Backward multiple imputation estimation of the conditional lifetime expectancy function with application to censored human longevity data

Jing Kong[a], Barbara E. K. Klein[b], Ronald Klein[b], and Grace Wahba[a,c,d,1]

[a]Department of Statistics, University of Wisconsin–Madison, Madison, WI 53706; [b]Department of Ophthalmology, University of Wisconsin–Madison, Madison, WI 53706; [c]Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, WI 53706; and [d]Department of Computer Sciences, University of Wisconsin–Madison, Madison WI 53706

Contributed by Grace Wahba, July 7, 2015 (sent for review June 7, 2015; reviewed by Pang Du and Ping Ma)

The conditional lifetime expectancy function (LEF) is the expected lifetime of a subject given survival past a certain time point and the values of a set of explanatory variables. This function is attractive to researchers because it summarizes the entire residual life distribution and has an easy interpretation compared with the popularly used hazard function. In this paper, we propose a general framework of backward multiple imputation for estimating the conditional LEF and the variance of the estimator in the right-censoring setting. Simulation studies are conducted to investigate the empirical properties of the proposed estimator and the corresponding variance estimator. We demonstrate the method on the Beaver Dam Eye Study data, where the expected human lifetime is modeled with smoothing-spline ANOVA given the covariates information including sex, lifestyle factors, and disease variables.

lifetime expectancy | imputation | smoothing-spline ANOVA | right-censored survival data | human longevity

Survival analysis has focused on the popular hazard function for decades, and one of the most famous models is Cox's proportional hazard model (1). However, the hazard function, defined as the risk of immediate failure, can be conceptually difficult to understand. The expected or remaining lifetimes are intuitively more attractive because of the easy interpretation and turn out to be a more relevant metric under many circumstances. For example, it is more transparent to patients if the doctor explains it as "on average, the lifetime is expected to be 80 y if one, also at 70 y with similar demographic and healthy background like you, takes this treatment," rather than in the language of "the average hazard is expected to decrease by 25% among the treated patients similar to you." Furthermore, in the analysis of reliability and actuarial data, a life insurance company may care more about the life expectancy of a person, and an engineering firm might want to know the expected remaining lifetime of a system given survival past certain time. This motivates us to focus more attention on direct estimation of key summary measures regarding remaining lifetimes. This paper targets lifetime expectancy function and the mean residual life function.

The lifetime expectancy function (LEF) of a survival time T (with $T>0$), denoted by $e(t)$, is defined as

$$e(t)=E(T|T>t)=t+\int_t^{\tau_T}\frac{S(u)}{S(t)}du,$$

where $S(t)=P(T>t)$ is the survival function and $\tau_T=\inf\{t:S(t)=0\}$. Denote $m(t)$ the mean residual life function (MRLF), which is the expected remaining lifetime given survival up to time t and

$$m(t)=e(t)-t=E(T-t|T>t).$$

$e(t)$ uniquely determines $S(t)$ as the following equation shows (2):

$$S(t)=\frac{e(0)}{e(t)-t}\exp\left\{-\int_0^t[e(u)-u]^{-1}du\right\}.$$

Ref. 2 provides necessary and sufficient conditions, such that $m(t)$ is a proper MRLF [or that $e(t)$ is a proper LEF]. That is, $F(t)=P(T\le t)$ is a proper continuous distribution function if and only if $m(t)$ satisfies:

1. $m(t)\ge 0$ for all $t\ge 0$;
2. $e(t)=m(t)+t$ is nondecreasing in t;
3. if there exists a τ such that $m(\tau)=0$ then $m(t)=0$ for all $t\ge\tau$, otherwise, $\int_0^\infty m(t)^{-1}dt=\infty$;
4. $m(t)$ is a right continuous function and has a left limit with positive increments at discontinuities.

In practice, real data always contain additional information besides the survival time itself and researchers are interested in how the variables contribute to lifetimes. This is when the conditionality of LEF or MRLF plays a role. For example, in the context of mobile devices, modeling the conditional LEF that the users keep active with certain apps or games after installation helps the providers target and stratify their customers, and offers insights about the effectiveness of different features related to the product. In the situation of property purchase, it is of interest to both seller and buyer to know how long it takes for a house to be sold after being listed for sale by a certain agent or on a real

Significance

The expected lifetime of a subject given survival past a certain time, denoted as lifetime expectancy, is often estimated in the data context with right censoring, which is a form of missing data problem commonly arising in biomedical applications, e.g., clinical trials, meaning that time-to-event is observed only if it occurs prior to some prespecified time. We report an advanced and more flexible method where users are free to choose a base model to estimate lifetime expectancy by imputing the right-censored times in backward order to address data "missingness." We use this innovative tool to explore the interesting and important issue of human aging providing individual attributes including gender, smoking, body mass index, socioeconomic status, and diseases.

Author contributions: B.E.K.K. and R.K. designed research; B.E.K.K. and R.K. performed research; J.K. and G.W. contributed new reagents/analytic tools; J.K. and G.W. analyzed data; and J.K. and G.W. wrote the paper.

Reviewers: P.D., Virginia Polytechnic Institute and State University; and P.M., University of Georgia.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: wahba@stat.wisc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1512237112/-/DCSupplemental.

Table 1. Variable description in the SS-ANOVA model

Variable	Units	Description
lastage	years	Censored age at death
survflag	yes/no	Survival indicator
baseage	years	Age at baseline
Sex	F/M	Sex
edu	years	Highest year school/college completed
BMI	kg/m^2	Body mass index
Smoke	yes/no	History of smoking
Income	yes/no	Household personal income > 20 K
Diabetes	yes/no	History of diabetes
Cancer	yes/no	History of cancer
Heart	yes/no	History of cardiovascular disease
Kidney	yes/no	History of chronic kidney disease

estate website considering the size, building year, location, and estimated price of the house. Moreover, as we will depict in our real data analysis, lifestyle factors such as smoking and socioeconomic status, disease, and healthy metrics are all informative toward one's expected lifespan.

In this paper, we propose a framework for estimating the conditional LEF $e(t|x)=E(T|T>t,X=x)$ when covariates X information is available and the survival times are subject to right censoring. Following the same idea with the Buckley–James estimator (3) to address censoring by imputation, our method replaces the censored survival times in backward order with a heuristic guess of a fitted LEF using a user-specific base model and the covariates. One is then able to model LEF with a completely imputed dataset. We provide variance estimation and confidence interval for the estimated LEF based on the idea of multiple imputation (4). When there is no covariate, our estimator is proven to be the same as the one derived by inverting the Kaplan–Meier estimator for the survival function (5). Considerable research has been done on estimation of the conditional MRLF (6, 7). Ref. 8 discussed different semiparametric conditional MRLF estimations and ref. 9 covered nonparametric estimation for MRLF with covariates; we show that this method is equivalent to our framework by choosing kernel regression as the base model. We investigate the behaviors of our proposed estimator in practical settings via different simulation studies. Finally, we demonstrate our method to model human lifetimes with the Beaver Dam Eye Study data (10), where survival information and a number of useful variables, from demographic records to medical measurements, are included.

Semiparametric and Nonparametric Estimation of Conditional MRL Function

There are several papers in the literature that discuss how to estimate MRL function $m(t|x)$ with right censoring conditional on $x=(x_1,\cdots,x_p)^T$, which is the p-dimensional vector of explanatory variables. It is easy to obtain the corresponding LEF $e(t|x)$ by $t+m(t|x)$. First, ref. 8 considered the semiparametric proportional MRL model

$$m^p(t|x)=m_0^p(t)\exp(\beta^T x),$$

where $m_0^p(t)$ is a baseline MRL function and β is a p-dimensional vector of regression coefficients. Ref. 6 proposed to estimate $m(t|x)$ as an additive expectancy regression model. The model takes a semiparametric form of

$$m^a(t|x)=m_0^a(t)+\gamma^T x,$$

where $m_0^a(t)$ is a baseline MRL function and γ is a p-dimensional vector of regression coefficients. Ref. 7 framed the general family of semiparametric transformation models

$$m^g(t|x)=g\{m_0(t)+\beta^T x\},$$

which includes the previous proportional and additive models as special cases.

As discussed in ref. 9, the nondecreasing property of $e(t|x)$ may be violated under the existing semiparametric models. The authors in ref. 9 considered taking a different perspective to satisfy this natural constraint. They first calculate the nonparametric estimation $\hat{S}_P(t|x)$ of the conditional survival function using a generalized Kaplan–Meier estimator according to refs. 11, 12, and then take the inversion to obtain the nonparametric estimator $\hat{m}_P(t|x)$ for the conditional MRL function. A smoothed estimation of MRL is available by inverting the smoothed $\hat{S}_P(t|x)$ based on Bernstein polynomials. It is straightforward that $\hat{m}_P(t|x)$ is a valid MRL function because $\hat{S}_P(t|x)$ is a well-defined survival function.

Backward Multiple Imputation Framework for Estimating LEF

Backward Imputation Without Covariates. Let us first consider the cases without any covariate to intuitively understand the idea. Let T be a continuous nonnegative random variable and C be the censoring variable. We assume that T and C are independent. The observed data set consists of n independent and identically distributed (i.i.d.) replicates of $(Y_i,\delta_i), i=1,\cdots,n$, where $Y_i=\min(T_i,C_i)$ and $\delta_i=I_{\{T_i\le C_i\}}$ is the censoring indicator. Let $y_{(1)}<\cdots<y_{(M)}$ be the distinct ordered values of the n observations and $n_1,\cdots,n_M$ be the corresponding number of observations taking each specific value of $y_{(i)}, i=1,\cdots,M$. Denote $t_{(1)}<\cdots<t_{(K)}$ and $c_{(1)}<\cdots<c_{(J)}$ the distinct ordered event times and censored times, respectively. The notation $n(t_{(k)})$ or $n(c_{(j)})$ takes the number of observations at $t_{(k)}$ or $c_{(j)}$.

The $c_{(j)}$ s are right censored and we know that the true values should be greater than the censored times $c_{(j)}$. One reasonable

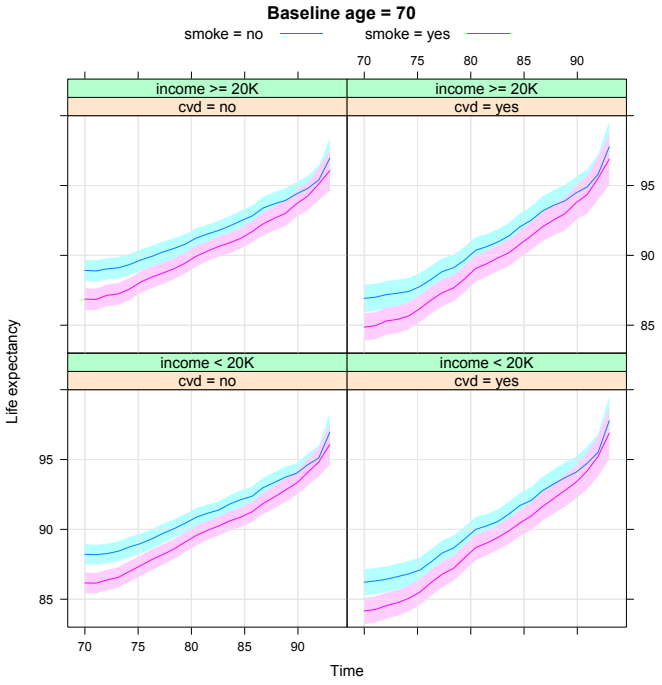

Fig. 1. Lifetime expectancy function estimation by smoking, heart disease, and income for the group with baseage = 70, sex = F, BMI = 28 (median of the population), edu = 12 (median of the population) and no other disease. The x axis is time t from 70 to 93. The y axis is $\hat{e}(t|X=x)$. The shaded area presents 95% normal confidence intervals.

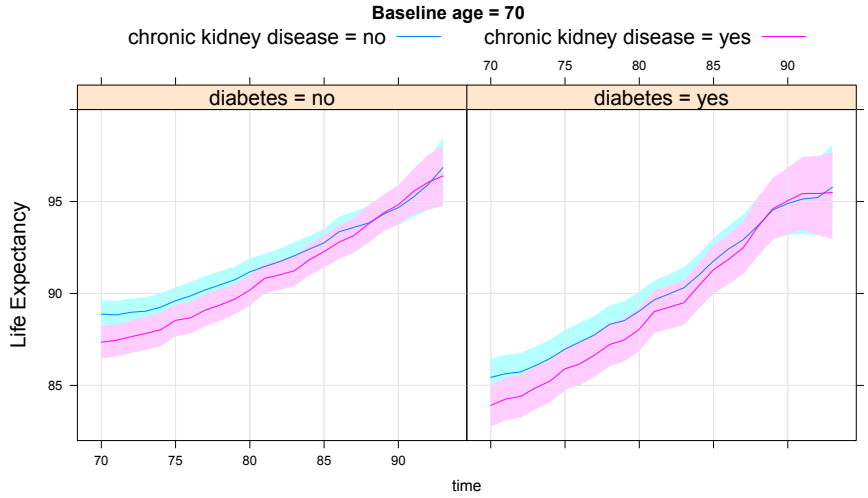

Fig. 2. LEF estimation by diabetes and chronic kidney disease for subjects with baseage = 70, sex = F, smoke = no, income ≥ 20 K, BMI = 28, edu = 12, and no heart disease, cancer, or stroke. The x axis is time t from 70 to 93. The y axis is $\hat{e}(t|X=x)$. The shaded area presents 95% normal confidence intervals.

guess for the true values is the lifetime expectancy at $c_{(j)}$, i.e., $e(c_{(j)}) = E(T|T > c_{(j)})$. This is the same idea as the single imputation of Little and Rubin (13) and as the Buckley–James estimator (3). We could use the sample lifetime expectancy, which is the mean of the observations greater than $c_{(j)}$, as an estimate for $e(c_{(j)})$. However, this does not work if censored data still exist to the right of the targeted $c_{(j)}$. We can address this problem by processing our guessing regime for $c_{(j)}$ s backwardly from J to 1. After imputing the censored values, it is easy to obtain sample lifetime expectancy at any time point t. The detailed steps are as follows:

Algorithm 1: Backward imputation without covariates

1. We do nothing if $c_{(J)}$ is the largest value in the dataset, i.e., $c_{(J)} = y_{(M)}$. Otherwise, we estimate $e(c_{(J)})$ by the sample mean of the observations beyond $c_{(J)}$, i.e.,

$$\hat{e}_B(c_{(J)}) = \frac{\sum_{i=1}^{n} y_i I_{\{y_i > c_{(J)}\}}}{\sum_{i=1}^{n} I_{\{y_i > c_{(J)}\}}} = \frac{\sum_{k=1}^{K} t_{(k)} n(t_{(k)}) I_{\{t_{(k)} > c_{(J)}\}}}{\sum_{k=1}^{K} n(t_{(k)}) I_{\{t_{(k)} > c_{(J)}\}}}.$$

   Replace $c_{(J)}$ by $\hat{e}_B(c_{(J)})$ and treat it as observed.

2. Repeat the above procedure backwardly for $j = J-1, \cdots, 1$ to replace $c_{(j)}$ by $\hat{e}_B(c_{(j)})$, which is the sample mean of the observations beyond $c_{(j)}$ in the imputed data. Because the process runs for j from J to 1, we will have imputed all of the censored values greater than $c_{(j)}$ and there is no "missingness" to estimate $e(c_{(j)})$ by the sample mean of the observations larger than $c_{(j)}$.
3. Let $\tilde{y}_1, \cdots, \tilde{y}_n$ be the data after backward imputation procedure. If y_i is observed or it is the largest observation and is censored, then $\tilde{y}_i = y_i$. Otherwise, y_i is one of the censored times and $\tilde{y}_i = \hat{e}_B(y_i)$. The backward procedure only obtains estimates of $e(t)$ at the censored times. In general, we estimate $e(t)$ for $t \geq 0$ by the following formula:

$$\hat{e}_B(t) = \frac{\sum_{i=1}^{n} \tilde{y}_i I\{y_i > t\}}{\sum_{i=1}^{n} I\{y_i > t\}}.$$

Relationship with Kaplan–Meier Estimator. Another way to obtain an estimator for $e(t)$ is by inverting an estimator for $S(t)$. We know that Kaplan–Meier estimator $\hat{S}_{KM}(t)$ is the MLE for $S(t)$ with respect to the empirical likelihood. Denote $\hat{e}_{KM}(t)$ the estimate for $e(t)$ by inverting $\hat{S}_{KM}(t)$. The following theorem proves the equivalence between $\hat{e}_B(t)$ and $\hat{e}_{KM}(t)$. This also demonstrates the equivalence between the spirit of backward imputation and the idea of redistribution-to-the-right to estimate survival function established by ref. 14.

Theorem 1. *Let T be a continuous nonnegative random variable which is independent of the censoring variable C. We observe n i.i.d. replicates of $(Y_i, \delta_i), i = 1, \cdots, n$, where $Y_i = \min(T_i, C_i)$ and $\delta_i = I_{\{T_i \leq C_i\}}$. Denote $\hat{e}_B(t)$ the backward imputation estimator for $e(t)$ as described in Algorithm 1 and $\hat{e}_{KM}(t)$ the inverted Kaplan–Meier estimator for $e(t)$, which takes the following explicit form*:

$$\hat{e}_{KM}(t) = \begin{cases} t_{(k)} + \dfrac{\sum_{l=k+1}^{K} (t_{(l)} - t_{(l-1)}) \hat{S}_{KM}(t_{(l-1)})}{\hat{S}_{KM}(t_{(k-1)})}, \\ \quad \text{if} \quad t_{(k-1)} < t < t_{(k)}; \\ t_{(k)} + \dfrac{\sum_{l=k+1}^{K} (t_{(l)} - t_{(l-1)}) \hat{S}_{KM}(t_{(l-1)})}{\hat{S}_{KM}(t_{(k)})}, \\ \quad \text{if} \quad t = t_{(k)}, k = 1, \cdots, K-1; 0, \ \text{if} \quad t \geq t_{(K)}. \end{cases}$$

Then $\hat{e}_B(t) = \hat{e}_{KM}(t)$ for $t \geq 0$.

Backward Imputation with Covariates. We want to make use of the covariates information for estimating LEF. We assume the censoring to be conditionally independent of the survival time given the covariates $X = x$. Now our observations are n i.i.d samples $(Y_i, \delta_i, x_i), i = 1, \cdots, n$, where $Y_i = \min(T_i, C_i)$ and $\delta_i = I_{\{T_i \leq C_i\}}$. Suppose we have a base regression model $f(x) = E(T|X = x)$ that uses the covariates information to predict the mean survival times when there is no censoring. We substitute the sample mean in the previous backward imputation procedure by the base regression model. This means that we treat the estimate for $e(c_{(j)}|x) = E(T|T > c_{(j)}, X = x)$ as our guess for the censored case $c_{(j)}$ with its covariates x. The following algorithm illustrates the detailed steps.

Algorithm 2: Backward imputation with covariates.

1. We do nothing if $c_{(J)}$ is the largest response value in the dataset, i.e., $c_{(J)} = y_{(M)}$. Otherwise, we obtain the fitted model $\hat{f}$ using the observations $\{(y_i, x_i) | y_i > c_{(J)}\}$. Note that all of the observations with $y_i > c_{(J)}$ should be uncensored in this step by the definition of $c_{(J)}$. Replace $c_{(J)}$ by $\hat{f}(x_0)$, where x_0 represents the observed covariates values for $c_{(J)}$, and treat it as observed.

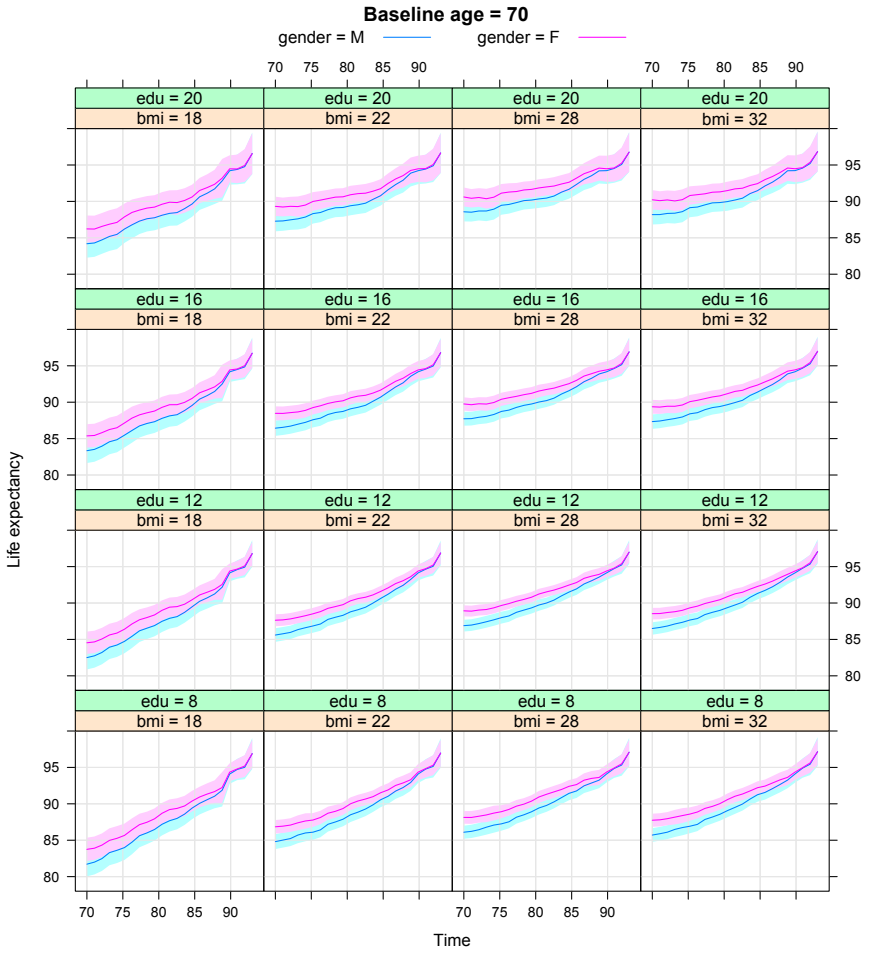

Fig. 3. LEF estimation by BMI, edu, and sex for the subgroup with baseage = 70, smoke = no, income ≥ 20 K, and no disease. The *x* axis is time *t* from 70 to 93. The *y* axis is $\hat{e}(t|X=x)$. The shaded area presents 95% normal confidence intervals.

2. Repeat the above procedure backwardly for $j=J-1,\cdots,1$ with the imputed data.
3. Let $\tilde{y}_1,\cdots,\tilde{y}_n$ be the data after backward imputation procedure. Obtain the fitted base model $\hat{f}$ using the data $\{(\tilde{y}_i,x_i)|y_i>t\}$ and we estimate $e(t|x)$ by $\hat{f}(x)$.

There are several advantages of this framework. One is that it allows time-varying effects of the covariates because we obtain the fitted $e(t|x)$ restricted to the subset of the data with the original censored survival time greater than the time point t. Another flexibility about this procedure is the freedom to choose the base model f that describes the data the best. The following theorem, with proof in *SI Appendix*, states that implementing kernel regression as the base model in backward imputation procedure is equivalent to the nonparametric estimator $\hat{e}_P(t|x)$ proposed by McLain and Ghosh (9). This also implies that $\hat{e}_P(t|x)$ shares the similar pros and cons to kernel regression. For example, one has to take care with the choice of kernel, the contamination in the distance due to irrelevant variables, and curse of dimensionality. One is able to address these issues by applying more appropriate base models in backward imputation method to accommodate different datasets.

Theorem 2. *Let $K:\mathbb{R}^p\to\mathbb{R}$ be the p-dimensional kernel function and h_n denote the bandwidth. Let $\hat{e}_B(t|x)$ be the estimator for $e(t|x)$ from backward imputation using kernel regression with K and h_n, and $\hat{e}_P(t|x)$ be the nonparametric estimator of the conditional LEF proposed in ref. 9 using the same K and h_n. Then $\hat{e}_B(t|x)=\hat{e}_P(t|x)$ for $t\geq 0$ given x.*

Variance Estimation with Multiple Imputation. The methods illustrated above are in the fashion of single imputation, which does not take into account the uncertainty about the predictions of the unknown censored values. It is likely that the variance estimation for $\hat{e}_B(t|x)$ is biased toward zero. We incorporate the idea of multiple imputation procedure (4) in our proposed method. Instead of filling in the conditional expected values for each

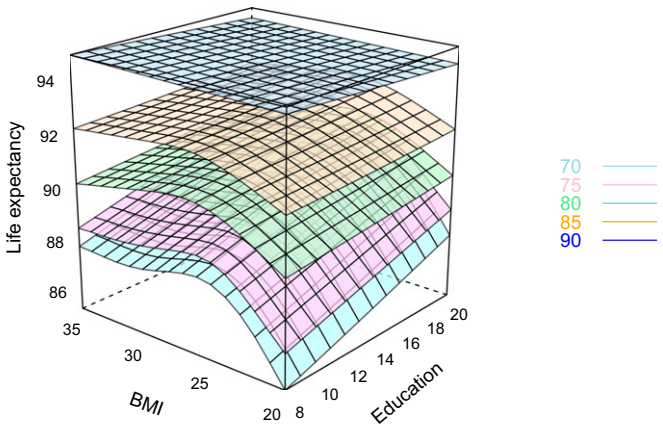

Fig. 4. BMI and edu effects on expected lifetime for baseage = 70, sex = F, smoke = no, income ≥ 20 K, and no disease with $t=70, 75, 80, 85$, and 90.

censored value as described above, we replace by a random sample drawn from the posterior predictive distribution under the base model each time. It introduces randomness that represent the uncertainty about the right value to impute. We repeat the backward multiple imputation a number of times and the results are combined finally to obtain a valid variance estimation and confidence interval for the estimate of conditional LEF. The procedures are shown below.

Algorithm 3: Backward multiple imputation with covariates.

1. Set up the number of multiple imputation m. For each replication, repeat steps 2–4.
2. We do nothing if $c_{(J)}$ is the largest response value in the dataset, i.e., $c_{(J)} = y_{(M)}$. Otherwise, we obtain the fitted model $\hat{f}$ using the observations $\{(y_i, x_i) | y_i > c_{(J)}\}$. Note that all of the observations with $y_i > c_{(J)}$ should be uncensored in this step by the definition of $c_{(J)}$. Replace $c_{(J)}$ by a random sample from the posterior predictive distribution of the fitted model at x_0, where x_0 represents the observed covariates values for $c_{(J)}$, and treat it as observed.
3. Repeat the above procedure backwardly for $j = J-1, \cdots, 1$ with the imputed data.
4. Let $\tilde{y}_1, \cdots, \tilde{y}_n$ be the data after backward imputation procedure. Obtain the fitted base model $\hat{f}$ using the data $\{(\tilde{y}_i, x_i) | y_i > t\}$ and we estimate $e(t|x)$ by $\hat{f}(x)$. Moreover, keep record of the estimated variance for $\hat{f}(x)$.
5. With m imputations, one collects m different sets of the point and variance estimates for $e(t|x)$. Let $\hat{Q}_i$ and $\hat{U}_i$ be the point and variance estimates of $e(t|x)$ from the ith imputed data set, $i = 1, \cdots, m$. Note that $\hat{Q}_i$ and $\hat{U}_i$ are functions of x and we eliminate the dependency on x in the notation for simplicity.
6. The point estimate for $e(t|x)$ from multiple imputations is $\overline{Q} = 1/m \sum_{i=1}^{m} \hat{Q}_i$.
7. Let $\overline{U}$ be the within-imputation variance and B be the between-imputation variance $\overline{U} = 1/m \sum_{i=1}^{m} \hat{U}_i$, $B = 1/(m-1) \sum_{i=1}^{m} (\hat{Q}_i - \overline{Q})^2$. The variance estimation for the estimated $e(t|x)$ is the total variance $T = \overline{U} + (1 + (1/m))B$.
8. The statistic $(Q - \overline{Q}) T^{-1/2}$ is approximately distributed as a t distribution with degrees of freedom $v_m = (m-1)[1 + (\overline{U}/(1 + m^{-1})B)]^2$. When v_m is large, one may approximate by a normal distribution. The confidence interval for $e(t|x)$ can be derived accordingly.

A brief illustration about the idea of multiple imputation and simulation studies to demonstrate the effectiveness of the backward multiple imputation method are presented in *SI Appendix*.

Application to Beaver Dam Eye Study Data

Data Description. The Beaver Dam Eye Study (BDES) (10) is an ongoing population-based study of age-related ocular disorders with 5-, 10-, 15-, and 20-y follow-ups. Subjects at baseline, examined between 1988 and 1990, were a group of 4,926 people aged 43–86 y from Beaver Dam, Wisconsin. The survival statuses, including ages at death, for this population were updated by 31 Dec 2013 with 2,014 individuals who were alive. BDES provides us an excellent opportunity to study the lifetime expectancy with our proposed methods.

A number of variables, including measurements on individual health and lifestyles, were recorded in the study. We took advantage of a couple of the most important ones which were used in ref. 15 to examine the association with human mortality. To maintain the largest sample sizes, we focused on the baseline data. Table 1 lists the description of all of the variables involved in the study. A number of variables with weak signals for longevity are discussed in *SI Appendix*.

Model Fitting and Results. The smoothing-spline ANOVA model (16–18) has a successful history in modeling BDES data (15, 19). Our base model is an SS-ANOVA model with the following form:

$$\begin{aligned}(imputed)\ lastage\ =\ & \mu + f_1(baseage) + \beta_{gender} I_{\{gender=F\}} \\ & + f_2(edu) + f_{12}(baseage : edu) \\ & + f_3(BMI) + \beta_{smoke} I_{\{smoke=no\}} \\ & + \beta_{inc} I_{\{inc>20K\}} + \beta_{diabetes} I_{\{diabetes=no\}} \\ & + \beta_{cancer} I_{\{cancer=no\}} + \beta_{heart} I_{\{heart=no\}} \\ & + \beta_{kidney} I_{\{kidney=no\}} \qquad (*)\end{aligned}$$

Functions f_1, f_2, and f_3 are cubic splines and f_{12} uses the tensor product construction. The remaining covariates are unpenalized and modeled as linear terms with $I_{\{\cdot\}}$ as indicator functions. We incorporated this base model in *Algorithm 3* with multiple imputation replications $m = 200$ to estimate the conditional LEF in the population of BDES. One adjustment we applied for *Algorithm 3* was that we used model (*) if the sample size involved in step 2 of *Algorithm 3* was greater than 100; otherwise, we simply used sample mean of ages of death among all of the samples involved in this step. We observed that both the estimations for LEF and the variance estimation became stable after 20 multiple imputations.

Figs. 1–4 display the predicted conditional LEFs for the cohort with baseline age of 70. In Fig. 1, we examine the effects of smoking, cardiovascular disease, and income for the subgroup of females with midvalued body mass index (BMI), education, and no other disease. The natural constraint of monotonic nondecreasing over t for $e(t|x)$ is closely satisfied in practice. From the plots, it appears that smoking and having a history of heart disease have negative influences on longevity in this population. Higher household income slightly protects longevity. Fig. 2 discovers how diabetes and chronic kidney disease change expected survival given the rest of covariates. It turns out that diabetes is a strong risk factor that reduces human longevity. Chronic kidney disease, although not as harmful as diabetes, also exerts a negative effects on survival times among this subgroup of people.

In Fig. 3, we present how the expected lifetime changes with BMI, education, and sex for nonsmoking rich and healthy individuals with baseage of 70. The plots suggest that females tend to have longer lifespans compared with males. Higher education and midvalued BMI are protective for longevity. The covariates effects fade out as t gets large with several possible reasons. First, the sample size is limited when restricting to subjects over 85. Second, it is likely that those long-lived individuals have survived from the risk factors so that we could not find the significance for the covariates. Notice that the nondecreasing constraint of LEF can be violated when the sample size is small, as may be just barely noticeable upon very close inspection of panels 3 and 4 on the top row near the tail.

Fig. 4 takes a different perspective from the previous three plots and focuses on the two continuous variables BMI and education for a cohort of rich and healthy female nonsmokers who entered the study when they were 70. The five surfaces correspond to five time points, $t = 70, 75, 80, 85$ and 90. Each surface represents the estimated expected lifetime across different values of BMI and education. When t is small, we observe the quadratic influence of BMI where very low BMI values are very

Table 2. Comparison of the estimates of $e(t|x)$ and its estimated SD by bootstrap and backward multiple imputation

Quantiles	$Q_{0.1}$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$Q_{0.9}$
$\frac{\hat{e}_{BM}(t\|x)}{\hat{e}_{BOOT}(t\|x)}$	0.9971	0.9987	1.0006	1.0031	1.0053
$\frac{\widehat{std}\{\hat{e}_{BM}(t\|x)\}}{std\{\hat{e}_{BOOT}(t\|x)\}}$	0.7561	0.8694	0.9897	1.0929	1.1959

harmful and the optimal value happens at around 26 or 27 and tails down slowly for higher values. Note that Beaver Dam is a small town in the Midwest and may not be representative of some population groups in other areas of the country. The education displays a monotonic increasing effect on lifetime in this cohort. The higher the completed education is, the longer the expected lifetime is. When t gets large, we again find that the influences of BMI and education disappear. More results about some other weakly related variables and other baseline age cohorts are discussed in *SI Appendix*.

Validation Using Bootstrapped Samples. This is an observational study and the true $e(t|x)$ is unknown. We used the bootstrap method to get the empirical distributions of $e(t|x)$ for different values of t and x to check if the results coming from backward multiple imputation match the mean and SD of the empirical distribution. The following steps cover the bootstrap details.

1. Obtain bootstrap samples by resampling with replacement.
2. Use backward imputation, *Algorithm 2*, with SS-ANOVA on the bootstrapped samples.
3. Estimate $e(t|x)$ with the imputed bootstrap data for the combinations of t and x used to generate Figs. 1–3.
4. Repeat steps 1–3 for 1,000 times to get empirical distribution of $e(t|x)$ for each combination of t and covariates values.

From the above bootstrap procedure, we obtained estimated mean and SE of $e(t|x)$ from the empirical distributions, denoted as $\hat{e}_{BOOT}(t|x)$ and $\widehat{std}\{\hat{e}_{BOOT}(t|x)\}$, respectively. We could compare those with the ones derived from the previous backward multiple imputation, denoted as $\hat{e}_{BM}(t|x)$ and $\widehat{std}\{\hat{e}_{BM}(t|x)\}$. For baseline age of 70, there were 6,400 combinations of t and covariates values where t runs from 70 to 93. Table 2 summarizes the differences between the ratios of the two estimators and the ratios of the two corresponding estimated SDs. The ratios between the estimators are closely centered at 1. The ratios between the two SDs spread out a little bit but still concentrate around 1, meaning that $\hat{e}_{BM}(t|x)$ and $\widehat{std}\{\hat{e}_{BM}(t|x)\}$ match the empirical distribution fairly well. Furthermore, the bootstrapped distributions of $\hat{e}(t|x)$ are very much close to normal distribution; for details see *SI Appendix*, Fig. S9, which validates our use of normal confidence intervals.

Discussion

In this article, we presented our backward multiple imputation framework for estimating the conditional LEF. In the case without covariates, our estimator is proven to be equivalent to the estimation for LEF by inverting the Kaplan–Meier survival function estimator. In the case with covariates, one is free to select a base model that best captures the data. One is able to recover the nonparametric estimator for conditional LEF proposed in ref. 9 based on the generalized Kaplan–Meier estimator by using kernel regression in our framework. The simulation studies demonstrated the performance of our methods and validated the use of multiple imputation for variance estimation under three different settings. The application to the Beaver Dam Eye Study data illustrated the use of SS-ANOVA model together with our backward multiple imputation method. We presented the fitted results for the cohorts with baseline age of 70 where a number of variables, including sex, smoking, education level, BMI values, and several diseases, were shown to be significantly associated with human longevity.

There are a couple of issues that we will consider as our future direction. First, as pointed out by ref. 9, many existing models for estimating MRLF or LEF do not satisfy the nondecreasing property of $e(t|x)$. We know that kernel regression ensures the validation of this condition. The real application using SS-ANOVA in BDES data provided satisfactory nondecreasing curves in most situations, with slight violation when sample sizes were small. It is of our practical and theoretical interest to explore what base models guarantee this property as well. One possible solution is to impose constraints in the backward imputation procedure so that the predicted $e(t_1|x) \le e(t_2|x)$ for $t_1 \le t_2$. In this paper, we discussed the use of multiple imputation to obtain the variance estimation for the estimated LEF. Still more research about the other ways to construct the variance estimator under certain base models, including asymptotic distributions of the LEF estimator, is needed to reduce the burden in computation. Please see Supporting Information regarding access to data and computer code.

ACKNOWLEDGMENTS. This work was supported by NIH Grant EY09946 and National Science Foundation Grant DMS1308847 (to J.K. and G.W.), NIH Grant EY06694 and Research to Prevent Blindness Senior Scientist-Investigator Awards, New York (to R.K. and B.E.K.K.), and a Consortium of NIH Institutes under Award U54AI117924 (to G.W.).

1. Cox DR (1972) Regression models and life-tables. *J R Stat Soc, B* 34(2):187–220.
2. Hall WJ, Wellner J (1981) Mean residual life. *Statistics and Related Topics*, eds Csörgö M, et al. (North-Holland, Amsterdam), pp 169–184.
3. Buckley J, James I (1979) Linear regression with censored data. *Biometrika* 66(3): 429–436.
4. Rubin DB (2004) *Multiple Imputation for Nonresponse in Surveys* (John Wiley & Sons, New York), Vol 81.
5. Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 53(282):457–481.
6. Chen YQ, Cheng S (2006) Linear life expectancy regression with censored data. *Biometrika* 93(2):303–313.
7. Sun L, Zhang Z (2009) A class of transformed mean residual life models with censored survival data. *J Am Stat Assoc* 104(486):803–815.
8. Oakes D, Dasu T (1990) A note on residual life. *Biometrika* 77:409–410.
9. McLain AC, Ghosh SK (2011) Nonparametric estimation of the conditional mean residual life function with censored data. *Lifetime Data Anal* 17(4):514–532.
10. Klein R, Klein BE, Linton KL, De Mets DL (1991) The Beaver Dam eye study: Visual acuity. *Ophthalmology* 98(8):1310–1315.
11. Dabrowska DM (1989) Uniform consistency of the kernel conditional Kaplan-Meier estimate. *Ann Stat* 17(3):1157–1167.
12. Gonzalez-Manteiga W, Cadarso-Suarez C (1994) Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. *Commun Stat Theory Methods* 4(1):65–78.
13. Little RJ, Rubin DB (2014) *Statistical Analysis with Missing Data* (John Wiley & Sons, Hoboken, NJ).
14. Efron B (1967) *The Two Sample Problem with Censored Data* (Prentice-Hall, Englewood Cliffs, NJ), Vol 4, pp 831–853.
15. Kong J, Klein BE, Klein R, Lee KE, Wahba G (2012) Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality. *Proc Natl Acad Sci USA* 109(50):20352–20357.
16. Wahba G (1990) *Spline Models for Observational Data* (SIAM, Philadelphia), Vol 59.
17. Gu C (2013) *Smoothing Spline ANOVA Models* (Springer Science & Business Media, New York), Vol 297.
18. Wang Y (2011) *Smoothing Splines: Methods and Applications* (CRC Press, Boca Raton, FL).
19. Lu F, Keleş S, Wright SJ, Wahba G (2005) Framework for kernel regularization with application to protein clustering. *Proc Natl Acad Sci USA* 102(35):12332–12337.