Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–13

DOI: 10.1177/0962280217693262
journals.sagepub.com/home/smm
SAGE

A doubly robust approach for cost–effectiveness estimation from observational data

Jiaqi Li,[1] Anil Vachani,[2] Andrew Epstein[2] and Nandita Mitra[1]

Abstract

Estimation of common cost–effectiveness measures, including the incremental cost–effectiveness ratio and the net monetary benefit, is complicated by the need to account for informative censoring and inherent skewness of the data. In addition, since the two components of these measures, medical costs and survival are often collected from observational claims data, one must account for potential confounders. We propose a novel doubly robust, unbiased estimator for cost–effectiveness based on propensity scores that allow the incorporation of cost history and time-varying covariates. Further, we use an ensemble machine learning approach to obtain improved predictions from parametric and non-parametric cost and propensity score models. Our simulation studies demonstrate that the proposed doubly robust approach performs well even under mis-specification of either the propensity score model or the outcome model. We apply our approach to a cost–effectiveness analysis of two competing lung cancer surveillance procedures, CT vs. chest X-ray, using SEER-Medicare data.

Keywords

Net monetary benefit, doubly robust, incremental cost effectiveness ratio, machine learning, propensity score

1 Introduction

Policy makers are often interested in the cost–effectiveness (CE) of healthcare interventions in their decision making. Many countries, including the United Kingdom, Australia and Canada, require CE evidence before a drug is granted reimbursement status.[1] However, proper CE analysis is complicated by the need to account for features of cost data, including informative censoring and skewness. In addition, medical costs are often collected from claims data, which are susceptible to confounding. In this paper, we aim to estimate CE measures from observational data accounting for the unique features of cost. Common CE measures include the incremental cost effectiveness ratio (ICER), net monetary benefit (NMB) and CE acceptability curves. These measures require one to estimate cost and effectiveness (e.g. survival time, quality adjusted survival) separately.[2–4]

Historically, cost estimation methods have focused on two unique features of these data: distributional skewness and informative censoring. To account for the skewness and heteroscedasticity often present in cost data, researchers have used ordinary least square regression (OLS), OLS for log cost, generalized linear models (GLM), generalized gamma models, median regression and the Weibull model. Several studies[5–7] have evaluated the performance of these various cost models. Dodd et al.[5] found the generalized gamma model to be the most robust; nevertheless, there is no one-size-fits-all model and Mihaylova et al.[8] concluded that the choice of cost model should depend on the specific structure of the cost data at hand. Another important feature of cost data is excess zeros, especially when analyzing monthly cost data. Two-part models[9,10] have been proposed to accommodate these structural zeros.

[1]Department of Biostatistics & Epidemiology, University of Pennsylvania, Philadelphia, PA, USA
[2]Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Corresponding author:
Jiaqi Li, Department of Biostatistics & Epidemiology, 423 Guardian, Dr Room 503, Philadelphia, PA 19104, USA.
Email: jiaqili@mail.med.upenn.edu

Censoring, which occurs when a study terminates before all patients reach their end-points, is also a common attribute of cost data. Although survival time is usually non-informatively censored (e.g. end-of-study censoring), total cost is informatively censored as patients may have drastically different rates of cost accrual. In practice, a patient with higher costs at the time of censoring is likely to have higher costs at the time of event as well. Lin et al.[11] and Bang and Tsiatis[12] proposed weighted estimators to handle such informative censoring. Several studies have investigated the properties of these methods,[13–16] and some[17–20] have extended the weighting method to linear regression, GLMs and median regression to model the relationship between cost and covariates.

As noted above, effectiveness (typically survival time) is often subject to non-informative censoring. Most CE studies use the area under the Kaplan–Meier (KM) estimate of the survival function to approximate mean survival time.[4,21,22] However, Willan and Briggs[2] suggested using the weighting technique often used for cost estimation to estimate effectiveness instead. This is a less intuitive but more flexible approach; to our knowledge, there is no study comparing the performances of these two approaches.

Current methodological guidance for CE estimation is primarily in the setting of randomized controlled trials.[2–4,22,23] In practice, however, most CE analyses rely on data from observational studies,[24] thus necessitating the need to develop CE estimation models for observational data. Previous studies Anstrom and Tsiatis[25] and Li et al.[26] have investigated the propensity score (PS)-based models for observational cost data only. Mitra and Indurkhya,[27] Indurkhya et al.[28] and Goldfeld[29] proposed PS adjustment for estimating the NMB from claims data.

The goal of this study is to develop novel doubly robust (DR) estimators based on PSs to estimate CE measures from observational data. We build on our previous work on cost estimation[26] and here developed DR estimators for CE, which is of great interest to policy makers and health economists. DR estimation combines outcome regression with weighting by PSs and is robust to mis-specification of one (but not both) of these two components. DR has also been shown to perform better than other PS-based methods such as stratification and weighting.[30] Thus, our proposed estimator has the advantage of being robust against mis-specification of PS and regression models over the aforementioned traditional CE estimation approaches. In addition, our proposed estimators allow the incorporation of cost history and can handle periodic cost data such as monthly or quarterly payment claims data. We have also extended our estimators to work with time-varying covariates. CE studies, especially those comparing cancer therapies, often aim to capture long-term cost and survival effects. During this period, patient characteristics such as health status and comorbidities may vary over time thus necessitating the need for CE models appropriate for time-varying covariates. Furthermore, given the heterogeneous nature of cost distributions, the many possible choices of cost models described above, as well as the challenge of accurately estimating PSs, we propose applying an ensemble machine learning approach based on cross-validation[31] to best estimate the outcome and the PS components in DR.

We begin with a background introduction to common CE measures. We follow by comparing and extending two effectiveness estimation methods and then propose a DR effectiveness estimator in Section 3.1. We then introduce two DR cost estimators the simple weighted and the partitioned model in Section 3.2. We present results from extensive simulation studies that compare the performance of the proposed DR estimators in Section 4. Finally, we apply these DR estimators to a CE analysis of two lung cancer surveillance procedures, CT scans vs. chest X-ray, using SEER-Medicare data.

2 Background: Common CE measures

CE analysis is often used to evaluate the merits of a new healthcare intervention (treatment, $Z=1$) compared to an existing one (control, $Z=0$). CE measures integrate estimates of costs and effectiveness in a single statistic derived from two components: Δ_E and Δ_C, where $\Delta_E = \text{Effectiveness}_{Z=1} - \text{Effectiveness}_{Z=0}$ and $\Delta_C = \text{Cost}_{Z=1} - \text{Cost}_{Z=0}$. The duration of interest can be considered to be $(0, \tau)$ so that cost and effectiveness measures are bounded by τ and cost refers to the total cost incurred from time 0 to τ. In this section, we briefly introduce three of the most commonly used CE measures.

2.1 ICER

ICER is an intuitive statistic that is defined as $\text{ICER} = \frac{\Delta_C}{\Delta_E}$. A major limitation of the ICER is its discontinuity when the denominator Δ_E approaches zero. In addition, estimating the variance of ICER is problematic due to the acknowledged statistical problems associated with ratio statistics. Non-parametric bootstrapping, Fieller's theorem and Bayesian approaches[32–34] can be applied to estimate the variance of ICER.

2.2 CE acceptability curve

An important concept in CE analysis is called willingness to pay (WTP, denoted by λ), which is the maximal monetary value decision makers are willing to pay for a unit of Δ_E. Typically, λ measures the dollar amount one is willing to pay for one year of additional life. A CE acceptability curve displays the probability that the treatment is cost-effective compared with the control for a range of λ values. To plot the CE acceptability curve, we use bootstrapping to estimate $Pr(\lambda\Delta_E - \Delta_C > 0)$. In practice, we simply count the proportion of bootstrapped samples that yields $\lambda\Delta_E - \Delta_C > 0$ for a range of λ values.

2.3 NMB

Recently, health economists have advocated the use of the NMB: $\text{NMB}(\lambda) = \lambda\Delta_E - \Delta_C$. NMB is a linear combination of Δ_C and Δ_E; it measures the excess benefit given a fixed level of λ. The NMB does not suffer from the singularity problem that the ICER does and it is straight forward to estimate its variance as $var(\text{NMB}(\lambda)) = \lambda^2 var(\Delta_E) + var(\Delta_C) - 2\lambda cov(\Delta_E, \Delta_C)$.

3 Methods: DR CE estimation

3.1 Δ_E estimation

In CE studies, effectiveness usually refers to survival time or quality adjusted life years. From here onwards, we will simply use survival time to represent effectiveness. Specifically, we are interested in estimating the *mean* survival time difference Δ_E between two treatment groups in the duration of interest $(0, \tau)$ in the presence of censoring.

3.1.1 Comparison and extension of current techniques

Consider a randomized controlled trial, where t_i and C_i represent the survival and censoring time for subject i, respectively, $T_i = \min(t_i, C_i, \tau)$ and censoring indicator $\delta_i = I(T_i \le C_i) + I(T_i > C_i) \times I(C_i \ge \tau)$. We start by reviewing two popular estimation techniques, the area under the survival curve and inverse probability weighting to estimate mean time difference $\Delta_E = E(T|Z = 1 - T|Z = 0)$ in the duration of interest $(0, \tau)$. Here, our goal is to accurately estimate mean survival time under censoring.

In the first approach, we integrate the area under the survival curve such as the KM curve. Let $S(t)$ be the survival function, by definition

$$\Delta_E = \int_0^\tau S_{Z=1}(t)\mathrm{d}t - \int_0^\tau S_{Z=0}(t)\mathrm{d}t \tag{1}$$

In practice, we integrate the area under the estimated survival curve: $\hat{\Delta}_E = \sum_{i=1}^{\tau} \hat{S}_{Z=1}(t_i) \times (t_{i+1} - t_i) - \sum_{j=1}^{\tau} \hat{S}_{Z=0}(t_j) \times (t_{j+1} - t_j)$. The area under the survival curve approach is easy to use and hence very popular in CE studies. It does require the non-informative censoring assumption and does not accommodation adjustment of confounders. Common regression-based survival models such as the Cox proportional hazard model focus on hazard ratio estimation, making mean survival time estimation difficult. In addition, to our knowledge, there is no DR method available for survival models.

An alternative way to handle censoring in mean survival time estimation is to utilize inverse probability weighting.[2] This weighting technique is often used for cost estimation[12] and is an application of the general representation theorem for missing data.[35,36] Here we apply the same concept to estimate mean survival time as follows

$$\Delta_E = \frac{1}{n_1}\sum_{i=1}^{n_1} \frac{T_i\delta_i Z_i}{K_{Z=1}(T_i)} - \frac{1}{n_0}\sum_{j=1}^{n_0} \frac{T_j\delta_j(1 - Z_i)}{K_{Z=0}(T_j)} \tag{2}$$

where $K_{Z=z}(u) = P(C \ge u|Z = z)$ and n_z is the total number of subjects in treatment group Z. This estimator is simply a weighted average of observed survival times T_i for patients who are not censored. The weight is given by the inverse of the probability of not being censored at the time of death for those who died prior to τ and the

inverse of the probability of not being censored at τ for those who survived to τ. We can easily show that $\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{T_i\delta_i Z}{K_{Z=1}(T_i)}$ is an unbiased estimator of $E(T|Z=1)$ as follows:

$$\begin{aligned}
E\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{T_i\delta_i Z}{K_{Z=1}(T_i)}\right] &= E\left[\frac{1}{n_1}\sum_i E\left[\frac{T_i\delta_i}{K_{Z=1}(T_i)}\middle| T_i, Z=1\right]\right] \\
&= E\left[\frac{1}{n_1}\sum_i\left[\frac{T_i|Z=1}{K_{Z=1}(T_i)}E(I(C_i \geq T_i|T_i, Z=1)\right]\right] \\
&= E\left[\frac{1}{n_1}\sum_i T_i|Z=1\right] = E[T|Z=1]
\end{aligned}$$

In practice, we use KM to estimate $K(u)$ based on the data $(T_i, 1-\delta_i)$ so $\hat{\Delta}_E = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{T_i\delta_i Z_i}{\hat{K}_{Z=1}(T_i)} - \frac{1}{n_0}\sum_{j=1}^{n_0}\frac{T_j\delta_j(1-Z_i)}{\hat{K}_{Z=0}(T_j)}$.

Hence, we have shown that the inverse probability weighting technique provides an unbiased estimate of mean survival time and thus Δ_E. Similarly, this approach assumes censoring to be non-informative. In other words, censoring in time is independent of other covariates. This assumption is considered to be valid for most observational studies,[15,25,29] especially in large population-based registries where censoring is administrative due to end of study. Nevertheless, if censoring is informative, in other words, censoring is dependent on covariates $\boldsymbol{X}$ in the case of induced dropout or non-compliance, we propose to modify equation (2) by using $P(\delta_i = 0|\mathbf{X}_i, Z)$ instead of $K_z(T_i)$.[37,38] Specifically, we first divide $(0, \tau)$ into discrete time intervals. At each time point j, we estimate $P(\delta_i^j = 0|\mathbf{X}_i, Z, \delta_i^{j-1} = 0)$ using a logistic regression model $\delta_i^j \sim \mathbf{X}_i$ for all subjects with $Z = z$ alive at time j. We can then use $P(\delta_i = 0|\mathbf{X}_i, Z) = \prod_{j<T_i} P(\delta_i^j = 0|\mathbf{X}_i, Z, \delta_i^{j-1} = 0)$ instead of $K(T_i)$ in equation (2). Similarly, this extension can incorporate time-varying covariates $\boldsymbol{X}$. When $\mathbf{X}_j$ takes on different values at different time points j, as often seen in long-term cancer studies, we can estimate the probability of not being censored at each time point $P(\delta_i = 0|\mathbf{X}_{ij}, Z)\prod_{j<T_i} P(\delta_i^k = 0|\mathbf{X}_{ij}, Z, \delta_i^{j-1} = 0)$. In 3.2.2, we discuss how our proposed Δ_C estimators can accommodate time-varying covariates.

Although both techniques are unbiased, the weighting technique has some advantages. Specifically, it allows for covariate adjustment, accommodates informative censoring, and, as we will discuss next, is a natural fit for DR estimation. In section 4.1, we carry out simulation studies to compare the performance of these two approaches. Of note, these two methods for Δ_E estimation have been widely used but their empirical performance has not been compared head-to-head until now.

3.1.2 DR method for Δ_E estimation

As CE studies are often based on observational data, we use the conventional counterfactual notation modeling a causal framework. Let $t_i^{(0)}$ and $t_i^{(1)}$ denote the survival time if the patients were in the control and treatment group respectively. Let $\mathbf{X}_i$ be a vector of measured confounders we wish to adjust for. Lastly, PSs are denoted by $e = e(\mathbf{X})$. We assume strong ignorability and non-informative censoring (see Li et al.[26] for further explanation of the assumptions).

We propose the following DR estimator for Δ_E that uses the concept of inverse probability weighting

$$\hat{\Delta}_E = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i T_i\delta_i}{\hat{e}_i\hat{K}(T_i)} - \frac{(Z_i-\hat{e}_i)m_1(\mathbf{X}_i)\delta_i}{\hat{e}_i\hat{K}(T_i)}\right] - \left[\frac{(1-Z_i)T_i\delta_i}{(1-\hat{e}_i)\hat{K}(T_i)} + \frac{(Z_i-\hat{e}_i)m_0(\mathbf{X}_i)\delta_i}{(1-\hat{e}_i)\hat{K}(T_i)}\right] \tag{3}$$

For simplicity, we use $\hat{K}(u)$ to denote the treatment-specific estimated probability of being uncensored at u, $\hat{K}_z(u)$. Moreover, $m_0(\mathbf{X}_i)$ and $m_1(\mathbf{X}_i)$ are the postulated models for the true regressions $E(T|Z=0, \mathbf{X})$ and $E(T|Z=1, \mathbf{X})$. The outcomes models $m_1(\mathbf{X})$ and $m_0(\mathbf{X})$ can be specified in various ways including:

- Normal model: $E(T_i|Z_i = z, X_i) = \mathbf{X}_i\beta$ weighted by $\frac{\delta_i}{\hat{K}(T_i)}$
- Lognormal model: $E(\log(T_i)|Z_i = z, X_i) = \mathbf{X}_i\beta$ weighted by $\frac{\delta_i}{\hat{K}(T_i)}$
- Gamma model: $E(T_i|Z_i = z, X_i) = \exp(\mathbf{X}_i\beta)$ weighted by $\frac{\delta_i}{\hat{K}(T_i)}$

The proposed DR estimator is consistent if the PS model e or the outcome models $m_1(\mathbf{X}) = E(T|Z=1, \mathbf{X})$ and $m_0(\mathbf{X}) = E(T|Z=0, \mathbf{X})$ are correctly specified. Note that weights $\frac{\boldsymbol{\delta}}{\hat{K}(\mathbf{T})}$ are applied to both PS weighting and

outcome models to account for censoring. The variance of $\hat{\Delta}_E$ can be estimated using large sample theory to get the sandwich variance estimator[26] or non-parametric bootstrapping.

Funk et al.[39] noted that the DR property can lead to biased estimates if both the outcome and the PS model are mis-specified, where in reality the true PS and outcome models are never known. In order to best estimate PS and outcome models in DR estimation, we employ a machine learning algorithm, Super Learner (SL).[31] SL is an ensemble learning approach based on V-fold cross-validation that allows one to specify several candidate prediction models and then use them to produce an asymptotically optimal combination. Specifically, each of the candidate algorithms is fitted on the training set and outcomes are predicted using the validation set. Averaging across all validation sets, we calculate the estimated cross-validated risk score for each algorithm. The SL algorithm finds the optimal weighted combination of all the models, which is shown to be asymptoticly efficient and is guaranteed to perform at least as well as the best estimators from the candidate models. SL is available as an R package SL and as an SAS macro.

Recent work suggests using non-parametric machine learning models such as Classification and Regression Trees (CART), random forests, neural networks and boosting[40,41] for PS estimation. An R package TWANG is available to estimate PS using boosting. Thus, we can use SL to estimate PS from models of different functional forms as well as the aforementioned non-parametric PS estimation algorithms. Since survival time distributions can be very different for different diseases, we can also utilize SL to combine prediction from several possible survival time models to estimate $m_1(\mathbf{X})$ and $m_0(\mathbf{X})$.

3.2 Δ_C estimation

Let $Y_i(u)$ be the known accumulated cost up to time u and Y_i be the total cost that subject i accrues up to τ. Hence, total cost $Y_i = Y_i(t_i)$ is not observed if $\delta_i = 0$, when a subject is censored before τ. In other words, we only observe Y_i for uncensored subjects. For censored subjects, their cost will continue to accrue hence their total cost Y_i is unknown. In this section, we introduce two DR estimators, the simple weighted and the partitioned based on Bang and Tsiatis.[12] The simple weighted estimator is appropriate when cost history information $Y(u), u < T_i$ is not available and we only know total cost $Y_i(T_i)$. The partitioned estimator is appropriate when we have access to cost history information, for example, periodic insurance claims or monthly Medicare payment information.

3.2.1 DR for Δ_C estimation – Simple weighted

When cost history data are not available, we propose the simple weighted DR estimator for cost estimation. This estimator is very similar to the Δ_E estimator discussed in section 3.1.2.

$$\hat{\Delta}_C = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i Y_i \delta_i}{\hat{e}_i \hat{K}(T_i)} - \frac{(Z_i - \hat{e}_i) m_1(\mathbf{X}_i)\delta_i}{\hat{e}_i \hat{K}(T_i)}\right] - \left[\frac{(1 - Z_i) Y_i \delta_i}{(1 - \hat{e}_i)\hat{K}(T_i)} + \frac{(Z_i - \hat{e}_i) m_0(\mathbf{X}_i)\delta_i}{(1 - \hat{e}_i)\hat{K}(T_i)}\right] \tag{4}$$

Thompson and Nixon[42] suggested that conclusions from CE analyses are sensitive to choice of the cost distribution. The true cost distribution is often unknown and can vary greatly across diseases and medical procedures. Different cost models such as the normal, GLM, lognormal and generalized gamma models can be considered as candidate outcome models for m_0 and m_1. We can then utilize SL to incorporate all possible outcome models, since SL is guaranteed to perform at least as well as the best fitted model among all candidates. A proof of the DR property of the simple weighted estimator can be found in Li et al.[26]

3.2.2 DR methods for Δ_C estimation – Partitioned

When cost history information is available, we propose a partitioned DR estimator. This estimator is based on the partitioned total cost estimation method.[12] Specifically, the duration of interest $(0, \tau)$ is partitioned into L subintervals $(t_j, t_{j+1}]$, $j = 1, 2, \ldots, L$, $0 = t_1 < t_2 < \cdots < t_{L+1} = \tau$. Intuitively, we estimate the cost difference within each interval and "sum up" contributions from all L intervals.

$$\hat{\Delta}_C = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{L}\left[\frac{Z_i Y_{ij} \delta_i^j}{\hat{e}_i \hat{K}_j(T_i^j)} - \frac{(Z_i - \hat{e}_i) m_1^j(\mathbf{X}_i)\delta_i^j}{\hat{e}_i \hat{K}_j(T_i^j)}\right] - \left[\frac{(1 - Z_i) Y_{ij} \delta_i^j}{(1 - \hat{e}_i)\hat{K}_j(T_i^j)} - \frac{(Z_i - \hat{e}_i) m_0^j(\mathbf{X}_i)\delta_i^j}{(1 - \hat{e}_i)\hat{K}_j(T_i^j)}\right] \tag{5}$$

where $Y_{ij} = Y_i(t_j) - Y_i(t_{j-1})$ is subject i's cost accrued in the interval $(t_{j-1}, t_j]$. $T_i^j = \min(T_i^{t_j}, C_i)$ and $\delta_i^j = I(\min(T_i, t_j) \le C_i)$, the censoring indicator for subject i at time t_j. $\hat{K}_j(T_i^j)$ are the KM survival estimates of

$K_j(T_i^j)$ based on the data $(\min(T_i^{t_j}, C_i), 1-\delta_i^j)$. $m_0^j(\mathbf{X}_i)$ and $m_1^j(\mathbf{X}_i)$ are the postulated models for the true regressions $E(Y_{ij}|Z_i=0,\mathbf{X}_i)$ and $E(Y_{ij}|Z_i=1,\mathbf{X}_i)$.

Similarly, we use SL to estimate the postulated outcome models and PS model. For outcome models, we estimate within each interval. For example, the normal estimating equation for the postulated outcome models is $\sum_{i=1}^n \frac{\delta_i^j}{\hat{K}_j(T_i^j)}(Y_{ij}-\beta_k' X_i)X_i = 0$ and thus $\hat{\beta}_k = \left\{\sum_{i=1}^n \frac{\delta_i^j}{\hat{K}_j(T_i^j)} X_i' X_i\right\}^{-1} \sum_{i=1}^n \frac{\delta_i^j}{\hat{K}_j(T_i^j)} Y_{ij} X_i$. See Appendix 1 for a proof of the DR property. Variance of $\hat{\Delta}_C$ can be estimated using large sample theory to obtain the sandwich variance estimator or via non-parametric bootstrapping.

In periodic cost data such as monthly claims data, it is common to observe subjects with zero costs in specific periods. Thus, in addition to the outcome models mentioned in section 3.1.2, we propose the two-part model first introduced by Duan et al.[10] to account for structural zeros:

Part 1: a logit/probit model to model the probability of zero cost: $I(Y_{ij}=0) \sim X_i\beta$.
Part 2: a normal/lognormal/gamma model for positive costs: $f(Y_{ij}|Y_{ij}>0) \sim X_i\beta$.

The partitioned Δ_C estimator allows for a natural extension for handling time-varying covariates and time-varying treatments. Let Z_{ij} and $\mathbf{X}_{ij}$ be the treatment indicator and the confounding covariates for subject i at time interval j, respectively. We can substitute Z_i and $\mathbf{X}_i$ in equation (5) with Z_{ij} and $\mathbf{X}_{ij}$. Note that here we may need to partition time differently so that the L time intervals are fine enough to capture all covariate, treatment and cost changes. In other words, at most one change in $Z_i, \mathbf{X}_i, Y_i$ occurs in each interval $(t_{j-1}, t_j]$. The PS e_{ij} needs to be estimated at each interval as well to accommodate time-varying treatments and covariates.

3.3 CE estimation

After we have estimated Δ_E and Δ_C according to the methods presented above, we can combine them to estimate common CE measures

$$\widehat{ICER} = \frac{\hat{\Delta}_C}{\hat{\Delta}_E} \tag{6}$$

and

$$\widehat{NMB}(\lambda) = \lambda\hat{\Delta}_E - \hat{\Delta}_C \tag{7}$$

For the CE acceptability curve, we can use bootstrapping to count the proportion of iterations with $\lambda\hat{\Delta}_E - \hat{\Delta}_C > 0$ for a range of λ values. The variance of NMB can be estimated directly as mentioned in Section 2, although bootstrapping[43] is preferred due to the complexity of the DR models.

4 Simulation studies

Using simulation studies, we first compare the two effectiveness estimation methods: inverse probability weighting vs. area under the survival curve. Then, we evaluate the performance of the DR models for Δ_E, Δ_C and NMB discussed in Section 3 under various settings. We chose to focus on NMB over other CE measures because of its attractive statistical properties and popularity in modern CE analyses. We report the percentage bias (bias), the coverage probability (cvrg) of the resulting 95% confidence interval and the empirical standard error (SE).

4.1 Δ_E estimation: Area under the survival curve vs. inverse probability weighting

In section 3.1.1, we reviewed these two popular techniques for mean time estimation. To our knowledge, there are no studies in the literature comparing their empirical performance. In this study, exponential (mean = 6), Weibull (shape = 2, scale = 6) and uniform ([0,10]) survival times were simulated. Two levels of censorship were introduced with censoring time being uniformly distributed on [0,20] and [0,12.5], corresponding to 20% and 30% censoring, respectively. The average 10-year mean survival time $E(T), \tau = 10$ serves as the parameter of interest. Sample size n, was chosen to be 100 and 1000. A total of 500 simulations were conducted and confidence intervals were

Table 1. $E(T)$ simulation results: Inverse probability weighting vs. area under the survival curve.

			Inverse probability weighting			Area under the survival curve		
n	Censoring	Survival time	bias	SE	cvrg	bias	SE	cvrg
100	Light	Uniform	−0.006	0.309	0.952	−1.243	0.311	0.950
		Weibull	−0.297	0.280	0.944	−3.360	0.313	0.896
		Exponential	−0.327	0.371	0.964	−1.230	0.371	0.960
	Heavy	Uniform	−0.673	0.349	0.940	−1.337	0.331	0.942
		Weibull	−0.009	0.336	0.940	−1.758	0.326	0.926
		Exponential	−0.218	0.438	0.962	−0.891	0.395	0.958
1000	Light	Uniform	0.002	0.098	0.952	−0.116	0.098	0.948
		Weibull	−0.038	0.089	0.958	−0.408	0.091	0.954
		Exponential	0.037	0.118	0.942	−0.052	0.118	0.938
	Heavy	Uniform	−0.013	0.105	0.954	−0.114	0.105	0.950
		Weibull	0.163	0.097	0.946	−0.063	0.098	0.942
		Exponential	−0.046	0.125	0.926	−0.138	0.125	0.918

constructed using non-parametric bootstrapping with the BCa correction. The simulation study settings were chosen according to those used by Bang and Tsiatis.[12]

As expected, mean survival time $E(T)$ estimation using inverse probability weighting and area under the survival curve both yielded very small bias (Table 1) and comparable coverage. Note that for subjects with large observation time, if the estimated probability of censoring $\hat{K}(T_i)$ was zero, then $\min \hat{K}(T_i)$ was used instead to avoid the denominator being zero. Similarly, for the area under the survival curve, the estimated probability of the largest censored observation was underestimated. Thus, we see some downward bias for the empirical estimation of mean survival time. The results from Table 1 show that inverse probability weighting has comparable, if not slightly better empirical performance compared to the area under the survival curve method, especially with smaller sample sizes.

4.2 CE estimation

We designed our simulation studies to mimic CE analyses from observational data. We first simulated three covariates $\mathbf{X} = (X_1, X_2, X_3)$, where X_1 was binary with success probability 0.5. X_2 and X_3 were normally distributed with means of 2 and 1, respectively and common standard error of 1. Using these covariates, treatment choice Z was defined using a logit index model with $D \sim$ Bernoulli(p) and $\text{logit}(p) = 0.5 + X_1 + 0.25X_2 + 0.5X_3 - 0.5X_1X_2 - 0.25X_2X_3 - 0.5X_3^2$. The sample size was set to be 1000. The mean average five-year Δ_E, Δ_C and NMB were chosen as parameters of interest.

We drew failure times from Weibull and exponential distributions. For exponential failure times, we set the mean to be $1/\exp(0.25 - Z + 0.5X_1 - 0.25X_2 - 0.25X_3)$. For Weibull, the shape parameter was 2 and the scale parameter was $\exp(-0.15 + Z - 0.5X_1 + 0.25X_2 + 0.25X_3)$. Censoring times were independently drawn from $U(0, 10)$ and $U(0, 6)$ for light and heavy censoring. The rate of censoring was approximately 20% for light censoring and 40% for heavy censoring.

For each subject, costs were generated for each month until the end of the study period τ. The total costs were generated from three different components: an initial diagnostic cost, an ongoing monthly cost and a cost accrued at the time of death if the subject's death was observed before τ. Four different cost distributions: normal, gamma, mixed and excess zeroes were generated as demonstrated in Table 2. For the excess zero cost model, proportion p_0 ($\text{logit}(p0) = -Z - 2X_1 - .5X_2 + .5X_3$) of the patients were assumed to experience zero cost each month.

We first estimated Δ_E and Δ_C with a commonly used in practice approach that we refer to as a "conventional" approach, in which Δ_C was estimated from a linear regression and Δ_E was derived using area under the survival curve. We then estimated Δ_C and Δ_E using inverse probability of treatment weighting (IPTW) based on PSs with censoring correction $\frac{\delta}{\hat{K}(T)}$.[26] PS models were either correctly specified or mis-specified. In the correctly specified case, covariates and their correct functional forms were included in the logistic regression model; in the mis-specified case, only the main effects X_1, X_2, X_3 were included. Then, our proposed DR model for Δ_E and our two DR models, simple weighted and partitioned for Δ_C were applied. PSs were estimated utilizing the SL algorithm. Candidate PS models included a logistic regression model, a logistic regression model with all interactions and three non-parametric algorithms: generalized

Table 2. Simulation setup: Four cost distributions.

	Initial cost	Ongoing cost	Dying cost
Normal	$N(\mu = 10 + 10Z + 5X_1 + 5X_2 + 5X_3, 5)$	$N(.1\mu, .5)$	$N(0.2\mu, .1)$
Gamma	$\text{Gamma}(\alpha = 2.5, \beta = \exp(-Z - .5X_1 - .2X_2 - .2X_3))$	$\text{Gamma}(\alpha, .1\beta)$	$\text{Gamma}(\alpha, .2\beta)$
Mixed	a 50/50 mixture of normal and gamma		
Excess zero	$N(\mu, 5)$	$P(p_0) = 0$, $P(1 - p_0) = N(.1\mu, .5)$	$N(0.2\mu, .1)$

Table 3. Simulation results: Cost, effectiveness and NMB estimation.

Effectiveness Δ_E

		Conventional			IPTW			DR		
PS	Time dis	bias	SE	cvrg	bias	SE	cvrg	bias	SE	cvrg
Correct	Weibull	−21.061	0.09	0.028	−0.319	0.578	0.99	−0.065	0.082	0.952
	Exp	−21.574	0.104	0.198	−1.024	0.332	0.964	−0.209	0.107	0.948
Mis	Weibull	−20.901	0.097	0.026	−13.192	0.1	0.372	−0.065	0.082	0.952
	Exp	−21.09	0.112	0.304	−15.026	0.124	0.634	−0.209	0.107	0.948

Cost Δ_C

			Conventional			IPTW			DR − simple			DR − partitioned		
PS	Time dis	Cost dis	bias	SE	cvrg	bias	SE	cvrg	bias	SE	cvrg	bias	SE	cvrg
Correct	Weibull	Normal	−38.134	4.545	0	0.404	44.871	0.99	0.268	4.021	0.938	−0.131	4.025	0.95
		Mixed	−34.142	3.779	0	−1.845	18.147	0.964	−0.121	3.433	0.958	0.064	3.428	0.95
		Gamma	−29.046	3.386	0	0.912	33.183	0.988	−0.354	3.132	0.956	0.349	3.207	0.958
		Zero	−32.713	4.234	0	−1.702	25.871	0.968	−0.036	3.809	0.95	−0.092	3.803	0.938
Mis		Normal	−38.134	4.545	0	−15.192	5.771	0.204	0.268	4.021	0.938	−0.131	4.025	0.95
		Mixed	−34.142	3.779	0	−14.071	4.623	0.198	−0.121	3.433	0.958	0.064	3.428	0.95
		Gamma	−29.046	3.386	0	−13.145	3.966	0.25	−0.354	3.132	0.956	0.349	3.207	0.958
		Zero	−32.713	4.234	0	−13.93	4.988	0.182	−0.036	3.809	0.95	−0.092	3.803	0.938

NMB

			Δ_E, Δ_C : Conventional			Δ_E, Δ_C : I PTW			Δ_E:DR,Δ_C: simple DR			Δ_E:DR,Δ_C: partitioned DR		
PS	Time dis	Cost dis	bias	SE	cvrg	bias	SE	cvrg	bias	SE	cvrg	bias	SE	cvrg
Correct	Weibull	Normal	−119.442	3.037	0.000	3.269	16.402	0.988	1.282	1.559	0.950	−1.007	1.550	0.946
		Mixed	−278.966	2.749	0.000	−2.363	3.279	0.970	−2.893	1.328	0.952	0.603	1.431	0.954
		Gamma	54.670	2.906	0.624	1.718	6.026	0.988	4.718	1.693	0.952	−1.671	1.659	0.962
		Zero	−107.682	3.300	0.008	−8.007	7.858	0.964	−0.232	2.123	0.948	−0.639	2.109	0.964
Mis		Normal	−119.442	3.037	0.000	−22.820	1.980	0.432	1.282	1.559	0.950	−1.007	1.550	0.946
		Mixed	−278.966	2.749	0.000	−32.122	1.469	0.804	−2.893	1.328	0.952	0.603	1.431	0.954
		Gamma	54.670	2.906	0.624	−11.882	1.693	0.900	4.718	1.693	0.952	−1.671	1.659	0.962
		Zero	−107.682	3.300	0.008	−18.326	2.401	0.804	−0.232	2.123	0.948	−0.639	2.109	0.964

PS: propensity score; IPTW: inverse probability of treatment weighting; DR: doubly robust; NMB: net monetary benefit.

additive model (GAM), k nearest network (KNN) and boosting. Next, the outcome parameters in all DR models were also estimated using the SL algorithm with the following candidate model:linear regression, GLM with log link, generalized gamma model with log link and gamma variance. In addition, a two-part model consisting of a logistic regression model for structural zeros and a linear model for the non-zero costs was included among the candidates. For simplicity, we used the default tuning parameters for all machine learning algorithms and SL. We estimated NMB with a standard willingness to pay of $\lambda = 50,000$/yr.

The results from Table 3 show that the conventional area under the survival curve method yields biased estimates because it failed to account for confounders. IPTW method was unbiased, but produced biased

estimates (−13.2% to 15.0%) when the PS model was mis-specified. In addition, even when the PS model was correct, IPTW had large standard errors and inflated coverage (0.96–0.99). The proposed DR method worked well and had very small bias (−0.23% to −0.05%), small standard error and good coverage. DR model was also robust to PS model mis-specification, since PSs were estimated using several parametric and non-parametric algorithms.

For Δ_C estimation, the results show that conventional linear regression produced biased estimates under all scenarios. IPTW yielded unbiased estimates with large standard errors but failed when the PS model was mis-specified. Both simple and partitioned DR had negligible biases ranging from −0.036% to 0.349%. Both had standard errors and their coverage probabilities were around 95%. Note that even for the excessive zero cost structure, the two DR methods had similar performance. This result again confirms the DR property of the proposed estimators, since the PSs were estimated correctly using SL.

For NMB estimation, the conventional method produced very biased estimates (−278% to 55%). Estimates from IPTW exhibited the same problem with large standard errors, inflated coverage and biased results when the PS model was mis-specified. Our proposed DR approaches had superior performance compared to all other approaches.

5 Lung cancer surveillance data

Lung cancer is responsible for the largest number of cancer-related deaths worldwide.[44] In addition, the overall economic burden of lung cancer on society is large and growing.[45] Patients who undergo curative resection for lung cancer are at risk of developing a recurrence or a new primary lung cancer in the future. Therefore, imaging surveillance has become standard of care after lung surgery. Currently, the two most common approaches to surveillance are use of chest X-ray or chest CT. The optimal surveillance strategy is unknown; there are no randomized trials that have directly compared the effect of imaging strategy (CT vs. X-ray) on overall survival following lung cancer resection. Of note, surveillance is different from screening; screening is the use of X-ray or CT for detection of lung cancer in patients at risk, while surveillance is the use of X-ray or CT to look for cancer recurrence in patients who have already received surgery for their lung cancer. Although the CE of lung cancer screening[46] has been extensively studied, there have been no trials comparing X-ray to CT in the surveillance setting. Therefore, we turn to observational data and apply our CE estimators to compare the three-year CE of chest X-ray vs. CT using a cohort of patients derived from the SEER-Medicare registry.

We included stage I–IIIA non-small cell lung cancer patients diagnosed between 2007 and 2009 and treated with curative intent surgery. See Ciunci et al.[47] for a detailed description of inclusion/exclusion criterion. At the end of the study, 59.1 percent of the study cohort was censored. Payment data were extracted from Medicare claims from the inpatient MEDPAR, outpatient SAF and non-institutional Carrier files covering 2007 through 2010. For each patient, we calculated total spending as the sum of payments made to the provider by Medicare, the patient, and other payers. Payments were calculated in consecutive 30-day periods starting 181 days after the surgery index date and lasting until patient death or censoring (31 December 2010). We did not adjust for inflation due to the short time span covered in this study (2007 to 2010). The final cohort sample size was 3389; 1058 of whom had chest X-ray and 2331 had CT surveillance. Three-year total cost was highly right skewed, with a maximum observed cost of $722,100. The average observation duration was 22 months. Figure 1 shows the KM survival plot of CT and X-ray. Patients on CT had higher survival probability and comparable cost compared to patients on X-ray.

In this study, both the choice of surveillance strategy and three-year cost may have been influenced by demographic variables including age, sex, median income, marital status, as well as health and disease status variables including Charlson score, stage, surgery type, histology, chemotherapy and radiation. For instance, younger patients with less comorbidities as measured by Charlson comorbidity index are more likely to receive CT and these factors are also more likely to be associated with survival and cost. We first estimated CE measures, including ICER and NMB, using the "conventional" method, where Δ_C is estimated from a linear model and Δ_E is derived using area under the survival curve. We then estimated Δ_C and Δ_E using IPTW based on PSs. Lastly, DR models proposed in section 3.1.2 and section 3.2.1 were applied. PSs were estimated utilizing the SL algorithm. Candidate PS algorithms included a logistic regression model, a logistic regression model with all interactions, boosting, GAM, and KNN. Again, we used the default tuning parameters for all machine learning algorithms. Similarly, the regression parameters in the DR model were estimated using the SL algorithm in conjunction with linear regression, GLM with log link and generalized gamma model with log link and gamma variance. Approximate confidence intervals for ICER and NMB (WTP = $50,000/yr) were constructed using non-parametric bootstrapping with BCa correction.

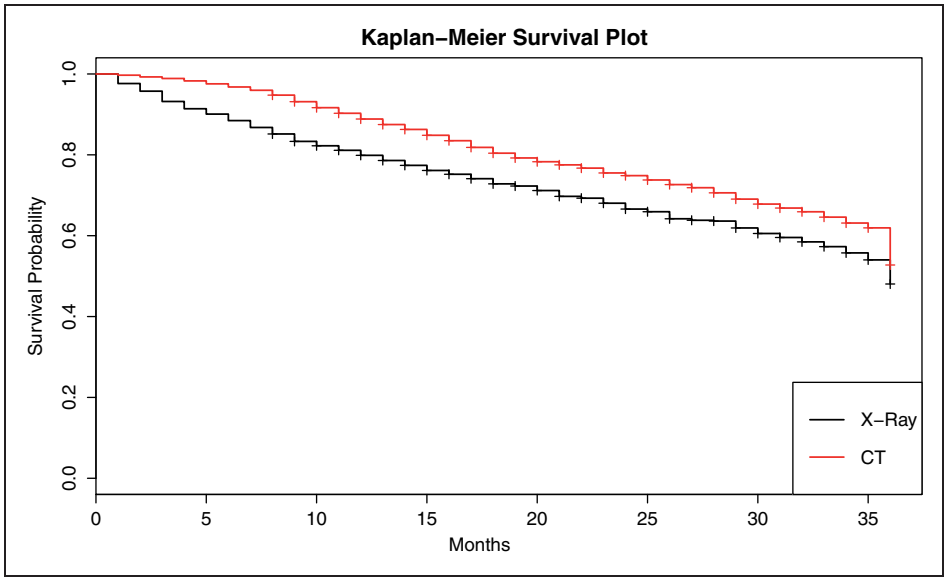

Figure 1. Kaplan–Meier survival plot.

Table 4. CE analysis of CT vs. X-ray (reference) for lung cancer surveillance.

			NMB (WTP=50,000)	
Method	$\Delta_E(mths)$	$\Delta_C(\$)$	Estimate	95% CI
Conventional	3.14	410	156,482	5939, 19,359
IPTW	3.12	−2539	158,390	7330, 23,455
DR	3.65	−3512	185,990	11,490, 27,472

IPTW: inverse probability of treatment weighting; DR: doubly robust; WTP: willingness to pay; NMB: net monetary benefit.

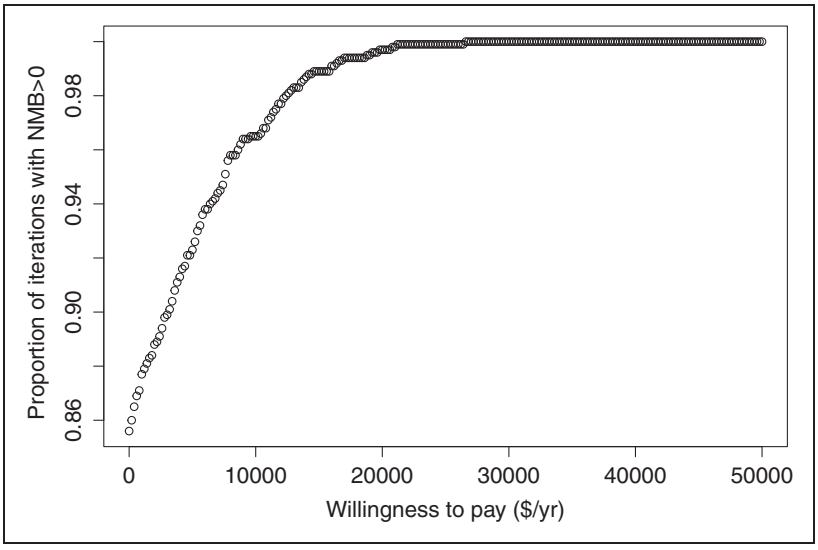

Figure 2. Cost-effectiveness acceptability curve of CT vs. X-ray (reference) for lung cancer surveillance.

As shown in Table 4, patients on CT were estimated to live on average 3.12 to 3.65 months longer than patients on X-ray. This result is consistent with Ciunci et al.[47] which demonstrated CT is associated with lower hazard of death. Δ_C was vastly different between the commonly used "conventional" method and IPTW or DR. Although all three approaches suggest that CT is significantly more cost-effective than X-ray, the DR approach produced a much higher NMB estimate, indicating that CT is notably more cost-effective. In addition, DR yielded tighter a 95% confidence interval than IPTW.

In Figure 2, we plotted the CE acceptability curve from bootstrapped samples under a wide range of WTP values. This figure provides a visual demonstration of when CT becomes significantly more cost-effective compared to X-ray. We see that around $\lambda = \$8000/\text{yr}$, over 95% of the bootstrap iterations yield positive NMB. In other words, CT was significantly more cost-effective compared to X-ray with a WTP of more than \$8000/yr.

6 Summary

In policy making and health services evaluation where an emphasis is placed on estimating not only the effectiveness but the CE of interventions, it is imperative to estimate CE measures accurately and robustly. We propose DR estimators based on PSs to estimate the ICER and the NMB from censored observational data. These estimators draw on the strengths of PS weighting and outcome regression fitting utilizing machine learning algorithms. Thus, we have demonstrated the merit of both causal inference models and modern machine learning approaches in CE analysis. We note that the partitioned DR Δ_C estimator, although theoretically more efficient than the simple weighted one, is more computationally intensive. Hence, we suggest using the simple weighted estimator. With smaller sample sizes, the partitioned DR may perform better, but further investigation is needed.

As in any observational study, unobserved or hidden bias may be of concern. Hence, in addition to DR based CE analysis, we suggest conducting sensitivity analyses to assess the effect of unmeasured confounders on the treatment effect.[48]

Acknowledgements

We used the linked SEER-Medicare database and acknowledge the efforts of the Applied Research Program; National Cancer Institute; Office of Research, Development and Information; Centers for Medicare and Medicaid Services; Information Management Services; and SEER program tumor registries in the creation of the SEER-Medicare database.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Clement FM, Harris A, Li JJ, et al. Using effectiveness and cost-effectiveness to make drug coverage decisions: a comparison of Britain, Australia, and Canada. *Jama* 2009; **302**: 1437–1443.
2. Willan AR and Briggs AH. *Statistical analysis of cost-effectiveness data*. Volume 37John Wiley & Sons, 2006.
3. Gomes M, Ng ESW, Grieve R, et al. Developing appropriate methods for cost-effectiveness analysis of cluster randomized trials. *Med Decis Making* 2012; **32**: 350–361.
4. Zhao H and Tian L. On estimating medical cost and incremental cost-effectiveness ratios with censored data. *Biometrics* 2001; **57**: 1002–1008.
5. Dodd S, Bassi A, Bodger K, et al. A comparison of multivariable regression models to analyses cost data. *J Eval Clin Pract* 2006; **12**: 76–86.
6. Basu A, Manning WG and Mullahy J. Comparing alternative models: log vs cox proportional hazard? *Health Econ* 2004; **13**: 749–766.
7. Basu A and Rathouz PJ. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* 2005; **6**: 93–109.
8. Mihaylova B, Briggs A and Hagan AO. Review of statistical methods for analysing healthcare resources and costs. *Health Econ* 2011; **916(August 2010)**: 897–916.
9. Leung SF and Yu S. On the choice between sample selection and two-part models. *J Econometr* 1996; **72**: 197–229.
10. Duan N, Manning WG, Morris CN, et al. Comparison of for alternative care models for the demand medical. *J Bus Econ Stat* 1983; **1**: 115–126.

11. Lin DY, Feuer EJ, Etzioni R, et al. Estimating medical costs from incomplete follow-up data. *Biometrics* 1997; **53**: 419–434.
12. Bang H and Tsiatis A. Estimating medical costs with censored data. *Biometrika* 2000; **87**: 329–343.
13. Zhao H, Bang H, Wang H, et al. On the equivalence of some medical cost estimators with censored data. *Stat Med* 2007; **26**: 4520–4530.
14. Zhao H, Cheng Y and Bang H. Some insight on censored cost estimators. *Stat Med* 2011; **30**: 2381–2388.
15. Raikou M and McGuire A. Estimating medical care costs under conditions of censoring. *J Health Econ* 2004; **23**: 443–470.
16. Young TA. Estimating mean total costs in the presence of censoring. *Pharmacoeconomics* 2005; **23**: 1229–1242.
17. Lin D. Linear regression analysis of censored medical costs. *Biostatistics* 2000; **1**: 35–47.
18. Lin D. Regression analysis of incomplete medical cost data. *Stat Med* 2003; **22**: 1181–1200.
19. Baser O, Gardiner JC, Bradley CJ, et al. Estimation from censored medical cost data. *Biometrical J* 2004; **46**: 351–363.
20. Bang H and Tsiatis A. Median regression with censored cost data. *Biometrics* 2002; **58**: 643–649.
21. Bang H. Medical cost analysis: application to colorectal cancer data from the seer Medicare database. *Contemp Clin Trials* 2005; **26**: 586–597.
22. Willan AR, Lin D and Manca A. Regression methods for cost-effectiveness analysis with censored data. *Stat Med* 2005; **24**: 131–145.
23. Glick HA, Doshi JA, Sonnad SS, et al. *Economic evaluation in clinical trials*. England: Oxford University Press, 2014.
24. Kreif N, Grieve R and Sadique MZ. Statistical methods for cost-effectiveness analyses that use observational data: a critical appraisal tool and review of current practice. *Health Econ* 2013; **22**: 486–500.
25. Anstrom KJ and Tsiatis AA. Utilizing propensity scores to estimate causal treatment effects with censored time-lagged data. *Biometrics* 2001; **57**: 1207–1218.
26. Li J, Handorf E, Bekelman J, et al. Propensity score and doubly robust methods for estimating the effect of treatment on censored cost. *Statistics in medicine* 2016; **35**: 1985–1999.
27. Mitra N and Indurkhya A. A propensity score approach to estimating the cost–effectiveness of medical therapies from observational data. *Health Econ* 2005; **14**: 805–815.
28. Indurkhya A, Mitra N and Schrag D. Using propensity scores to estimate the cost–effectiveness of medical therapies. *Stat Med* 2006; **25**: 1561–1576.
29. Goldfeld K. Twice-weighted multiple interval estimation of a marginal structural model to analyze cost-effectiveness. *Stat Med* 2014; **33**: 1222–1241.
30. Lunceford JK and Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med* 2004; **23**: 2937–2960.
31. Van der Laan MJ, Polley EC, et al. Super learner. *Stat Appl Genet Mol Biol* 2007; **6**(1).
32. Polsky D, Glick HA, Willke R, et al. Confidence intervals for cost-effectiveness ratios: a comparison of four methods. *Health Econ* 1997; **6**: 243–252.
33. Willan AR and O'Brien BJ. Confidence intervals for cost-effectiveness ratios: an application of Fieller's theorem. *Health Econ* 1996; **5**: 297–305.
34. Heitjan DF, Moskowitz AJ and Whang W. Bayesian estimation of cost-effectiveness ratios from clinical trials. *Health Econ* 1999; **8**: 191–201.
35. Robins JM and Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: *AIDS Epidemiology*. Birkhäuser Boston, 1992, pp.297–331.
36. Robins JM, Rotnitzky A and Zhao LP. Estimation of regression coefficients when some of regression coefficients estimation regressors are not always observed. *J Am Stat Assoc* 1994; **89**: 846–866.
37. Cain LE and Cole SR. Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident aids or death. *Stat Med* 2009; **28**: 1725–1738.
38. Cole SR and Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008; **168**: 656–664.
39. Funk MJ, Westreich D, Wiesen C, et al. Doubly robust estimation of causal effects. *Am J Epidemiol* 2011; **173**: 761–767.
40. Westreich D, Lessler J and Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol* 2010; **63**: 826–833.
41. Lee BK, Lessler J and Stuart EA. Improving propensity score weighting using machine learning. *Stat Med* 2010; **29**: 337–346.
42. Thompson SG and Nixon RM. How sensitive are cost-effectiveness analyses to choice of parametric distributions? *Med Decis Making* 2005; **25**: 416–423.
43. Briggs AH, Wonderling DE and Mooney CZ. Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation. *Health Econ* 1997; **6**: 327–340.
44. Siegel R, Naishadham D and Jemal A. Cancer statistics, 2012. *CA* 2012; **62**: 10–29.
45. Goodwin PJ and Shepherd FA. Economic issues in lung cancer: a review. *J Clin Oncol* 1998; **16**: 3900–3912.
46. Black WC, Gareen IF, Soneji SS, et al. Cost-effectiveness of CT screening in the national lung screening trial. *N Engl J Med* 2014; **371**: 1793–1802.

47. Ciunci CN, Mitra N, Yang J, et al. Patterns and effectiveness of surveillance after curative intent surgery in older stage I–IIIa non-small cell lung cancer patients. *J Clin Oncol*, Suppl; abstr 7541.
48. Handorf Ea, Bekelman JE, Heitjan DF, et al. Evaluating costs with unmeasured confounding: a sensitivity analysis for the treatment effect. *Ann Appl Stat* 2013; **7**: 2062–2080.

Appendix 1. DR property of partitioned Δ_C

From 3.2.2, consider $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{L}\left[\frac{Z_i Y_{ij}\delta_i^j}{\hat{e}_i \hat{K}_j(T_i^j)} - \frac{(Z_i-\hat{e}_i)m_1^j(\mathbf{X}_i)\delta_i^j}{\hat{e}_i \hat{K}_j(T_i^j)}\right]$. By the Law of Large Numbers, $\hat{\mu}_1$ estimates

$$E\left[\sum_{j=1}^{L}\frac{ZY^j\delta^j}{eK_j(T^j)} - \frac{(Z-e)m_1^j(\mathbf{X})\delta}{eK_j(T^j)}\right] = \sum_{j=1}^{L}E\left[\frac{ZY^j\delta^j}{eK_j(T^j)} - \frac{(Z-e)m_1^j(\mathbf{X})\delta}{eK_j(T^j)}\right]$$

Take an arbitrary interval j, $\hat{\mu}_{1,j}$ estimates

$$\begin{aligned}
&E\left[\frac{ZY^j\delta^j}{eK_j(T^j)} - \frac{(Z-e)m_1^j(\mathbf{X})\delta}{eK_j(T^j)}\right]\\
&= E\left[\frac{ZY^{(1),j}\delta^j}{eK_j(T^j)} - \frac{(Z-e)m_1^j(\mathbf{X})\delta}{eK_j(T^j)}\right]\\
&= E\left[\frac{\delta^j}{K_j(T^j)}Y^{(1),j} + \frac{(Z-e)}{e}\frac{\delta^j}{K_j(T^j)}\left(Y^{(1),j} - m_1^j(\mathbf{X})\right)\right]\\
&= E\left[Y^{(1),j}\right] + E\left[\left(\frac{Z}{e}-1\right)\frac{\delta^j}{K_j(T^j)}\left(Y^{(1),j} - m_1^j(\mathbf{X})\right)\right]\\
&= \mu_{1,j} + E\left[\left(\frac{Z}{e}-1\right)\frac{\delta^j}{K_j(T^j)}\left(Y^{(1),j} - m_1^j(\mathbf{X})\right)\right]
\end{aligned}$$

Hence for $\hat{\mu}_{1,j}$ to be unbiased, we need the second term $S = E\left[\left(\frac{Z}{e}-1\right)\frac{\delta^j}{K_j(T^j)}\left(Y^{(1),j} - m_1^j(\mathbf{X})\right)\right]$ to be zero. This condition is satisfied when the propensity score model is correctly specified: $E(Z|Y^{(1)},\mathbf{X}) = E(Z|\mathbf{X}) = e(\mathbf{X},\beta) = e$ so

$$\begin{aligned}
S &= E\left[E\left[\left(\frac{Z}{e}-1\right)\frac{\delta^j}{K_j(T^j)}\left(Y^{(1),j} - m_1^j(\boldsymbol{X})\right)|\,Y^{(1)},\mathbf{X}\right]\right]\\
&= E\left[\left(\frac{E(Z|Y^{(1)},\mathbf{X})}{e}-1\right)\frac{\delta}{K_j(T^j)}\left(Y^{(1),j} - m_1^j(\mathbf{X})\right)\right] = 0
\end{aligned}$$

When the outcome model $m_1^j(\mathbf{X})$ is correctly specified, $m_1^j(\boldsymbol{X}) = E(Y^j|Z=1,\mathbf{X}) = E(Y^{(1),j}|Z=1,\mathbf{X}) = E(Y^{(1),j}|Z,\mathbf{X})$ so

$$\begin{aligned}
S &= E\left[E\left[\left(\frac{Z}{e}-1\right)\frac{\delta_j}{K_j(T^j)}\left(Y^{(1),j} - m_1^j(\mathbf{X})\right)|Z,\mathbf{X}\right]\right]\\
&= E\left[\left(\frac{Z}{e}-1\right)\left(E\left(\frac{\delta_j}{K_j(T^j)}Y^{(1),j}|Z,\mathbf{X}\right) - \frac{\delta_j}{K_j(T^j)}m_1^j(\mathbf{X})\right)\right]\\
&= E\left[\left(\frac{Z}{e}-1\right)\left(E(Y^{(1),j}|Z,\mathbf{X}) - \frac{\delta_j}{K_j(T^j)}m_1^j(\mathbf{X})\right)\right] = 0
\end{aligned}$$

Hence $\mu_{1,j}$ is unbiased if either the propensity score model e or the outcome model m_1^j is correctly specified. Adding all the L intervals together, the DR property holds for Δ_E as long as either the propensity score model e or the outcome models m_0^j and m_1^j are correct.