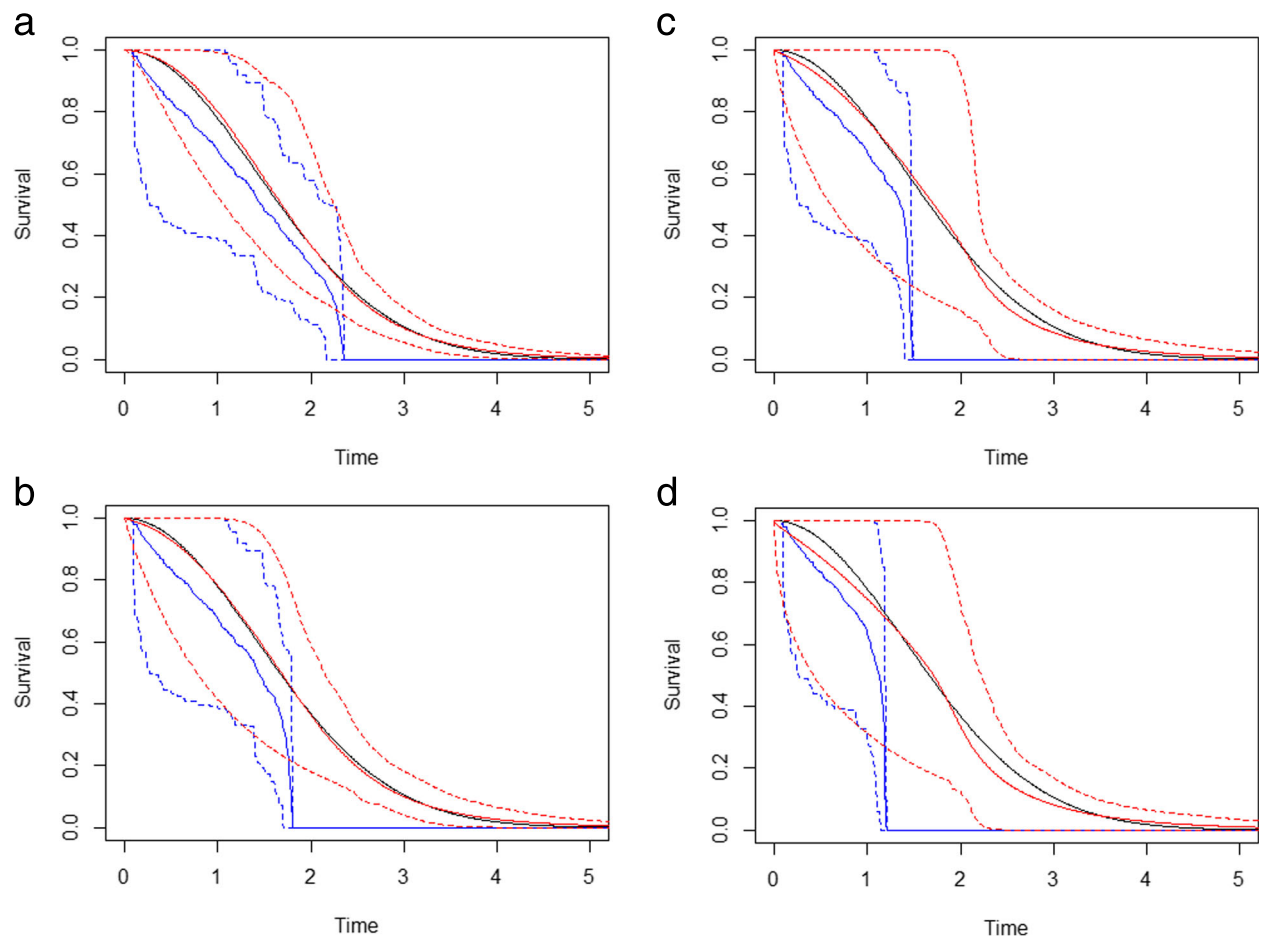

Fig. 3 Graphical comparison of the NPMLE estimate (blue lines), stacked model estimate (without the Weibull submodel included) (red lines) relative to the underlying Weibull failure time survival function (black line) using samples of size 125 over 100 simulation runs with varying amounts of administrative censoring

stacked survival function estimates, with their respectively bootstrapped 95% pointwise confidence intervals, in red. From the various plots, we find that the NPMLE tends to capture the general shape of the underlying survival function and captures the survival curve within its 95% pointwise confidence intervals. Other than the Gamma distribution, the individual parametric models do not capture the shape of the underlying survival function and for various models, their 95% confidence intervals do not capture the underlying survival curve at certain time points. Over 100 simulation runs, using the stacked model estimates, the mean weight for the corrected NPMLE was 0.4372 whereas the average weights for the other parametric models were 0.1475 (Weibull), 0.2581 (Log-Logistic), 0.01849 (Log-Normal) and 0.1387 (Gamma). The average DISSEs are listed in Table 4. Although the stacked model estimate did not perform as well as the corrected NPMLE with respect to the average DISSE, the stacked estimator yielded an improvement over the other parametric models. In addition, unlike the parametric estimators, there were no time points at which the stacked model estimator's bootstrapped 95% pointwise confidence interval did not capture the underlying survival function.

Application

We demonstrated our proposed stacking estimator by using it to estimate survival with dementia from forward recurrence time data obtained from the CSHA, as described at the beginning of Section 1. We estimated the underlying survival functions for the probable Alzheimer's disease group (389 participants, approx. 21% censoring), possible Alzheimer's disease group (253 participants, approx. 24% censoring) and vascular dementia group (172 participants, approx. 19% censoring), separately, and computed bootstrapped 95% pointwise confidence intervals. The stacked estimator included the corrected NPMLE,

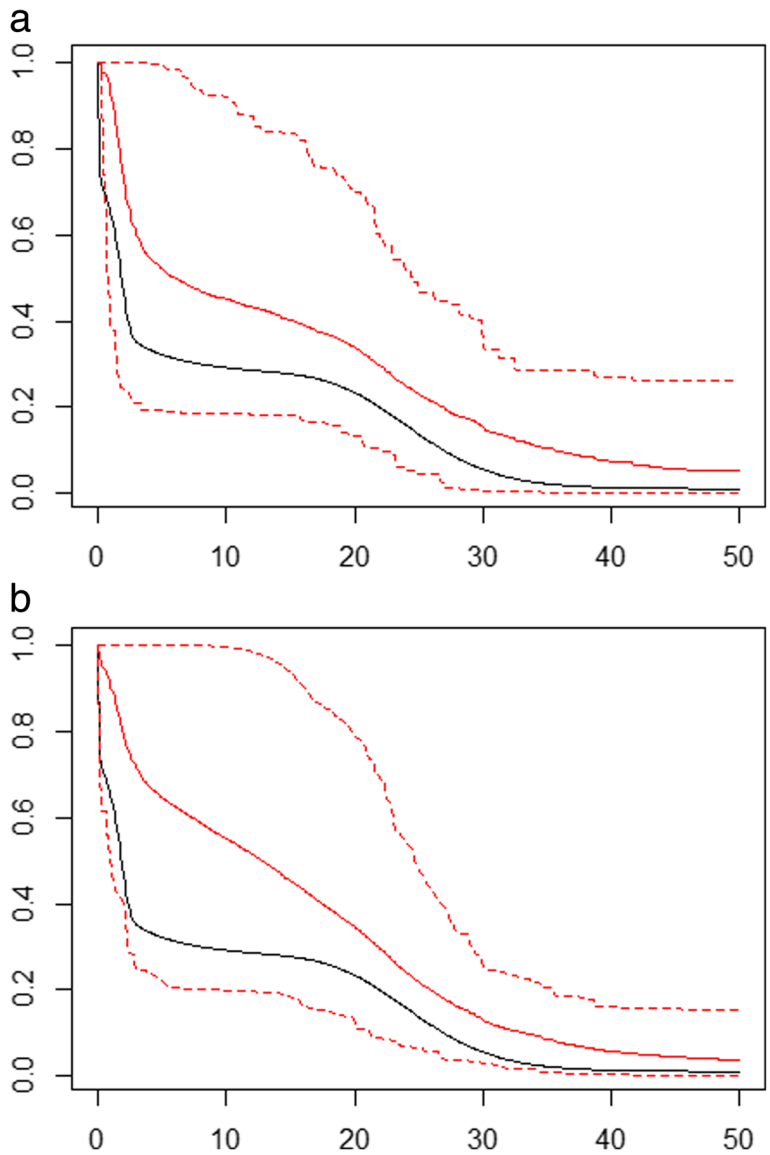

Fig. 4 Graphical comparison of the individual model mean survival estimates (solid red line) with bootstrapped 95% pointwise confidence intervals (dotted red lines) relative to the underlying mixture failure time survival function (solid black line) using samples of size 125 over 100 simulation runs with 30% random censoring (Panel a - NPMLE, Panel b - Stacked Estimator)

Weibull, Log-Normal and Gamma estimators (See Additional file 3 for individual parametric estimates of survival). We did not include the Log-Logistic distribution in the stacking procedure as it admits only a decreasing hazard function, and is therefore not suitable as a model for survival with dementia. In Fig. 1, we plot the stacked estimates with 95% pointwise confidence limit curves in red along with the corrected NPMLE and the associated 95% pointwise confidence limit curves in black.

From Fig. 1, the stacked estimate generally captures the same shape as the NPMLE until approximately 60 months. The 60 month mark corresponds to the approximate follow-up time for subjects in the study and thus the non-parametric estimate does not capture the underlying survival function behaviour past this point. On the other hand, since the stacked estimator is defined as a linear combination of both the corrected NPMLE and parametric estimators with support unconstrained by the observed data, the resulting stacked estimate captures the survival function tail behaviour past 60 months. Additionally, from Fig. 1, we see that the bootstrapped confidence intervals based on the stacked estimator and the NPMLE

Table 4 Average discrete integrated squared survival errors (DISSE) for individual and stacked models for a mixture failure time distribution with 30% random censoring for samples of size 125 over 100 simulation runs

Model	Average DISSE
NPMLE	1.478
Weibull	2.755
Log-Logistic	5.612
Log-Normal	4.655
Gamma	3.075
Stacked Model (all)	2.273

are approximately of equal width in the case of the probable Alzheimer's group. In the possible Alzheimer's group, the widths based on the stacked estimator are visibly reduced, while the widths are strikingly reduced in the vascular dementia group. We note that in the possible Alzheimer's disease group, decline in the first 60 months appears to be more rapid than in the other two groups. We speculate that the possible Alzheimer's disease group included a variety of non-Alzheimer's disease dementias, some of which are characterized by rapid decline.

In the probable and possible Alzheimer's disease groups, the Weibull model received most of the weight, while in the vascular dementia group, the Gamma model was heavily favoured. This demonstrates the ability of the stacking model to shift its assignment of weight to a different model in the stack, a model (such as the Gamma) for example, that may not have been considered alone initially. For a listing of the individual submodel weights of the stacked models, see Table 5. The median survival estimated for all three dementia types was roughly 4.2 years when using a stacking model for the residual lifetimes. In comparison, the estimated median survival for the three dementia groups combined was roughly 4.5 years when using the full data that included the current lifetimes [3]. The latter (full) data, naturally, produced much narrower pointwise confidence intervals.

Discussion

We originally hoped to improve the corrected NPMLE when estimating the survival function using only the observed residual lifetimes from a prevalent cohort study with follow-up. Our goal was to introduce parametric models into a stacking estimator, while retaining the corrected NPMLE, speculating that the parametric models would mitigate the two major shortcomings of the corrected NPMLE: (i) the wide pointwise confidence intervals that are often produced, and (ii) the failure of the corrected NPMLE to capture the tail behaviour of the survival function, particularly when follow-up is short. However, we found that when comparing the estimators using, essentially, their average DISSEs, the corrected NPMLE did not perform well either alone or as a member of the stack when follow-up was short.

The sample mean discrete integrated squared survival error takes into account both bias and variance. Nevertheless, our application suggests that even though confidence interval width is concerned only with variance, the stacked estimator produces narrower (sometimes considerably narrower) confidence intervals than those of the corrected NPMLE. A potential objection to the use of parametric models in the setting of this article, is their lack of robustness to model misspecification. By building the stack with several different parametric models, we believe that to a large extent, these fears can be allayed. It is comforting to see that in our example, the survival function produced by the stacking estimator is smooth, in that it does not have difficult-to-explain kinks.

An alternative approach for modelling the underlying survival function is through a parametric mixture model. Rather than fitting the individual parametric models separately for the residual lifetime density functions and then subsequently estimating the weights, one can define a mixture model likelihood function and then maximize the likelihood to estimate the failure time parameters and model weights simultaneously [19]. In contrast, it is possible to define a parametric mixture model of the submodel survival functions and then estimate the unknown parameters by maximizing the corresponding likelihood function. In both proposed approaches however, it remains an area of future research as to how to incorporate the non-parametric estimates into the mixture models and how these various procedures compare when predicting the underlying survival function. It is worth noting that multi-state models are often applied to survival (or event history) data. However, their use is somewhat limited under stationarity and it is hard to see how their introduction would enhance the proposed methods.

Table 5 Weights of stacked survival models applied to the three dementia type strata of the Canadian Study of Health and Aging

	Individual submodel type of stacked estimator			
CSHA Strata	NPMLE	Weibull	Log-Normal	Gamma
Probable Alzheimer's Disease	1.282×10^{-8}	0.9928	2.523×10^{-7}	0.007240
Possible Alzheimer's Disease	1.000×10^{-8}	6.556×10^{-7}	1.353×10^{-7}	0.9999
Vascular Dementia	1.126×10^{-8}	0.7179	2.022×10^{-7}	0.2821

Abbreviations
CSHA: Canadian study of health and aging; DISSE: Discrete integrated squared survival errors; NPMLE: Nonparametric maximum likelihood estimator; MLE: Maximum likelihood estimator; PDF: Probability density function

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s12874-021-01496-3.

Additional file 1: Supp1-RFunctionsCode.R
Additional file 2: Supp2-RSimulationsCode.R
Additional file 3: Supplementary Figures.

Acknowledgements
Not Applicable.

Authors' contributions
DBW directed the research on the stacking model algorithm, the application to the CSHA data set and aided in the preparation of the manuscript. JHM, VA and ZL implemented the stacking model algorithm and aided in the preparation of the manuscript. The authors read and approved the final manuscript.

Authors' information
Not Applicable.

Funding
The CSHA was supported by the Seniors Independence Research Program, through the National Health Research and Development Program (NHRDP) of Health Canada (Project 6606-3954-MC[S]). The progression of dementia project within the CSHA was supported by Pfizer Canada through the Health Activity Program of the Medical Research Council of Canada and the Pharmaceutical Manufacturers Association of Canada; by the NHRDP (project 6603-1417-302[R]); by Bayer; and by the British Columbia Health Research Foundation (projects 38 [93-2] and 34 [96-1]).
The first author was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) PGSD-3 award.

Availability of data and materials
Dr. Christina Wolfson (christina.wolfson@mcgill.ca, Department of Epidemiology, Biostatistics and Occupational Health, McGill University) can be contacted concerning access to the data that support the findings of this study. The data are not publicly available. All simulation R code applied in manuscript available in Additional Files 1 and 2.

Declarations

Ethics approval and consent to participate
Not Applicable.

Consent for publication
Not Applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Mathematics and Statistics, McGill University, Montreal, CA. [2]Department of Mathematics, Statistics and Computer Science, Macalester College, St.Paul, USA.

Received: 6 August 2021 Accepted: 16 December 2021
Published online: 07 January 2022

References
1. Wolfson C, Wolfson DB, Asgharian M, M'Lan CE, Østybe T, Rockwood K, Hogan DB. A reevaluation of the duration of survival after the onset of dementia. N Engl J Med. 2001;344(15):1111–16.
2. OL O. Canadian Study of Health and Aging: study methods and prevalence of dementia. Can Med Assoc J. 1994;150(6):899–913.
3. Asgharian M, M'Lan CE, Wolfson DB. Length-biased sampling with right censoring; an unconditional approach. J Am Stat Assoc. 2002;97(457): 201–09.
4. Wang M-C. Nonparametric estimation from cross-sectional survival data. J Am Stat Assoc. 1991;86(413):130–43.
5. McVittie JH, Wolfson DB, Stephens DA. Parametric modelling of prevalent cohort data with uncertainty in the measurement of the initial onset date. Lifetime Data Anal. 2020;26(2):389–401.
6. Keiding N, Fine JP, H HO, Slama R. Accelerated failure time regression for backward recurrence times and current durations. Stat Probab Lett. 2011;81:724–29.
7. Keiding N, Kvist K, Hartvig H, Tvede M, Juul S. Estimating time to pregnancy from current durations in a cross-sectional sample. Biostatistics. 2002;3(4):565–78.
8. Keiding N, Hansen OKH, Sørensen DN, Slama R. The current duration approach to estimating time to pregnancy. Scand J Stat. 2012;39(2): 185–204.
9. Westling T, Carone M. A unified study of nonparametric inference for monotone functions. Ann Statist. 2020;48(2):1001–24.
10. Grenander U. On the theory of mortality measurement, part ii. Skand Akt. 1956;39:125–53.
11. Woodroofe M, Sun J. A penalized maximum likelihood estimator of $f(0+)$ when f is non-increasing. Statistica Sinica. 1993;3:501–15.
12. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1981;53(282):457–81.
13. Huang J, Wellner JA. Estimation of a monotone density or monotone hazard under random censoring. Scand J Stat. 1995;22(1):3–33.
14. Denby L, Vardi Y. The survival curve with decreasing density. Technometrics. 1986;28(4):359–67.
15. Huang Y, Zhang C-H. Estimating a monotone density from censored observations. Ann Stat. 1994;22(3):1256–74.
16. Wey A, Connett J, Rudser K. Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models. Biostatistics. 2015;16(3):537–49.
17. Wey A, Vock DM, Connett J, Rudser K. Estimating restricted mean treatment effects with stacked survival models. Stat Med. 2016;35(19): 3319–32.
18. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Stat Med. 2005;24(11):1713–23.
19. Smyth P, Wolpert D. Linearly combining density estimators via stacking. Mach Learn. 1999;36:59–83.

Publisher's Note