Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Conditional life expectancy estimation by ordered fractions of population with censored data

Celia García-Pareja*[1] | Michele Santacatterina[2] | Anna Mia Ekström[3,4] | Matteo Bottai[1]

[1]Unit of Biostatistics, Karolinska Institutet, SE-17177, Stockholm, Sweden
[2]TRIPODS Center for Data Science for Improved Decision Making and Cornell Tech, Cornell University, NY-10044, New York
[3]Department of Public Health Sciences, Karolinska Institutet, SE-17177, Stockholm, Sweden
[4]Department of Infectious Diseases, Karolinska University Hospital, SE-14186, Stockholm, Sweden

Correspondence
*Celia García-Pareja, Unit of Biostatistics, IMM, Box 210, SE-17177, Karolinska Institutet, Stockholm, Sweden. Email: celia.garcia.pareja@ki.se

Summary

Life expectancy is one of the most commonly used measures to describe population's health. However, its estimation might be hindered in the presence of censoring, where the time variable is only observed until a certain quantile. To deal with this issue, available methods for life expectancy estimation entail, for example, the ad-hoc choice of a cut-off time point until which the time variable has been observed, and that yields estimated quantities that have difficult interpretation. We propose a novel quantile-based approach for estimating life expectancy in terms of observed covariates of interest and in the presence of censored data, in which we divide the population under study into survival-ordered fractions defined by a set of proportions, and provide conditional life expectancy estimators for each fraction separately. Our approach provides a detailed picture of the distribution of time-to-death while preserving the appealing interpretability of life expectancy estimates and allows for easy groups' comparisons. In a simulation study, we show the use of our methodology in combination with widely used survival analysis models and show its advantages with respect to other available methods. In particular, our measure proves to be of great use for detecting differences in life expectancy across specific fractions that currently used methods fail to detect. Finally, we apply our method to life expectancy estimation on late presenters, that is, among those who start treatment late, on data from a clinical trial on people living with human immunodeficiency virus in the United States.

KEYWORDS:
Life expectancy estimation, mean survival time, restricted mean, compound expectation, Cox model, AFT model

1 | INTRODUCTION

Life expectancy is a widely known health summary measure, used by the scientific community and popular among the lay public, and it is extensively used in public health research and in policy-decision making to predict treatment compliance and healthcare costs[1,2]. Life expectancy estimation is of importance, for example, among people living with human immunodeficiency virus (PLHIV). Although survival among PLHIV has improved over the years, it is still behind that observed in the general population[3]. Late presentation to health-care services due to social barriers, such as stigma and discrimination[4] is an evidently related risk factor. Its impact on life expectancy is, however, still undetermined. A major challenge in estimating life expectancy among

PLHIV is posed by the loss to follow-up in cohort studies and censoring at the end of the follow-up period. These issues are common to many longitudinal studies and particularly frequent in HIV cohorts[5].

However, life expectany estimation poses some difficulties because it relies on global properties of the underlying conditional distribution of time-to-death, which is usually not fully observed in the presence of censoring. To solve this issue, in his seminal paper[6] Irwin proposed estimating the restricted mean instead, where life expectancy is computed only up to a certain cut-off chosen observed survival time. Karrison[7] examined the restricted mean as an index for comparing survival in different groups and proposed an extension of the method to incorporate covariates based on a piecewise exponential model for the baseline hazard. Royston and Parmar[8] considered the restricted mean as a useful general measure to report differences between two survival curves.

Nonetheless, methodology based on the restricted mean presents several doubts. Criteria for choosing the cutoff point up to which the restricted mean is to be estimated across different groups remain still unclear. The population in those groups can be completely different in terms of patients' frailty and in those cases comparison is no longer meaningful. In this paper, we focus on the advantages of using the Conditional Compound Expectation (CCE)[9] in survival settings, where censoring often prevents the observation of the entire underlying distribution, and present it as an alternative for life expectancy estimation in combination with widely used survival models.

The rest of this paper is structured as follows. In Section 2 we present the rationale behind the CCE and comment on its differences with respect to the restricted mean. Section 3 provides an overview on how to estimate the CCE in combination with commonly used survival models, more specifically, with the Cox and accelerated time failure (AFT) models and Section 4 is devoted to present the finite sample performance of the CCE estimation using the aforementioned models. In Section 5 we show the application of our proposed approach to real data. Finally, in Section 6 we comment on some final remarks and possible future work.

2 | CONDITIONAL COMPOUND EXPECTATION WITH CENSORED DATA

Let T be a non-negative random variable representing time-to-death. On what follows, we exploit the relation between the conditional mean $\mu(x) := \mathrm{E}(T|x)$ and the conditional quantile function $Q(\cdot|x)$ of T given a set of observed covariates $x = (x^1, \dots, x^m)$,

$$\mu(x) = \mathrm{E}(T|x) = \int_0^\infty S(t|x)\mathrm{d}y = \int_0^1 Q(p|x)\mathrm{d}p, \tag{1}$$

where $S(\cdot|x)$ denotes the survival function of T and we assume $\mathrm{E}(T|x) < \infty$ for all x.

If we consider a grid of quantile levels $\boldsymbol{\tau} = \{\tau_k : 0 < \tau_k < 1, \tau_{k-1} < \tau_k\}$, we can divide $\mu(x)$ into K components $\{\mu_k(x)\}_{k=1}^K$ such that

$$\mu(x) = \sum_{k=1}^K \mu_k(x) := \sum_{k=1}^K \left(\int_{\tau_{k-1}}^{\tau_k} Q(p|x)\mathrm{d}p \right), \tag{2}$$

where here we take $\tau_0 = 0$ and $\tau_K = 1$ for convenience. Such division allows to treat different portions of population separately, where each portion relates to a certain fraction of interest in the population and is delimited by the corresponding interval $[\tau_{k-1}, \tau_k]$. Once the components have been defined, one might in turn, compute the conditional mean of T given x for each of these fractions separately, that is,

$$\overline{\mu}_k(x) = \frac{\mu_k(x)}{\tau_k - \tau_{k-1}},$$

where $\{\overline{\mu}_k(x)\}_{k=1}^K$ is what has been coined as the conditional compound expectation (CCE) of T given x[9]. An advantage of the CCE is that its estimation allows for the choice of any suitable estimator of the underlying quantile function, and the CCE's estimates will simply inherit its properties, namely, unbiasedness, consistency and asymptotic normality providing a useful framework to derive confidence intervals and other inferential tools[9]. In what follows, we highlight the advantages of using the CCE in survival scenarios, where censoring might prevent the observation of the complete distribution of time-to-death, hence hindering estimation of life expectancy.

Let $Y = \min(T, C)$ be a non-negative random variable, where T depicts time-to-death and C is a random censoring mechanism that prevents T to be fully observed. In the presence of C, τ_K can be set to the largest proportion of observed deaths, that is, the one corresponding to the last observed quantile, and the CCE can be computed up to this τ_K. While in this case, the CCE does

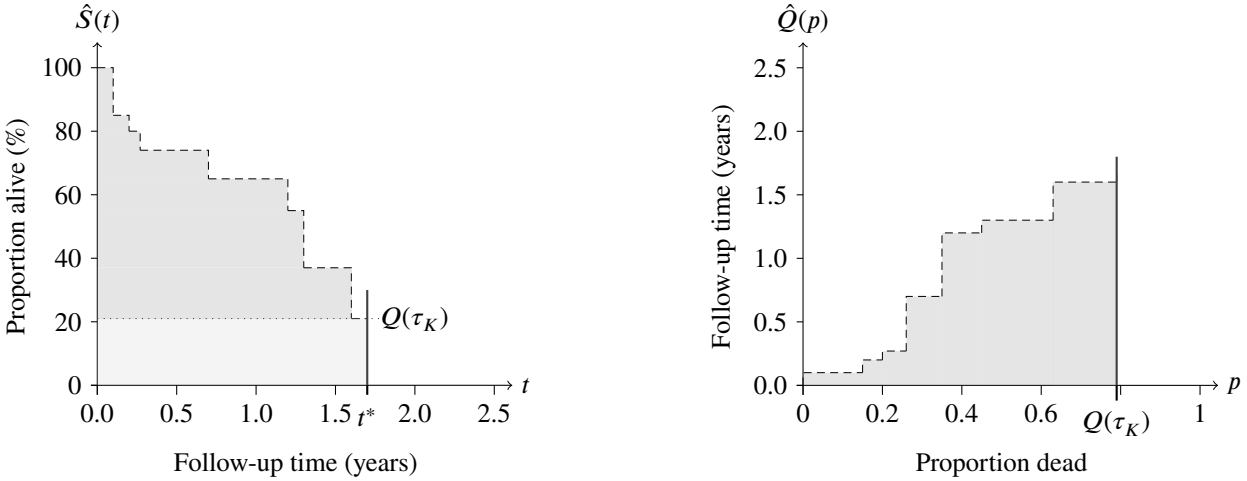

FIGURE 1 Kaplan Meier estimated survival curve (left) and respective quantile function (right). The dark grey shaded areas depict the value of the component $\mu_K(x)$, while the light and dark grey shaded areas together (left) depict the value of the restricted mean up to $t^\star$.

not provide a life expectancy estimator for the entire population under study, it does provide life expectancy estimators for each fraction delimited by the chosen grid $\boldsymbol{\tau}$. This has also advantages in terms of inference, as shown later on in Section 4, because life expectancy estimates for components with smaller k are typically more precise than those for components with larger k. Indeed, it is to be expected that larger observed quantiles will be more affected by the censoring variable, in the sense that the number of observed censored events typically increases over time, so that estimation of the corresponding components will be affected accordingly.

It is worth noting that while the restricted mean[6] also allows estimation up to a last observed quantile t^*, our approach provides easier interpretation of the obtained estimates and meaningful comparison between groups. In brief, while for the restricted mean at least one member in each group has lived up to the t^*, the cut-off does not provide any further information about the type of individuals in each of the groups, and that are actually being compared. In the case of the CCE, comparisons are made across components with the same value of k, that is, between groups that represent the same fraction of population in terms of survival. For example, for $k = 1$, the comparison would be done between groups corresponding to the weakest patients, that is, the same fraction to have died first in each group.

Moreover, note that for $\tau_K < 1$, life expectancy for the first τ_K-th fraction of the population to die, that is, the CCE with $\boldsymbol{\tau} = \{0, \tau_K\}$ and one unique component $\mu_K(x)$, does not correspond to the restricted mean computed up to the last observed quantile $t_x^\star = Q(\tau_K|x)$ (see Figure 1 for illustration). This in turn, highlights a central difference between both measures in terms of interpretation of the resulting estimates. Indeed, while

$$\overline{\mu}_K(x) = \frac{1}{\tau_K} \int_0^{\tau_K} Q(p|x)\mathrm{d}p$$

can be easily interpreted as the life expectancy of a characterized fraction of the population under study, the corresponding

$$\mu^\star(x) = \int_0^{t_x^\star} S(t|x)\mathrm{d}t,$$

which refers to the life expectancy for a population followed for $t_x^\star$ units of time, does not prove as informative.

TABLE 1 Baseline hazard, cumulative hazard, its inverse, survival and quantile functions for commonly used parametric survival models

	Distributions		
Functions	Exponential(λ)	Weibull(λ,υ)	Gompertz(λ,s)
Hazard	$\lambda_0(t) = \lambda$	$\lambda_0(t) = \lambda\upsilon t^{\upsilon-1}$	$\lambda_0(t) = \exp(st)$
Cum hazard	$\Lambda_0(t) = \lambda t$	$\Lambda_0(t) = \lambda t^{\upsilon}$	$\Lambda_0(t) = \frac{\lambda}{s}(\exp(st) - 1)$
Inv Cum hazard	$\Lambda_0^{-1}(t) = \lambda^{-1}t$	$\Lambda_0^{-1}(t) = (\lambda^{-1}t)^{1/\upsilon}$	$\Lambda_0(t) = \frac{1}{s}\log(\frac{s}{\lambda} + 1)$
Survival	$S_0(t) = \exp(-\lambda t)$	$S_0(t) = \exp(-\lambda\lambda t^{\upsilon})$	$S_0(t) = \exp(\frac{\lambda}{s}(1 - \exp(st))$
Quantile	$Q_0(p) = -\frac{\log(1-p)}{\lambda}$	$Q_0(p) = \left(-\frac{\log(1-p)}{\lambda}\right)^{1/\upsilon}$	$Q_0(p) = \frac{1}{\alpha}\log\left(1 - \frac{s\log(1-p)}{\lambda}\right)$

3 | ESTIMATION STRATEGIES WITH WIDELY USED SURVIVAL MODELS

As pointed out above, one of the main advantages of the CCE is that it allows using a suitable estimator of the underlying quantile distribution in any scenario. In this section, we propose estimating strategies involving commonly used survival models that we will use further on in a simulation study.

3.1 | Cox model

The Cox proportional hazards model is amongst the most widely spread survival models and it is commonly used in a large number of applications. In the Cox model, the conditional hazard function $\lambda(t|x)$ is assumed to factorize on a baseline underlying hazard function $\lambda_0(t)$, shared by all individuals in the population, and an independent term that controls differences (hazard ratios) between a set of covariates of interest, that is,

$$\lambda(t|x) = \lambda_0(t)\exp\left(\sum_{i=1}^{m}\beta_i x_i\right).$$

While one of the model's strengths is that one does not need to specify $\lambda_0(\cdot)$ in order to estimate the β_i parameters, there exist many examples in which characterizing the underlying distributional shape might be of interest. Such characterizations have been largely explored in the literature and include, for example, assuming fully parametric models[10] or flexible parametric approaches[11],[12], to mention a few.

Estimation of the CCE assuming an underlying Cox proportional hazards model relies on the following equivalences. Recall that $-\log(S(t|x)) = \Lambda(t|x) = \int_0^t \lambda(s|x)ds$, and thus,

$$S(t|x) = \exp\left\{-\Lambda_0(t)\exp\left(\sum_{i=1}^{m}\beta_i x_i\right)\right\} \Leftrightarrow Q(p|x) = \Lambda_0^{-1}\left(-\log(1-p)\exp\left(-\sum_{i=1}^{m}\beta_i x_i\right)\right), \tag{3}$$

for invertible Λ_0.

Therefore, an estimator for the k-th component μ_k is simply

$$\widehat{\mu_k}(x) = \int_{\tau_{k-1}}^{\tau_k} \widehat{\Lambda_0}^{-1}\left(-\log(1-p)\exp\left(-\sum_{i=1}^{m}\hat{\beta}_i x_i\right)\right)\mathrm{d}p \tag{4}$$

where the notation $\hat{\cdot}$ denotes estimator.

Difficulties in the computation of $\widehat{\mu_k}(x)$ will be then highly dependent on the chosen underlying baseline hazard $\lambda_0(\cdot)$, that is, the assumed underlying baseline distribution. Table 1 shows distributional functions of interest for commonly used parametric models in survival that support the proportional hazards assumption, and whose inverse cumulative hazard exists and has closed form. In case flexible parametric models are preferred instead, $\Lambda_0^{-1}(t)$ will not (in general) be available in closed form, but one might resort to numerical approximations.

In case interest lies in comparisons between different groups, (4) can be evaluated at different covariate patterns, and their CCE differences can be estimated.

3.2 | Accelerated failure time model

The accelerated failure time (AFT) model assumes that the survival function for different individuals accelerates (or decelerates) survival time w.r.t. a specified underlying baseline survival $S_0(t)$ (or reference group), and this acceleration factor depends on the set of covariates of interest. More specifically, it assumes that

$$S(t|x) = S_0(t\exp(-\sum_{i=1}^{m}\beta_i x_i)).$$

Thus, the survival at time t for an individual with covariate pattern $x = (x_1, \dots, x_m)$ is the survival time for the reference group at time $t\exp(-\sum_{i=1}^{m}\beta_i x_i))$. The term $\exp(-\sum_{i=1}^{m}\beta_i x_i))$ is actually referred to as the deceleration factor.

Similarly to Cox proportional hazards model, the AFT model is also widely used in survival applications, and it turns out really convenient in our setting because it admits representation as a log-linear model

$$\log T = \mu + \sum_{i=1}^{m}\beta_i x_i + \sigma W, \tag{5}$$

where μ and σ are location and scale distributional parameters (respectively), and the residuals W have any distribution. For instance, if $\log T \sim \text{Logistic}(\mu + \sum_{i=1}^{m}\beta_i x_i, \sigma)$ and thus $W \sim \text{Logistic}(0,1)$, this means we are assuming a log-logistic model for T with $T \sim \text{logLogistic}(\exp(\mu + (\sum_{i=1}^{m}\beta_i x_i)), \frac{1}{\sigma})$, the parameters referring to scale and shape in this case.

For any assumed model for W, and because quantiles are invariant to any monotonically increasing transformation, from (5) we have

$$Q_{\log T}(p|x) = \mu + \sum_{i=1}^{m}\beta_i x_i + \sigma Q_W(p) = \log(Q_T(p|\mathbf{x})),$$

where $Q_Z(\cdot)$ denotes the quantile function of a random variable Z. Thus,

$$Q_T(p|x) = \exp(\mu + \sum_{i=1}^{m}\beta_i x_i + \sigma Q_W(p)) = \exp(\mu + \sum_{i=1}^{m}\beta_i x_i)\exp(\sigma Q_W(p)).$$

As shown further on, this factorization will allow to estimate the effect of each covariate in any component.

Estimation of the CCE assuming an AFT model is easily derived. For each k and quantile levels (τ_{k-1}, τ_k)

$$\widehat{\mu}_k(x) = \int_{\tau_{k-1}}^{\tau_k}\exp(\hat{\mu} + \sum_{i=1}^{m}\hat{\beta}_i x_i + \hat{\sigma}Q_W(p))\mathrm{d}p = \exp(\hat{\mu})\exp(\sum_{i=1}^{m}\hat{\beta}_i x_i)\int_{\tau_{k-1}}^{\tau_k}\exp(\hat{\sigma}Q_W(p))\mathrm{d}p, \tag{6}$$

where note that $\exp\hat{\beta}_i$ is a multiplicative effect on the covariate. For example, if we have $\widehat{Q}_T(p|\mathbf{x}) = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\sigma}Q_W(p))$ with x_1 binary and taking on values 0 or 1,

$$\widehat{\mu}_k(x_1 = 1) = \exp(\hat{\beta}_1)\widehat{\mu}_k(x_1 = 0),$$

that is, the k-th component for $x_1 = 1$ is $\exp(\hat{\beta}_1)$ times the component for the reference $x = 0$.

Note that in the log-Logistic case, the AFT model would also have interpretation in terms of proportional odds.

4 | SIMULATION STUDY

In this section, we present results from a simulation study to show the finite sample performance for CCE estimators assuming first, a Cox and then an AFT as underlying models for the data.

4.1 | Simulations setup

We considered two parametric models, namely, a) a Cox proportional hazards model with baseline Weibull distribution, that is, $T_1 \sim \text{Weibull}(0.05, 1.5)$ and b) an AFT model with baseline log-logistic distribution, that is, $T_2 \sim \text{logLogistic}(2,2)$. Censoring variables were assumed independent of T_1 and T_2 given covariates, and were drawn from uniform distributions, namely, $C_1 \sim \mathcal{U}(0,10)$ and $C_2 \sim \mathcal{U}(0,30)$, providing an average 50% censoring rate. Both scenarios included two covariates of interest

$X_1 \sim$ Bernouilli(0.5) and $X_2 \sim$ Normal(0, 1) and parameters $\beta_1 = 1$ and $\beta_2 = 3$. More specifically, we considered the Cox model

$$\lambda(T_1|X) = 0.05\frac{3}{2}T_1^{2/3}\exp(X_1 + 3X_2) \tag{7}$$

and the AFT model

$$\log T_2 = 2 + X_1 + 3X_2 + 0.5W, \tag{8}$$

where $W \sim$ Logistic(0, 1).

We simulated survival times under the Cox model (7) following the strategy proposed in Bender et al.[13], which derive from the equivalences shown in (3). More specifically, we have

$$S(t|x) = \exp\left\{-\Lambda_0(t)\exp\left(\sum_{i=1}^{m}\beta_i x_i\right)\right\} \Leftrightarrow F(t|x) = 1 - \exp\left\{-\Lambda_0(t)\exp\left(\sum_{i=1}^{m}\beta_i x_i\right)\right\},$$

where $F(\cdot)$ denotes the cumulative distribution function.

Now, we know that for any random variable Z, $F(Z) = U$ is uniformly distributed on [0, 1], and thus, $1 - U = \tilde{U}$ is also uniformly distributed. Then, for

$$\tilde{U} = \exp\left\{-\Lambda_0(t)\exp\left(\sum_{i=1}^{m}\beta_i x_i\right)\right\} \sim \mathcal{U}(0,1) \text{ we have } T = \Lambda_0^{-1}\left(-\log(\tilde{U})\exp\left(-\sum_{i=1}^{m}\beta_i x_i\right)\right),$$

which provides an easy simulation strategy for survival times T, by simply drawing uniform samples $\tilde{U} \sim \mathcal{U}(0, 1)$ and then substituting in the formula above for given $\Lambda_0^{-1}(\cdot)$ and β_i parameters. In our particular simulation scenario,

$$T_1 = \left(\frac{-\log(\tilde{U})\exp(-x_1 - 3x_2)}{0.05}\right)^{2/3}.$$

Simulation of survival times T_2 under the AFT model was done by simply drawing samples from $W \sim$ Logistic(0, 1) and then exponentiating the right hand side of (8).

Once survival times were simulated under models (7) and (8), we estimated the CCE of T_1 and T_2 considering a grid of deciles, i.e., $\boldsymbol{\tau} = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$, with each component's estimator of the form

$$\widehat{\mu}_k(x) = \int_{\tau_{k-1}}^{\tau_k}\left(\frac{-\log(1-p)\exp(-x_1 - 3x_2)}{0.05}\right)^{2/3}\mathrm{d}p \tag{9}$$

and

$$\widehat{\mu}_k(x) = \int_{\tau_{k-1}}^{\tau_k}\exp\left\{2 + x_1 + 3x_2 + 0.5\log\left(\frac{p}{1-p}\right)\right\}\mathrm{d}p,$$

for all k and each model respectively. Results of our simulation study are reported in the next subsection.

4.2 | Results

We report obtained results for 2000 simulations with sample size $n = 300$ for each of the models detailed in Section 4.1 (Table 2 and Table 3). Scenarios considering larger and smaller sample sizes yielded similar conclusions and are therefore not reported here. Parametric Cox models were estimated using function `phreg` from the `eha` CRAN repository[14].

In both scenarios, we observe that the standard errors increase with the order k of the components μ_k. This is due to the lack of data in the area of the distribution corresponding to larger survival times T, caused by the censored observations. Indeed, in[9] we argue that the CCE has higher variability in fractions of the distribution that are less populated with data. As we can see in this setting, this conclusion remains valid even when a parametric model is employed, because any method that fits the parameters prioritizes the fit in regions where the distribution of the data has been better observed.

The CCE also provides a clear description of the shape of the distribution. Sharp increases between consecutive components, for example, indicate large queues to the right of the distribution, e.g., the 2.5 factor multiplication between μ_9 and μ_{10} in Table 3 reflects the heavy tail of the log-logistic distribution and still provides information in terms on mean survival, i.e., life expectancy.

Therefore, CCE estimates are advantageous to subtract relevant information about life expectancy on certain fractions on population, whose precision is not as affected by censorship. This poses very useful in comparison with other methods that use the entire observed distribution of the data for inference purposes.

TABLE 2 Simulation results for the Cox-Weibull model, reporting the number k-th of component, component's true value μ_k under the model, estimated component's value averaging over 2000 simulations and square root of the mean-squared error (se). Estimated parameters for the baseline model are also reported.

Baseline CCE										
k	1	2	3	4	5	6	7	8	9	10
μ_k	0.070	0.101	0.143	0.202	0.288	0.416	0.616	0.952	1.590	2.989
$\hat{\mu}_k$	0.071	0.102	0.144	0.204	0.290	0.419	0.620	0.956	1.593	2.982
se	0.008	0.011	0.015	0.021	0.030	0.043	0.063	0.096	0.160	0.331

Model parameters				
Parameter	β_1	β_2	λ	υ
Average estimated value	1.009	3.042	0.050	1.519

TABLE 3 Simulation results for the AFT-loglogistic model, reporting the number k-th of component, component's true value μ_k under the model, estimated component's value averaging over 2000 simulations and square root of the mean-squared error (se). Estimated parameters for the baseline model are also reported.

Baseline CCE										
k	1	2	3	4	5	6	7	8	9	10
μ_k	0.161	0.310	0.427	0.543	0.669	0.819	1.010	1.288	1.787	4.594
$\hat{\mu}_k$	0.164	0.314	0.431	0.547	0.674	0.823	1.014	1.291	1.791	4.614
se	0.0211	0.034	0.045	0.056	0.070	0.088	0.113	0.153	0.234	0.906

Model parameters				
Parameter	β_1	β_2	μ	σ
Average estimated value	1.004	3.005	2.001	0.496

5 | TREATMENT EFFECT ON LIFE EXPECTANCY AMONGST DEEPLY IMMUNOSUPPRESSED PLHIV

In this section we present an illustrative example of the use of the CCE, analyzing data from a multicenter clinical trial on PLHIV, which are freely available in the `mdhglm` CRAN repository [15]. The trial was part of the Community Programs for Clinical Research on AIDS (CPCRA) and funded by the National Institute of Allergy and Infectious Diseases. The data and code used to generate the results presented in this section can be found in the Supplementary material to this paper.

5.1 | Study population

We analyzed 467 HIV patients that had presented zidovudine intolerance and were therefore deeply immunosupressed. The recruited patients were enrolled in a two-arms clinical trial and randomly assigned either to didanosine or zalcitabine. The aim of the trial was to determine disease progression and survival after randomization. Patients were followed-up for a maximum of 21 months. Figure 2 shows the estimated survival curves per treatment throughout the entire follow-up period. Further information on the data can be found in [16].

5.2 | Results

We estimated the CCE of time-to-death with respect to gender (β_1), pre-existing AIDS condition (β_2), CD4 cell count at baseline (β_3) and treatment (β_4), assuming an underlying Cox-Weibull model. The proportionality assumption was tested using Schoenfeld residuals against transformed time, obtaining a global p-value of 0.35. Marginal tests supported also the proportional hazards assumption for all considered covariates. The chosen grid of proportions $\boldsymbol{\tau}$ divided the population by deciles.

As shown in Table 4 CCE estimates increased considerably with k, but estimates for large k where far less reliable as noted by the increasing standard errors. Nonetheless, estimates from lower components showed a better precision. Interpretation of the reported results indicates, for example, that the weakest 10% of enrolled patients to die, have a life expectancy of 1.48 years after

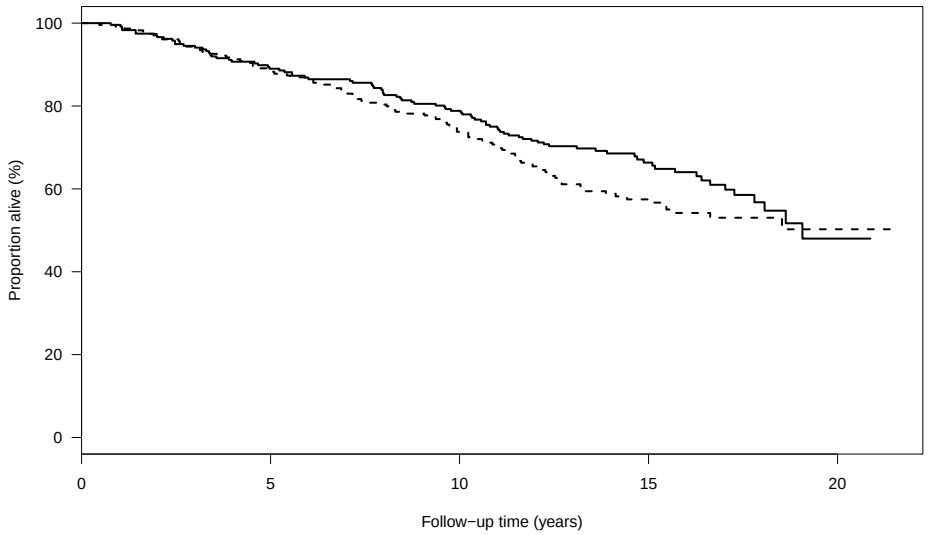

FIGURE 2 Kaplan Meier estimated survival curves by treatment (dashed line didanosine and solid line zalcitabine)

TABLE 4 Estimated CCE by deciles ($\hat{\mu}_k$) under Cox proportional hazards model assumption with baseline Weibull distribution, with log hazard ratios β_1 for gender (reference female), β_2 for AIDS precondition (reference no-AIDS), β_3 for CD4 cell count at baseline, and β_4 for treatment applied (reference didanosine). Bootstrapped standard errors (se) for he CCE estimates are also reported.

Baseline CCE										
k	1	2	3	4	5	6	7	8	9	10
$\hat{\mu}_k$	0.148	0.213	0.303	0.430	0.614	0.888	1.319	2.044	3.428	6.526
se	0.03	0.03	0.05	0.09	0.17	0.53	0.75	0.98	1.8	3.45

Model parameters						
Parameter	β_1	β_2	β_3	β_4	λ	υ
Estimated value	−0.245	0.864	−0.150	−0.238	0.017	1.469
se	0.147	0.215	0.242	0.024	0.006	0.301

enrollment, while our estimation showed that the strongest 10% would have a life expectancy of 65.26 years. The first statement is based on observed data, while the latter is mainly based on the parametric modelling assumption and has far less precision.

Effect of the different covariates on CCE estimates can also be computed (see equation (9)). For example, the multiplicative factor difference in CCE by treatment for all components was

$$\exp(0.238)^{1/1.469} = 1.18,$$

that is, those who were assigned zalcitabine lived, on average, 18% longer compared to those under didanosine (all other factors equal).

6 | FINAL REMARKS

In this paper, we have shown the advantages of the CCE for estimating life expectancy in the presence of censored data and for a given set of covariates of interest. One the main advantages of the CCE over other methods, (for instance, the restricted mean) is that its estimates have a clear interpretation in terms of life expectancy that relate to well-defined fractions of the population. Our method also exploits the information contained in the data and is able to report more precise estimates. Dividing the population

in fractions, allows obtaining precise estimation for lower components (corresponding to lower quantiles), a feature that is lost when considering all the available data at once.

Because CCE estimation is possible under any suitable model of choice for the data, other survival scenarios could also be considered. For example, all models included here assumed T to be conditionally independent from the censoring variable C, but any competing risks model could be adapted to our framework.

An interesting feature of our approach that has not been explored here, would entail the use of models that include time-varying covariates, so that different effects of distinct components could also be detected. Some of these models of interest are, for example, flexible parametric survival models that include time-varying covariate effects[11,12], Aalen's additive hazard model[17] or flexible parametric quantile models[18].

References

1. Lubitz J, Cai L, Kramarow E, Lentzner H. Health, life expectancy, and health care spending among the elderly. *New England Journal of Medicine* 2003; 349(11): 1048–1055.

2. Christensen K, Doblhammer G, Rau R, Vaupel JW. Ageing populations: the challenges ahead. *The lancet* 2009; 374(9696): 1196–1208.

3. Sabin CA. Do people with HIV infection have a normal life expectancy in the era of combination antiretroviral therapy?. *BMC medicine* 2013; 11(1): 251.

4. Mocroft A, Lundgren JD, Sabin ML, others . Risk factors and outcomes for late presentation for HIV-positive persons in Europe: results from the Collaboration of Observational HIV Epidemiological Research Europe Study (COHERE). *PLoS medicine* 2013; 10(9): e1001510.

5. Bisson GP, Gaolathe T, Gross R, et al. Overestimates of survival after HAART: implications for global scale-up efforts. *PloS one* 2008; 3(3): e1725.

6. Irwin J. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *The Journal of hygiene* 1949; 47(2): 188.

7. Karrison T. Restricted Mean Life with Adjustment for Covariates. *Journal of the American Statistical Association* 1987; 82(400): 1169-1176. doi: 10.1080/01621459.1987.10478555

8. Royston P, Parmar MK. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in medicine* 2011; 30(19): 2409–2421.

9. García-Pareja C, Bottai M. On mean decomposition for summarizing conditional distributions. *Stat* 2018; 7(1): e208. doi: 10.1002/sta4.208

10. Cox C, Chu H, Schneider MF, MuÃśoz A. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statist. Med.* 2007; 26(23): 4352–4374. doi: 10.1002/sim.2836

11. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statist. Med.* 2002; 21(15): 2175–2197. doi: 10.1002/sim.1203

12. Crowther MJ, Lambert PC. A general framework for parametric survival analysis. *Statist. Med.* 2014; 33(30): 5280–5297. doi: 10.1002/sim.6300

13. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statist. Med.* 2005; 24(11): 1713–1723. doi: 10.1002/sim.2059

14. Brostrom G. *Event History Analysis*. 2018. R package version 2.6.

15. Lee Y, Molas M, Noh M. *Multivariate Double Hierarchical Generalized Linear Models*. 2018. R package version 1.8.

16. Abrams DI, Goldman AI, Launer C, et al. A Comparative Trial of Didanosine or Zalcitabine after Treatment with Zidovudine in Patients with Human Immunodeficiency Virus Infection. *New England Journal of Medicine* 1994; 330(10): 657-662. PMID: 7906384doi: 10.1056/NEJM199403103301001

17. Aalen O. Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics* 1978; 6(4): 701–726.

18. Frumento P, Bottai M. Parametric modeling of quantile regression coefficient functions with censored and truncated data. *Biometrics* 2017; 0. Preprintdoi: 10.1111/biom.12675