Inverse probability weighting and doubly robust standardization in the relative survival framework

Elisavet Syriopoulou1,2 | Mark J. Rutherford1 | Paul C. Lambert1,2

1Biostatistics Research Group, Department of Health Sciences, University of Leicester, Leicester, UK
2Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Correspondence
Elisavet Syriopoulou, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, SE-17177 Stockholm, Sweden. Email: elisavet.syriopoulou@ki.se

Funding information
Cancer Research UK, Grant/Award Number: C1483/A18262; Research Trainees Coordinating Centre, Grant/Award Number: DRF-2017-10-116

A commonly reported measure when interested in the survival of cancer patients is relative survival. Relative survival circumvents issues with inaccurate cause of death information by incorporating the expected mortality rates of cancer individuals from population lifetables of the general population. A summary of the cancer population prognosis can be obtained using the marginal relative survival. To explore differences between exposure groups, such as socioeconomic groups, the difference in marginal relative survival between exposed and unexposed can be obtained and under assumptions interpreted as the average causal effect of exposure to survival. In a modeling context, this is usually estimated by applying regression standardization as the average of the individual-specific estimates after fitting a relative survival model. Regression standardization yields an estimator that consistently estimates the causal effect under standard causal inference assumptions and if the relative survival model is correctly specified. We extend inverse probability weighting (IPW) and doubly robust standardization methods in the relative survival framework as additional valuable tools for obtaining average causal effects when correct model specification might not hold for the relative survival model. IPW yields an unbiased estimate of the average causal effect if a correctly specified model has been fitted for the exposure (propensity score) whereas doubly robust standardization requires that at least one of the propensity score model or the relative survival model is correctly specified. An example using data on melanoma is provided and a simulation study is conducted to investigate how sensitive are the methods to model misspecification, including different ways for obtaining standard errors.

KEYWORDS
doubly robust standardization, inverse probability weighting, Monte Carlo simulation study, regression standardization, relative survival

1 | INTRODUCTION

When investigating survival using cancer registry data, the event of interest is usually death due to a specific cancer. However, other competing events that preclude the occurrence of the event of interest may also be present. In the presence...
of competing events, there are two main approaches for quantifying cancer survival: (i) cause-specific survival and (ii) relative survival (and its mortality analogue excess mortality). In the cause-specific approach, patients who died from causes different than the cancer of interest are censored at their time of death and only deaths due to the cancer of interest are recorded as events. Thus, accurate classification on the cause of death is required. However, the information on the cause of death that is available in cancer registries might be unreliable or not available at all, especially for older patients who are more prone to die from causes other than their cancer. The relative survival approach circumvents problems with the inaccurate cause of death information and it has been shown to provide more appropriate estimates. Rather than utilizing the cause of death information, excess mortality is obtained by incorporating information on the expected mortality rates for each individual in the cancer population based on the mortality rates of individuals with the same covariate pattern in the general population that is assumed to be free from the cancer of interest. This information is obtained from available life tables in the general population. In this article, the focus is on the relative survival framework. Applications of cancer registry data often focus on exploring variation in survival after a cancer diagnosis across population groups. For instance, it might be of interest to investigate relative survival differences between socioeconomic groups of melanoma patients. Other exposures of interest include sex, race, region, or comorbidities. Cancer-specific exposures might also be considered, for example, stage at diagnosis, treatment, and many more. Recent work has extended relative survival within a formal causal framework for studying survival differences across population groups. Similar to a standard survival framework, assumptions and statistical models are formulated using the counterfactual outcomes framework. For a binary exposure, these would be the outcomes that would be observed if a patient was exposed or unexposed. The average causal effect of interest is defined as the difference in marginal relative survival if everyone was exposed and everyone was unexposed and, under certain assumptions that are extended in the relative survival framework, is identifiable using observed data.

In the standard survival framework, estimates of the average causal effect can be obtained by applying regression standardization (g-computation) and inverse probability weighting (IPW), with each approach requiring different assumptions for the statistical models fitted. Doubly robust methods have also been suggested. In the relative survival framework, estimates of marginal relative survival and the average causal effect are usually obtained by applying regression standardization. This is performed in a similar way as for a standard survival setting, but this time a relative survival model is fitted and information on expected mortality rates is incorporated. Under certain assumptions, regression standardization yields an estimator that consistently estimates the causal effect, if the correct relative survival model has been fitted conditional on exposure and confounders for the exposure and survival outcome. To deal with settings where correct specification of the relative survival model cannot be assumed, extensions of other approaches to relative survival are required. Extending IPW to the relative survival framework is however challenging and requires further adjustments to be able to estimate marginal relative survival. Doubly robust standardization has also not been adapted to the relative survival framework even though doubly robust estimators represent an important advance in methods for estimating causal effects from observational data, as model misspecification is a very common issue.

In this article, we extend IPW and doubly robust standardization methods in the relative survival framework as additional valuable tools for obtaining marginal measures and average causal effects when correct specification of the relative survival model might not hold. The remainder of the article is structured as follows. In Section 2, we introduce excess mortality and relative survival measures and define the measures of interest, that is, the marginal relative survival and the average causal relative survival difference between exposed and unexposed. Then, estimating approaches are discussed. Regression standardization in the relative survival framework is described in Section 3. An extension that enables applying IPW within the relative survival is described in Section 4, followed by doubly robust standardization in Section 5. In Section 6, we utilize data on melanoma to compare estimates obtained after applying regression standardization, IPW and doubly robust standardization. A formal comparison of the methods is provided in Section 7 using a Monte Carlo simulation study. We conclude with a summary of the described methods and a discussion on potential strengths and limitations in Section 8.

# EXCESS MORTALITY AND RELATIVE SURVIVAL

Assume that we are interested in exploring the effect of an exposure on cancer survival. Let \( X \) denote a binary exposure with values 0 and 1 for the unexposed and exposed individuals, respectively. For the remaining of the article, lower case letters with subscript \( i \) will be used to denote the observed value of an individual \( i \) (\( X = x_i \)) and without subscript to denote an assigned value (\( X = x \)). To quantify cancer survival using the relative survival framework, information on both all-cause survival and expected survival if patients had not had the cancer is required. All-cause survival corresponds to
the observed survival in the cancer population. Expected survival is considered to be known (fixed) and is obtained by stratified population lifetables from the general population. In this way, individual-level data from the general population is used as a proxy for the expected survival if cancer patients did not have the cancer under study. Stratification of the population lifetables enables the use of expected survival probabilities of individuals with similar characteristics to the cancer patients, for example, sex, age, year, deprivation. Let $Z$ denote the set of all confounders, with $Z_1$ and $Z_2$ denoting the confounders between the exposure and the expected or relative survival, respectively. Often the confounders for which the population lifetable is stratified, $Z_1$, will be a subset of the confounders included in the relative survival model, $Z_2$. In this case, $Z_2$ will be the same as $Z$. For instance, sex and age affect other cause mortality rates (ie, expected mortality) but they also affect cancer mortality rates. However, in some cases we might want to allow other cause mortality to vary by calendar year but assume that calendar year does not affect mortality due to a specific cancer. Confounders $Z_2$ will also include cancer-related factors such as stage at diagnosis, treatment, etc.

The observed (all-cause) mortality rate of an individual $i$ with observed exposure value $X = x_i$ at time $t$ since diagnosis, $h(t|X = x_i, Z = z)$, is defined as the summation of their expected mortality if they did not have the cancer, $h^*(t|X = x_i, Z_1 = z_{1i})$, and their excess mortality due to the cancer, $\lambda(t|X = x_i, Z_2 = z_{2i})$:

$$h(t|X = x_i, Z = z_i) = h^*(t|X = x_i, Z_1 = z_{1i}) + \lambda(t|X = x_i, Z_2 = z_{2i}).$$

The expected mortality rates are considered to be known and are obtained from stratified population lifetables from the general population. Even though Equation (1) implies that the expected mortality is conditional on the exposure, our approach can be generalized in cases where the exposure $X$ has an effect only on cancer-related mortality rates. This would be for example the case if a cancer-related variable was under study, such as stage at diagnosis or treatment that have no direct effect on expected mortality.

By transforming to the survival scale, the observed survival of patient $i$, $S(t|X = x_i, Z = z_i)$, is given as the product of their expected survival if they did not had the cancer, $S^*(t|X = x_i, Z_1 = z_{1i})$, and their relative survival, $R(t|X = x_i, Z_2 = z_{2i})$:

$$S(t|X = x_i, Z = z_i) = S^*(t|X = x_i, Z_1 = z_{1i})R(t|X = x_i, Z_2 = z_{2i}).$$

A key point in Equation (2) is that the expected and relative survival probabilities vary across individuals with different values for $Z_1$ and $Z_2$, respectively. For the relative survival function $R(t|X = x_i, Z_2 = z_{2i})$ only confounders $Z_2$ are included in the relative survival model. However, the expected mortality rates of individuals are also incorporated in the model and these are obtained by the population lifetables which are stratified by confounders $Z_1$.

Relative survival can be interpreted as the probability of survival in a hypothetical world where it is possible to eliminate competing events (ie, net survival) under the following assumptions:

(i) the two competing events, death due to the cancer of interest and death due to other causes, are conditionally independent and so there are no other factors to affect both competing events than the factors we have adjusted for and (ii) there is appropriate information on the expected survival probabilities. The independence assumption is the same as the one required for the cause-specific approach and its validity cannot be tested formally. For the second assumption, it is important to have sufficiently stratified population lifetables by $Z_1$, so that the expected survival probabilities are representative of what the cancer patients would experience if they did not have their cancer. In cases where other factors need to be considered but they are not available at a population level, adjustments of the mortality rates have been suggested.

The setting of interest is also shown in a simplified DAG in Figure 1. We are interested in the effect of an exposure on cancer survival. However, we have no accurate information on the cause of death. Instead, for each individual we observe an event time $T$ and an event indicator $\delta$, with the term “event” referring to death due to any cause (including also deaths due to competing events). This is shown in the DAG using arrows from the exposure $X$ on both cancer mortality (excess mortality rate, $\lambda(t)$) and other cause mortality (expected mortality rate, $h^*(t)$). To obtain the effect of the exposure on cancer survival, we need to isolate the path of exposure through $\lambda(t)$. For this, we need to control for the path that is due to $h^*(t)$. In practice, we control for it using population lifetables that are stratified by confounders $Z_1$. Adjusting for $Z_1$ is essential in the setting of relative survival even in cases where the exposure does not affect the expected mortality (eg, cancer-related exposures). This is because, as shown in Equation (2), the relative survival approach utilizes expected mortality rates from lifetables as a proxy for other cause mortality instead of relying on cause of death information as for the cause-specific approach.
FIGURE 1 Directed acyclic graph showing the relationship between exposure $X$, excess $\lambda(t)$ and expected mortality rates $h^*(t)$, and outcome with event time and event indicator $(T, \delta)$, while adjusting for confounders $Z_1$ and $Z_2$. Confounders $Z_1$ are usually a subset of confounders $Z_2$.

2.1 Causal quantities of interest

Using the counterfactual outcomes framework, the estimand of interest is defined as the counterfactual marginal relative survival when setting $X = x$:

$$\theta(t|x) = E[R(t|X = x, Z_2)]$$

with the expectation taken over the marginal distribution of $Z_2$. Here the exposure is set $X = x$ for everyone in the study population, as opposed to keeping their observed value $X = x_i$.

Under assumptions, the counterfactual marginal relative survival provides a summary for the survival of the whole population, in a net-world setting where the cancer of interest is the only possible cause of death. To quantify survival in a real-world setting in which both cancer and other causes of death are present, the marginal all-cause survival and marginal crude probabilities of death (ie, the relative survival analogue of cumulative incidence functions) can be obtained instead. These are discussed in more detail elsewhere.\(^7\) If interest is on the hazard function, then the marginal excess rate can be estimated instead. Using the relationship that links the hazard and survival functions $h(t) = -\frac{\ln(S(t))}{dt}$, the marginal excess mortality rate is defined as

$$E[\lambda(t|X = x, Z_2)] = \frac{E[R(t|X = x, Z_2)\lambda(t|X = x, Z_2)]}{E[R(t|X = x, Z_2)]}.$$ 

Contrasts between the counterfactual marginal survival functions if everyone was exposed and everyone was unexposed can be formed to define the average causal effect that is identifiable under certain assumptions. There are many different contrasts that could be defined but in this article the average causal effect of interest is defined as the difference in marginal relative survival probabilities as a function of time:
with the first term setting \( X = 1 \) and the second term setting \( X = 0 \) for everyone in the study population. This allows the comparison of survival if everyone was exposed vs if everyone was unexposed.

### 2.2 Identifying assumptions

To link the counterfactual outcomes with the observed outcomes and obtain an estimate for the average causal difference defined in Equation (5), a number of assumptions are required.\(^{11,26}\) The assumptions that need to hold are similar to those of a standard survival setting, but this time they are extended to both competing events: death due to cancer and death due to other causes. The assumptions are (i) conditional exchangeability so that each outcome is independent of the exposure given confounders, (ii) consistency, that is, an individual’s potential outcome under a specific exposure corresponds to the actual outcome of this person under this exposure level, and (iii) positivity so that the probability of being in every level of the exposure group is positive for all individuals. The main limitation here is that conditional exchangeability for the other cause mortality can only be achieved by adjusting the available population lifetables of the general population for sufficient variables. For instance, assume that interest is in exploring the effect of certain comorbidities on cancer mortality. Individuals with comorbidities have different expected mortality to individuals with no comorbidities, so comparing the mortality of the cancer patients with comorbidities to the mortality of the total population is not appropriate. Expected mortality rates from the general population stratified by comorbidities should be utilized instead. If individual-level information on comorbidities is not available for the general population, the conditional exchangeability for other cause mortality is violated. However, in principle, lifetables can be constructed for any number of covariates and several methods have been suggested for adjusting the expected mortality rates when relevant covariates are not available at a population level.\(^{23-25}\) In addition to the standard causal inference assumptions mentioned above, assumptions relevant to relative survival need to hold and these were discussed in Section 2.

### 3 Regression standardization in the relative survival framework

The marginal relative survival, defined in Equation (3), can be estimated as the standardized relative survival, using regression standardization.\(^7\) First, a relative survival model should be fitted. Several models have been suggested for relative survival, and in principle the methods can be applied to any of these, but this article will focus on flexible parametric survival models.\(^{27-30}\) Flexible parametric models (FPMs) explicitly estimate the baseline log cumulative excess hazard by using restricted cubic splines for the logarithm of time \( t \) rather than assuming linearity with time.\(^{31,32}\) The complexity of the available data dictates the number of knots (or the number of degrees of freedom (df) that is equal to the number of knots minus 1) needed to accurately capture the shape for the underlying hazard. After fitting a relative survival model, predictions of relative survival are obtained for each individual in the study population. The standardized relative survival when setting \( X = x \) is derived as the average of these predictions. For a study population of \( N \) patients, this can be written as:

\[
\hat{\theta}(t|x) = \frac{1}{N} \sum_{i=1}^{N} \hat{R}(t|X = x, Z_2 = z_{2i}).
\]  

Similarly the marginal excess mortality rate of Equation (4) can be estimated using regression standardization by:

\[
E \left[ \hat{\lambda}(t|X = x, Z_2) \right] = \frac{\frac{1}{N} \sum_{i=1}^{N} \hat{R}(t|X = x, Z_2 = z_{2i}) \hat{\lambda}(t|X = x, Z_2 = z_{2i})}{\frac{1}{N} \sum_{i=1}^{N} \hat{R}(t|X = x, Z_2 = z_{2i})}. 
\]

The difference in marginal relative survival functions, defined in Equation (5), can also be estimated as the difference in standardized relative survival functions in which first we set \( X = 1 \) and then set \( X = 0 \) for everyone in the study population:
\[
\frac{1}{N} \sum_{i=1}^{N} \hat{R}(t|X = 1, Z_2 = z_{2i}) - \frac{1}{N} \sum_{i=1}^{N} \hat{R}(t|X = 0, Z_2 = z_{2i}).
\]

(8)

Under identifying assumptions discussed in Section 2.2, regression standardization yields an estimator that consistently estimates the average causal effect, if the correct relative survival model has been fitted for the outcome conditional on exposure and confounders. To deal with settings where correct specification of the relative survival model cannot be assumed, other estimating methods should be applied. Such methods include the IPW and doubly robust standardization and are discussed in the following sections.

4 | INVERSE PROBABILITY WEIGHTS IN THE RELATIVE SURVIVAL FRAMEWORK

IPW builds upon a regression model for the exposure including all relevant confounders that is commonly referred to as the propensity score model. The propensity score is used to derive weights that will then be applied in the survival model that only includes the exposure of interest and so the IPW approach requires a marginal structural model to be fitted. In this article, we focus on simple marginal structural models for a baseline covariate rather than a time-varying covariate. This will be straightforward in a standard survival setting, however, further consideration should be given in a relative survival setting. In a standard survival setting, the estimates of a survival model without confounders should be in good agreement with the estimates of a nonparametric Kaplan-Meier approach. However, this would not be the case for a relative survival model and the estimates would differ from the estimates obtained from a nonparametric approach (e.g., Pohar Perme estimator\(^1\)). The reason for this is that even though the excess mortality obtained from the relative survival model will be constant across individuals, the expected mortality rates that are incorporated in the model will vary between patients by the factors for which the lifetables are stratified. This issue is explored in more detail in the following section.

4.1 | A model for marginal relative survival

When no confounders are included in the model, a simple but incorrect way of obtaining the all-cause mortality when setting \(X = x\) is:

\[
h(t|X = x, Z_1 = z_{1i}) = h^*(t|X = x, Z_1 = z_{1i}) + \lambda(t|X = x),
\]

where the excess mortality remains constant across individuals but the expected mortality varies for individuals with different confounders \(Z_1\) from the lifetable (such as sex, age, and calendar year). However, most of the confounders \(Z_1\) would most probably influence also the excess mortality. This breaks the conditional exchangeability assumption for cancer mortality unless all factors \(Z_1\) do not influence \(\lambda(t|X = x)\). Thus, even in the case where interest is in estimates of marginal relative survival, confounders within the population lifetables should be modeled and the individual predictions should be averaged to yield the marginal estimates (i.e., applying regression standardization).

A more appropriate approach would be to incorporate in the model the marginal expected mortality rate at time \(t\), \(\overline{h}^*(t|X = x)\), instead of the individual expected rates. This would result in the following model for the marginal all-cause mortality:

\[
h_m(t|X = x) = \overline{h}^*(t|X = x) + \lambda(t|X = x)
\]

and under this approach, it would not be necessary to include confounders from the population lifetable in the relative survival model when interest is only on the marginal effect. The marginal relative survival can be obtained as

\[
R_m(t|X = x) = \exp \left[ - \int_0^t \lambda_m(u|X = x) du \right].
\]

A naive approach for calculating \(\overline{h}^*(t|X = x)\) is to obtain the average of the individual expected mortality rates among those at risk at time \(t\). However, relative (net) survival is survival in a hypothetical world where it is not possible to
die from causes other than cancer. The cancer population will be systematically different, with individuals with higher other cause mortality rates being underrepresented as time since diagnosis increases. Weights are required to upweight individuals with a higher probability of dying from other causes, that is, lower expected survival probability. An estimator that incorporates such weights for the mean expected hazard at risk time $t$ has recently been proposed by Lambert et al.:

$$
\overline{h^*}(t|X = x) = \frac{\sum_{j \in R(t)} w_j^*(t) h_j^*(t|X = x, Z_1 = z_{ij})}{\sum_{j \in R(t)} w_j^*(t)},
$$

(9)

where $R(t)$ is the set of those at risk at time $t$ and with weights $w_j^*(t)$ varying by individual and time and being equal to the inverse of the expected survival at time $t$:

$$
w_j^*(t) = \frac{1}{S^*(t|X = x, Z_1 = z_{ij})}.
$$

For situations where the exposure of interest has an effect only on the cancer-related mortality and not on the expected mortality (eg, stage at diagnosis), the weights can be obtained as the inverse of the expected survival probabilities conditioned only on $Z_1$ and no $X$.

The contribution of individual $i$ in the log-likelihood, incorporating the weights is then given by:

$$
\ln L_i = d_i w_i^*(t_i) \ln \overline{h^*}(t_i|X = x) + \lambda_m(t_i|X = x)) - \int_0^{t_i} w_i^*(u) \lambda_m(u|X = x) du,
$$

(10)

where $d_i$ is an event indicator for death due to any cause for individual $i$, with $d_i = 1$ if individual $i$ had the event and $d_i = 0$ otherwise.

Lambert et al. showed that marginal estimates of expected mortality rates can be obtained in Stata using the user-written command mrsprep to expand the data. In this way, time-dependent weights can be incorporated when subsequently fitting models. The integral of the log-likelihood is estimated by splitting the timescale into a number of intervals and then assuming that the weight is constant within each interval, with the number of intervals being chosen by the analyst. After obtaining the weights, standard parametric relative survival models can be fitted, assuming that the weights can be incorporated in the likelihood.

### 4.2 Inverse probability weighting

The IPW approach can be implemented in a relative survival setting by utilizing the above marginal expected mortality estimator, in the following steps. First, a regression model is fitted for the exposure given all relevant confounders (propensity score). For instance, a logistic regression might be fitted for a binary exposure. For each level of the exposure, predictions $\hat{P}(X = x|Z = z_i)$ are obtained for each individual $i$ based on the fitted model and these are used to derive relevant weights ($w_i^*$):

$$
\frac{1}{\hat{P}(X = 1|Z = z_i)} \quad \text{for the exposed and weights} \quad \frac{1}{1 - \hat{P}(X = 1|Z = z_i)} \quad \text{for the unexposed.}
$$

Stabilized weights that result in narrower confidence intervals might also be applied, by including the probability of being exposed or being unexposed in the numerator:

$$
\frac{\hat{P}(X = 1)}{\hat{P}(X = 1|Z = z_i)} \quad \text{for the exposed and weights} \quad \frac{\hat{P}(X = 0)}{1 - \hat{P}(X = 1|Z = z_i)} \quad \text{for the unexposed.}
$$

The marginal expected mortality is then derived, separately for each exposure group. This can be obtained by replacing the weights $w_i^*(t)$ of Equation (9) with weights $w_i(t)$:

$$
w_i(t) = w_i^*(t) \times w_i^d
$$
to incorporate the weights calculated from the propensity score. A marginal relative survival model is then fitted incorporating weights \(w_i(t)\) and the marginal expected mortality rates. The updated time-dependent weights, \(w_i(t)\), should also be included into the likelihood of Equation (10). Under the identifying assumptions, IPW yields an unbiased estimate of the average causal difference in marginal relative survival if the correct model has been fitted for the exposure conditional on the confounders.

5 | DOUBLY ROBUST STANDARDIZATION IN THE RELATIVE SURVIVAL FRAMEWORK

Doubly robust methods have also been suggested as a way to estimate the average causal effect and these have been showed to be more robust in the model misspecification.\(^\text{15-18}\) A trade-off between potentially reducing bias at the expense of precision has however been reported for doubly robust estimators, although they have been suggested to be more efficient than the usual IPW estimator.\(^\text{16}\) In a standard survival setting, doubly robust standardization is applied as a two-step procedure and it combines a survival model with weighting by the propensity score.\(^\text{15}\) First, a propensity score model is fitted in the same way as for IPW. Then a survival model including exposure and relevant confounders for the exposure and survival outcome is fitted by incorporating the weights. The second step is similar to the survival model fitted in regression standardization but this time the weights are being used in the fitting procedure. After fitting the survival model, predictions are obtained for each level of exposure and for every individual in the study population and these are averaged and form relevant contrasts.

Extending doubly robust standardization in the relative survival framework can be performed in a similar way. The main difference is that a relative survival model should be fitted instead of standard survival model. First, a propensity score model is fitted for the exposure given confounders \(Z\) to obtain the probability of being exposed. Predictions for each exposure level \(X = x\) given confounders, \(\hat{P}(X = x|Z = z_i)\), are then derived and these predictions are used to create the weights. These weights are calculated in the same way as weights \(w_i\) from the IPW approach in Section 4.2. The next step is to fit a relative survival model using the weights and including also the exposure and relevant confounders. Once again predictions are obtained for the relative survival of each individual in the study population by setting \(X = 1\) and \(X = 0\) to obtain the predicted counterfactual outcomes. The individual predictions are then averaged over the whole population and the difference in marginal relative survival functions between setting \(X = 1\) and \(X = 0\) is calculated as in Equation (8).

The implementation of doubly robust standardization is not more complex than that of regression standardization and IPW and it has the advantage of being more robust on model misspecification.\(^\text{15}\) Doubly robust standardization requires that at least one of the propensity score model or the survival model is correctly specified and yields an estimator that consistently estimate the average causal difference even if one of the two models is misspecified. As model misspecification is a very common issue with observational data, doubly robust estimators represent an important advance in methods for estimating the average causal difference in relative survival.

6 | ILLUSTRATIVE EXAMPLE

6.1 | Data

The methods described above are illustrated using data on melanoma and exploring relative survival differences between the least and most deprived patients. Factors such as deprivation are important health determinants and quantifying the overall health effects of these variables is often a natural starting point for inequalities research. As others have argued, understanding the magnitude of disparities across deprivation groups in a formalized causal framework gives a firm basis to further unpick the reasons for the differences, even if the ideal randomized experiment would be difficult to precisely define.\(^\text{34}\)

Data included all individuals diagnosed with melanoma in the region of East Midlands in England between 2008 and 2013 and were made available by Public Health England after request. The available information includes sex, age, and year at diagnosis as well as deprivation status. Deprivation status is a categorical variable with the deprivation quintiles calculated using the 2010 Index of Multiple Deprivation (IMD) based on the area of residency at the time of diagnosis.\(^\text{35}\) There are five ordinal deprivation groups, however, as the main purpose of the analysis was to demonstrate the methods,
only a subset of the population is utilized, that is, the least and most deprived groups. The final data included 1905 patients, 79% of which were in the least deprived group. More details on the study population can be found in Table 1. The purpose of this example is simply to demonstrate the methods and caution is required when interpreting the results as adjusting for sufficient confounders is essential for estimating the average causal effect.

### 6.2 Statistical models

Using the above data we estimated the marginal 5-year relative survival for the least and most deprived melanoma patients as well as the survival difference between the two deprivation groups by standardizing to the whole population and applying (i) regression standardization, (ii) IPW, and (iii) doubly robust standardization. For the analysis of the data, we fitted several flexible parametric survival models of increasing complexity. All models assumed 5 degrees of freedom for the baseline excess hazard. Year of diagnosis is often omitted from the analysis even though there is small variation in relative survival by year since diagnosis. Below we consider models both with and without year of diagnosis, to explore differences in the marginal relative survival estimates.

More specifically, for regression standardization we fitted the following models:

- Model 1: Included deprivation status and sex without allowing for any time-dependent effects.
- Model 2: Included deprivation status, sex as well as age at diagnosis as a continuous linear variable. The model allowed time-dependent effects only for deprivation and sex (3 df).
- Model 3: Included deprivation status, sex and age at diagnosis that was modeled as a continuous nonlinear variable by using restricted cubic splines (3 df). Time-dependent effects were allowed for all variables (3 df for each covariate with a time-dependent effect).
- Model 4: Same as model 3 but also included year of diagnosis in the model.

For IPW:

- Approach 1: A logistic regression was fitted for deprivation status including sex, year of diagnosis, and age (restricted cubic splines with 3 df) and this was utilized to obtain the relevant weights. Then, a FPM including only deprivation and 5 df for the baseline excess hazard was fitted. A time-dependent effect was also allowed for deprivation (3 df). This approach corresponds to a naive IPW as the model incorporates individual expected mortality rates and is expected to give biased estimates.
• Approach 2: Same logistic regression as in approach 1. Then we utilized the derived weights (in the mrsprep) to obtain the marginal expected mortality, separately for the least and most deprived. Finally, a marginal relative survival model, including only deprivation status is fitted by incorporating the marginal expected mortality as discussed in Section 4).

• Approach 3: Same as approach 2 but the marginal relative survival model includes also the time-dependent effect of deprivation (3 df).

For the doubly robust standardization:

• Approach 1: First, a logistic regression was fitted for deprivation status including sex. Then a relative survival FPM (5 df for the baseline excess hazard) is fitted using the weights including also sex and age (with restricted cubic splines, 3 df). Time-dependent effects are also allowed (3 df for each covariate with a time-dependent effect).

• Approach 2: Same as above but this time age at diagnosis (modeled as a nonlinear continuous variable using restricted cubic splines with 3 df) as well as year of diagnosis was also included in the logistic regression. Year of diagnosis was also included in the relative survival model.

Expected mortality rates were incorporated in the relative survival models using available population lifetables stratified by sex, age, calendar year, and deprivation status. 36

Robust standard errors were also derived. For regression standardization, standard errors were obtained by applying either the delta method (models 1-4) or M-estimation (only for model 4). M-estimation is a generalized form of the delta method that takes into consideration the observed variation in the confounder distribution of a population. 37,38 For IPW and doubly robust standardization, standard errors were obtained by applying the delta method, as M-estimation is not currently implemented in Stata. More details on the standard errors can be found in Appendix A.

6.3 Results for melanoma data

When regression standardization was applied (model 4) the 5-year relative survival of the least deprived was found to be approximately 91.0% (Table 2). This estimate was very similar when year of diagnosis was omitted from the relative survival model (model 3). The estimates of regression standardization were also very similar with the 5-year relative survival estimates obtained with doubly robust standardization. Doubly robust standardization yielded similar estimates under different fitted models (approaches 1 and 2). Under IPW and when individual expected mortality rates were incorporated in the marginal model, the 5-year relative survival of the least deprived was equal to 92.8% (approach 1 of naive IPW). However, by appropriately incorporating the marginal expected mortality in the model as discussed in Section 4, the 5-year relative survival was similar to the one found from the other approaches (91.0% under approach 2 of IPW). Allowing for time-dependent effects in the relative survival model that incorporates the weights had a small effect on the estimates (approach 3 of IPW). A similar pattern was also observed for the 5-year relative survival of the most deprived as well as the 5-year difference between the least and most deprived.

The standard errors obtained from regression standardization with the M-estimation were slightly larger than the ones from the delta method. This is because M-estimation incorporates some uncertainty for the covariates variation. Similar standard errors were also observed with doubly robust standardization. The standard errors of the naive IPW approach were smaller than all other approaches but the standard errors of the IPW approach, where the marginal expected mortality is incorporated in the model, were slightly larger than those obtained from the delta method of regression standardization.

7 A MONTE CARLO SIMULATION STUDY

A Monte Carlo simulation study was performed to compare different methods for obtaining marginal measures of interest. To plan the simulation study, the ADEMP structured approach was applied, according to which special consideration should be given to the aims, data-generating mechanisms, methods, estimands, and performance measures. 39-41
| Method                                      | Relative survival | SE     | CI             |
|---------------------------------------------|-------------------|--------|----------------|
| **Least deprived**                          |                   |        |                |
| RegStand (model 1)                          | 0.925             | 0.009  | (0.907, 0.942) |
| RegStand (model 2)                          | 0.916             | 0.010  | (0.896, 0.936) |
| RegStand (model 3)                          | 0.908             | 0.011  | (0.886, 0.930) |
| RegStand (model 4: delta method)            | 0.910             | 0.011  | (0.888, 0.932) |
| RegStand (model 4: M-estimation)           | 0.910             | 0.013  | (0.885, 0.935) |
| IPW (approach 1)                            | 0.928             | 0.009  | (0.910, 0.945) |
| IPW (approach 2)                            | 0.910             | 0.012  | (0.886, 0.935) |
| IPW (approach 3)                            | 0.908             | 0.013  | (0.884, 0.933) |
| DR (approach 1)                              | 0.908             | 0.011  | (0.886, 0.930) |
| DR (approach 2)                              | 0.911             | 0.011  | (0.889, 0.933) |
| **Most deprived**                           |                   |        |                |
| RegStand (model 1)                          | 0.835             | 0.024  | (0.789, 0.882) |
| RegStand (model 2)                          | 0.809             | 0.027  | (0.755, 0.862) |
| RegStand (model 3)                          | 0.800             | 0.027  | (0.746, 0.853) |
| RegStand (model 4: delta method)            | 0.804             | 0.027  | (0.751, 0.857) |
| RegStand (model 4: M-estimation)           | 0.804             | 0.030  | (0.744, 0.863) |
| IPW (approach 1)                            | 0.843             | 0.025  | (0.794, 0.892) |
| IPW (approach 2)                            | 0.798             | 0.030  | (0.740, 0.857) |
| IPW (approach 3)                            | 0.804             | 0.031  | (0.743, 0.864) |
| DR (approach 1)                              | 0.799             | 0.027  | (0.746, 0.930) |
| DR (approach 2)                              | 0.812             | 0.029  | (0.756, 0.868) |
| **Difference**                              |                   |        |                |
| RegStand (model 1)                          | −0.089            | 0.025  | (−0.139, −0.040) |
| RegStand (model 2)                          | −0.108            | 0.028  | (−0.163, −0.053) |
| RegStand (model 3)                          | −0.109            | 0.028  | (−0.164, −0.054) |
| RegStand (model 4: delta method)            | −0.107            | 0.028  | (−0.161, −0.052) |
| RegStand (model 4: M-estimation)           | −0.107            | 0.030  | (−0.166, −0.047) |
| IPW (approach 1)                            | −0.085            | 0.026  | (−0.137, −0.034) |
| IPW (approach 2)                            | −0.112            | 0.032  | (−0.175, −0.048) |
| IPW (approach 3)                            | −0.105            | 0.033  | (−0.170, −0.039) |
| DR (approach 1)                              | −0.109            | 0.028  | (−0.164, −0.054) |
| DR (approach 2)                              | −0.099            | 0.029  | (−0.155, −0.042) |

Note: Standard errors (SE) and 95% confidence intervals (CI) are also provided. Abbreviations: DR, doubly robust standardization; IPW, inverse probability weighting; RegStand, regression standardization.
7.1 | Aims

The main aim of the simulation study was to assess how sensitive point estimates obtained from regression standardization, IPW and doubly robust standardization are to model misspecification. As a secondary aim, different ways for obtaining standard errors for the point estimates are explored. The Stata code used for the simulation study is available online at https://github.com/syriop-elisa/simulation_IPW_DRstand.

7.2 | Data generating mechanisms

Each simulated dataset included 2000 observations. For each dataset, three confounders, $L_1, L_2, L_3$ were generated from Normal distributions. Variable $L_1$ was generated from a Normal distribution with mean 60 and standard deviation 13. Variables $L_2$ and $L_3$ were generated from Normal distributions with mean 0 and standard deviation 1. There were three data-generating mechanisms regarding the correlation between the confounders:

- **DGM-1:** High correlation with correlation matrix equal to
  \[
  \begin{pmatrix}
  1 & 0.8 & 0.8 \\
  0.8 & 1 & 0.8 \\
  0.8 & 0.8 & 1
  \end{pmatrix}
  \]

- **DGM-2:** Medium correlation with correlation matrix equal to
  \[
  \begin{pmatrix}
  1 & 0.5 & 0.5 \\
  0.5 & 1 & 0.5 \\
  0.5 & 0.5 & 1
  \end{pmatrix}
  \]

- **DGM-3:** No correlation with correlation matrix equal to
  \[
  \begin{pmatrix}
  1 & 0 & 0 \\
  0 & 1 & 0 \\
  0 & 0 & 1
  \end{pmatrix}
  \]

A binary exposure variable was also generated from a binomial distribution with the probability of being exposed equal to the inverse logit function of \( \log \left( \frac{\beta_0 + \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3}{1 - \beta_0} \right) \), with $\beta_0 = \logit(0.5)$, $\beta_1 = \log(1.01) = 0.010$, $\beta_2 = \log(1.3) = 0.262$, and $\beta_3 = \log(1.4) = 0.336$. The binary exposure resembles the deprivation status information that is also available on the lifetables.

As information from population lifetables needs to be incorporated in the model, and these are stratified by sex, year, age, and deprivation the relevant variables should be simulated. For simplicity, all individuals were assumed to be male and diagnosed in one specific calendar year, that is, 2009 while variable $L_1$ was allowed to take values between 18 and 99 and was centered around 60 to resemble age.

Both time to death from cancer and death due to other causes were generated for each individual, with the minimum value taken as the time to death. For time to death due to cancer, a Weibull distribution was assumed for the baseline survival:

\[
T = \left( -\frac{\log(U)}{\lambda \exp(\beta^T Z)} \right)^{\frac{1}{\gamma}}
\]

with $Z$ denoting the set of all covariates and $U$ is a random variable with $U \sim U(0, 1)$. For time to death due to other causes, exponential distributions were assumed for the baseline survival:
Further information on simulating survival times can be found elsewhere. More specifically, time to death from cancer was generated from a Weibull distribution with shape parameter $\gamma = 0.5$ and scale parameter $\lambda = 0.2$. Time to death from other causes was generated from exponential distributions using rates from population lifetables in England in 2009. To account for the increase in attained age, a different rate was applied for each year of follow up. If a value larger than 1 was generated, then the individual was assumed to be alive at the start of the next interval. If the value was less than one, then it was assumed that the individual had died in the interval.

The effects of $L_1$, $L_2$, and $L_3$ were assumed to be proportional over time with excess hazard ratios equal to 1.02, 1.3, and 1.5 per one unit increase, respectively. The effect of the exposure (deprivation) was also assumed to be proportional with the hazard ratio equal to 1.2 for the exposed vs unexposed.

We run 1000 iterations and more details on this can be found in the Supplementary Material.

7.3 | Estimands

The estimands of interest include marginal measures for a specific exposure group as well as the difference between exposed and unexposed, both at 1 year and 5 years since diagnosis. In particular, the estimands of interest are:

- 1-Year marginal relative survival of the exposed
- 1-Year marginal relative survival of the unexposed
- Difference in 1-year marginal relative survival between exposure groups
- 5-Year marginal relative survival of the exposed
- 5-Year marginal relative survival of the unexposed
- Difference in 5-year marginal relative survival between exposure groups

7.4 | Obtaining the true values

To calculate the true values, a large sample of 100,000,000 observations was generated following the data-generating mechanisms described in Section 7.2 to generate the exposure and confounder data. Then, the relative survival of each individual was obtained using the following Weibull formula at the timepoints of interest:

$$ S(t) = \exp[-\exp(\beta_0 X + \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3) \lambda t^\gamma] $$

and replacing the $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$, $\lambda$, and $\gamma$ with the values assumed to generate the data. The true values were then derived by obtaining the mean over the whole study population.

7.5 | Methods

The four methods compared are:

- Regression standardization (RegStand)
- Inverse probability weighting (IPW)
- Doubly robust standardization (assuming a correct model for the survival outcome) (DRsurv)
- Doubly robust standardization (assuming a correct model for the exposure outcome) (DRexp)

As the main aim of the simulation study was to assess the impact of model misspecification on the point estimates, a range of models is fitted for each method with confounders gradually omitted. In particular:

- Scenario 1: All confounders $L_1, L_2, L_3$ are included in the relevant model.
- Scenario 2: Only confounders $L_1, L_2$ are included in the relevant model.
- Scenario 3: Only confounder $L_1$ is included in the relevant model.
- Scenario 4: No confounders are included in the relevant model.

Variable $L_1$, that is, age, was included in the population lifetable and is a confounder that affects both other cause and cancer mortality.

The above scenarios apply, either to the survival model of the RegStand approach, the exposure model of the IPW approach, the exposure model of the DRsurv approach or the survival model of the DRexp approach. Note, however, that for DRsurv and DRexp at least one correct model is fitted under each scenario.

7.5.1 Modeling details

For the RegStand approach, several relative survival FPMs were fitted, with each one including a different set of confounders (scenarios 1-4). Each FPM was fitted assuming 3 df for the baseline excess hazard.

For the IPW approach, several logistic regression models were fitted for the exposure to derive the weights. For each logistic regression model, a different set of confounders was included in the model (scenarios 1-4). Next, for the relative survival model, a FPM was fitted (including only the exposure) using the weights, with 3 df for the baseline excess hazard. The marginal expected mortality rates were incorporated into the model as described in Section 4. The proportional hazards effect that was assumed for the exposure and was simulated from the conditional model may not be proportional for the relative survival model.44 Thus, time-dependent effects for the effect of the exposure were allowed in the relative survival model. These were modeled using restricted cubic splines with 3 degrees of freedom.

For the DRsurv approach, several logistic models were fitted assuming a different set of confounders (scenarios 1-4) and weights were obtained. Then, a FPM with 3 df for the baseline excess hazard was fitted using the weights and including exposure and all confounders $L_1, L_2, L_3$. Thus, for DRsurv the relative survival model was always correctly specified.

For the DRexp method, a logistic model was fitted for the exposure including all confounders $L_1, L_2, L_3$ and weights were obtained. Then, several FPMs with 3 df for the baseline excess hazard were fitted using the weights and assuming a different set of confounders (scenarios 1-4). Thus, for DRexp the exposure model was always correctly specified.

7.6 Standard errors

The main aim of the simulation study was to compare the bias on the estimands of interests (see the Supplementary Material), after applying different methods. As a secondary aim, the standard errors obtained from each method were also compared.

Standard errors for the regression standardization were obtained by applying either the delta method (RegStand-d) or the M-estimation (RegStand-m).37,38 IPW has been found to result in biased estimators for the standard errors when a conventional model-based variance estimator from the maximum likelihood estimator is applied. A simulation study showed that bootstrapping yields appropriate standard errors while there is a bias when using robust standard errors.45 In practice, when dealing with large datasets, there is a trade-off between computational time and small bias in standard errors. In the simulation study described in this article, the standard errors of the IPW, DRsurv, and DRexp were obtained with the delta method while using robust clustered standard errors, as the M-estimation is not implemented yet in Stata for these approaches. More details on the delta method and M-estimation are given in Appendix A. The Stata code that was used for the simulation analysis is available at https://github.com/syriop-elisa/simulation_IPW_DRstand, including information on how the standard errors were calculated.

7.7 Results of the simulation study

7.7.1 Comparing point estimates

Bias in the marginal relative survival of the exposed, the marginal relative survival of the unexposed as well as the difference in survival between the exposed and unexposed can be found in Figures 2 to 4, respectively. Data points in
FIGURE 2 Bias for the marginal relative survival of the exposed, both at 1 year and 5 years after diagnosis by method, confounders scenario, and data generating mechanism. Absolute bias larger than 0.01 is shown in orange [Colour figure can be viewed at wileyonlinelibrary.com]

orange refer to an absolute bias larger than 0.01. This is an arbitrary value that was chosen as a reference for the models comparison as the Monte Carlo errors are very small, resulting in narrow confidence intervals that indicate significant bias even when the magnitude of the bias is negligible in practice. All methods gave unbiased results when the full model with all confounders $L_1, L_2, L_3$ was fitted, either for the relative survival model (RegStand approach), exposure model (IPW approach) or both models (DRsurv and DRexp approach).

For the relative survival estimates of the exposed, the regression standardization approach had a small bias under scenario 2 when confounder $L_3$ was omitted from the relative survival model (Figure 2). The bias was larger when both $L_2$ and $L_3$ were omitted from the survival model. The bias was also larger when a lower correlation was assumed between confounders $L_1, L_2, L_3$. However, this was not the case for scenario 4, where all confounders were omitted. In this scenario, the bias was larger than all other scenarios but it was decreasing with a lower correlation between the confounders. This is because, under a high correlation scenario, the model that includes one out of three variables is still explaining a lot of the variability. Omitting this variable from the model would then yield a large bias. That is why, there is such a big gap between the model where $L_2, L_3$ are omitted vs $L_1, L_2, L_3$ are omitted. However, when confounders are not correlated, the bias from a model with only $L_1$ and the bias from a model without any confounders would have a smaller difference. In general, the bias was larger at 5 years.
A similar pattern was observed for the relative survival estimates of the exposed that were obtained with the IPW approach (Figure 2). Bias under IPW was however slightly smaller than the bias of regression standardization for all scenarios apart from scenario 4. When all confounders are omitted from the exposure model, a larger bias was observed for IPW in comparison with the bias that is observed when all confounders were omitted from the relative survival model of regression standardization.

The DRsurv approach, in which the correctly specified survival model was always fitted but different scenarios were assumed for the exposure model, was unbiased even when all confounders are omitted from the exposure model (Figure 2). The DRexp, in which the exposure model was always correctly specified but different scenarios were considered for the relative survival model, was unbiased for scenarios 1, 2, and 3. However, for scenario 4 when all confounders are omitted from the relative survival model, the estimates were biased. This is the only positive bias, as for all the other methods, scenarios and data-generating mechanisms a negative bias was observed and the true effect was underestimated. Bias was larger at 5 years and under DGM-1 when a high correlation was assumed between $L_{1}, L_{2}, L_{3}$. The DRexp approach under scenario 4 is, actually, the equivalent of a naive IPW approach where individual expected mortality rates are incorporated in the model, rather than the marginal expected mortality rates as discussed in Section 4, so bias should not be surprising. Confounder $L_{1}$ which resembles age is now omitted from the relative survival model while the expected mortality rates from the lifetable vary also by age. To obtain unbiased estimates under scenario 4, marginal expected mortality rates should have been incorporated in the relative survival model instead.
A similar pattern of results was also observed for the unexposed (Figure 3). This time, positive values are observed under all scenarios. The absolute values for the bias were slightly larger than the equivalent bias for the exposed.

The bias in the estimand of the difference after applying the regression standardization and IPW approaches had a similar pattern as for the exposed and the unexposed (Figure 4). The magnitude of the bias was found to be very similar between regression standardization and IPW approaches. Once again, increasing bias was observed for increasing level of model misspecification when more confounders were omitted from the relative survival model (regression standardization) or the exposure model (IPW). Bias was larger at 5 years after diagnosis and smaller when a higher correlation was assumed for $L_1, L_2, L_3$. However, when all confounders were omitted from the model, a larger bias was observed under DGM-1 (high correlation). Finally, the bias of the difference was larger than that of the exposed and the unexposed.

Doubly robust standardization approaches, DRsurv and DRexp, yielded unbiased estimates of the difference. Under scenario 4, DRexp gave biased estimates for the exposed and the unexposed (Figures 2 and 3). A negligible bias was however observed for the difference as the absolute values for the bias of the exposed and unexposed were approximately equal (Figure 4).

Detailed tables with the actual values of the bias, together with values for the Monte Carlo standard errors, can be found in the Supplementary Material.
7.7.2 Comparing standard errors

Figure 5 shows the relative error in average model standard error for the estimand of the difference in marginal relative survival between the exposed and unexposed. The relative error provides a comparison of the model (modSE) and empirical (empSE) standard errors and is estimated as \(100 \left( \frac{\text{modSE}}{\text{empSE}} - 1 \right)\). More information on performance measures can be found in the Supplementary Material. Even though the performance measures are shown for all scenarios, focus should be given to the unbiased estimates. The relative error was small for most approaches, suggesting that the model yields an appropriate standard error. The standard errors obtained from IPW were larger. For instance, under DGM-1 that assumes a high correlation and when the full model is fitted for the propensity score, the model standard error that is obtained for the difference at 5 years is 16% higher than the empirical standard error. The relative errors for IPW were also higher when a higher correlation was assumed for the data generating mechanism. In general, further investigation is required to ensure appropriate standard errors under the IPW approach.

Detailed information on the empirical and model standard errors for the difference can be found in Supplementary Tables S5 and S6. For regression standardization, the delta method and the M-estimation gave very similar empirical standard errors. Similar empirical standard errors were observed also for the doubly robust standardization approaches. However, the empirical standard errors obtained for the IPW approach were larger, resulting in loss in precision. Larger empirical standard errors were also observed for the difference at 5 years since diagnosis. The values are similar across different data-generating mechanisms. A similar pattern was observed for the model standard errors and these were
similar between regression standardization and the doubly robust standardization approaches, but larger for the IPW method.

Figure 6 shows the coverage for all scenarios with absolute bias less than 0.01. Coverage was found to be good under the full model. Larger coverage was observed for the IPW approach, reflecting the larger standard errors. For instance, the coverage for the 5-year difference was 0.980 for the IPW under DGM-1 and the full model. More details on the actual values of the performance measures (with Monte Carlo errors) for the difference at 1 and 5 years can be found in Supplementary Tables S5 and S6, respectively.

The pattern was also similar for the exposed and the unexposed. For regression standardization and doubly robust standardization the differences between the empirical and the model standard errors were slightly larger than those observed for the difference, especially when a higher correlation was imposed in confounders $L_1, L_2, L_3$. However, the relative error was lower for the IPW in comparison with the one for the difference. For example, under DGM-1 the model standard error that was obtained for the exposed at 5 years was 5% higher than the empirical standard error. Details on the performance measures for the exposed can be found in Supplementary Tables S1 and S2 for 1 and 5 years, respectively. Similarly, for the unexposed these are available in Tables S3 and S4 for 1 and 5 years, respectively.
We introduced IPW and doubly robust standardization in the relative survival framework as additional methods for obtaining marginal estimates of interest. Regression standardization requires a correctly specified relative survival model, conditional on exposure and confounders, and, under the identifiability assumptions, it yields an estimator that consistently estimates the average causal effect. IPW requires a correct model for the exposure conditional on the confounders. Doubly robust standardization combines a model for relative survival with a model for the exposure and thus it only requires one of the two models to be correctly specified to obtain an unbiased estimate.

IPW requires fitting a marginal survival model, but this is not directly applicable to a relative survival model. For standard relative survival models, individual expected survival probabilities are incorporated in the model rather than the marginal expected probabilities. This will also be the case when no confounders are included in the model for relative survival. To deal with this issue and enable the extension of IPW within relative survival, marginal expected mortality rates should be incorporated instead. Lambert et al.\(^3^3\) recently proposed an estimator for marginal expected mortality. In this article, we utilize this estimator in the IPW approach, including also the weights of the propensity score model.

All three approaches, regression standardization, IPW, and doubly robust standardization, were compared via a simulation study. In particular, the simulation study explored the bias of model misspecification for the relative survival model (regression standardization), the exposure model (inverse probability weighting) or either model (doubly robust standardization). The estimates of interest included both the marginal relative survival of the exposed and unexposed as well as the difference between them, both at 1 and 5 years after diagnosis. Different data-generating mechanisms were also assumed, with a varying level of correlation between the confounders that were omitted from the relevant models.

Both regression standardization and inverse probability approaches yielded small biases when one confounder was omitted from the relative survival or propensity score models, respectively. The bias was increased similarly when two confounders were omitted from the models. However, when all three confounders were omitted from the model IPW gave larger biases. The bias was also larger with increasing time since diagnosis. Doubly robust standardization was unbiased for all scenarios, apart from when all confounders were omitted from the relative survival model, even though these were modeled within the exposure model. This bias corresponds to the bias that arises when fitting a naive IPW approach, so it is not surprising. Omitting variables from the population lifetables (with the expected mortality rates) from the relative survival model will always yield biased estimates, as individual expected mortality rates will be incorporated in the model. That is why, for the IPW approach, the marginal expected mortality rates were obtained first and then were incorporated in the relative survival model rather than using individual-specific expected rates.

Standard errors for each of the approaches described above were also explored. For regression standardization, standard errors were estimated by either applying the delta method or M-estimation, using analytical formulas and therefore avoiding numerical approximation and bootstrap procedures. M-estimation incorporates also the variation in the confounder distribution of the population. The results of the simulations showed that the standard errors were very similar between the delta method and the M-estimation. Small differences between the delta method and M-estimation standard errors were also observed in a sensitivity analysis in which a smaller number of observations were generated (\(N = 200\) per simulated dataset). The model standard errors were also in good agreement with the empirical standard errors. The standard errors of the IPW and doubly robust standardization approaches were only estimated by applying the delta method, as M-estimation is currently not implemented for relative survival in Stata. However, the small difference that was observed between the two methods for regression standardization suggests that M-estimation would probably yield similar estimates with the delta method. In general, when the correct model was fitted all methods gave unbiased results but IPW resulted in larger standard errors and loss in precision. When the relative survival model or the propensity score was misspecified, doubly robust standardization gave unbiased estimates with a very small loss in precision. For IPW, the model standard errors were also found to be larger than the empirical standard errors for most scenarios which warrant further investigation to obtain appropriate standard errors. Finally, all methods yielded good coverage for the full models, with slightly higher coverage for the IPW approach which reflects the larger standard errors.

In this article, only time-fixed exposures and confounders were considered. However, in the presence of time-varying exposures or confounders, causal parameters can also be estimated using appropriate methodology. IPW and standardization methods can be described in terms of marginal structural models to account for covariates that change over time and this consists part of future work.\(^{46-48}\)

In general, all methods performed well when correctly specified models were fitted. However, in practice, when using observational data model misspecification is very common. Thus, doubly robust standardization might be preferable when this is applicable.
ACKNOWLEDGEMENTS
This work was supported by a National Institute for Health Research Doctoral Research Fellowship to E.S. (Reference: DRF-2017-10-116). This article presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. P.C.L. was also funded by Cancer Research UK (Grant number C1483/A18262).

CONFLICT OF INTEREST
The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT
Even though it is not possible to release the data of the illustrative example, as this is only available upon request from Public Health England, we make the Stata code used for the simulation study available for replication and the potential to adapt the parameters to different settings at https://github.com/syriop-elisa/simulation_IPW_DRstand.

ORCID
Elisavet Syriopoulou https://orcid.org/0000-0002-5749-4094

REFERENCES
1. Schaffar R, Rachet B, Belot A, Woods L. Cause-specific or relative survival setting to estimate population-based net survival from cancer? an empirical evaluation using women diagnosed with breast cancer in Geneva between 1981 and 1991 and followed for 20 years after diagnosis. Cancer Epidemiol. 2015;39(3):465-472. https://doi.org/10.1016/j.canep.2015.04.001
2. Mariotto AB, Noone AM, Howlader N, et al. Cancer survival: an overview of measures, uses, and interpretation. J Natl Cancer Inst Monogr. 2015;14(49):145-186.
3. Dickman PW, Adami H-O. Interpreting trends in cancer patient survival. J Internal Med. 2006;260:103-117.
4. Sarfaty D, Blakely T, Pearce N. Measuring cancer survival in populations: relative survival vs cancer-specific survival. Int J Epidemiol. 2010;39:598-610. https://doi.org/10.1093/ije/dyp392
5. Damgaard Skyrud K, Bray F, Møller B. A comparison of relative and cause-specific survival by cancer site, age and time since diagnosis. Int J Cancer. 2014;135(1):196-203. https://doi.org/10.1002/ijc.28645
6. Schaffar R, Rapiti E, Rachet B, Woods L. Accuracy of cause of death data routinely recorded in a population-based cancer registry: impact on cause-specific survival and validation using the Geneva cancer registry. BMC Cancer. 2013;13:609.
7. Syriopoulou E, Rutherford MJ, Lambert PC. Marginal measures and causal effects using the relative survival framework. Int J Epidemiol. 2020;49:619-628. https://doi.org/10.1093/ije/dyz268
8. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. J Am Stat Assoc. 2005;100(469):322-331.
9. Hernán MA. A definition of causal effect for epidemiological research. J Epidemiol Commun Health. 2004;58(4):265-271.
10. Robins J. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. J Chronic Dis. 1987;40:139S-161S.
11. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. Am J Epidemiol. 2011;173:731-738. https://doi.org/10.1093/aje/kqw472
12. van der Wal WM, Prins M, Lumbraeras B, G eskus RB. A simple G-computation algorithm to quantify the causal effect of a secondary illness on the progression of a chronic disease. Stat Med. 2009;28(18):2325-2337. https://doi.org/10.1002/sim.3629
13. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Commun Health. 2006;60(7):578-586.
14. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res. 2011;46:399-424. https://doi.org/10.1080/00273171.2011.568786
15. Vansteelandt S, Keiding N. Invited commentary: G-computation-lost in translation? Am J Epidemiol. 2011;173(7):739-742.
16. Jonsson Funk M, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. Am J Epidemiol. 2011;173(7):761-767. https://doi.org/10.1093/aje/kqw439
17. Robins J, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: performance of double-robust estimators when “inverse probability” weights are highly variable. Stat Sci. 2007;22(4):544-559.
18. Shinozaki T, Matsuyama Y. Doubly robust estimation of standardized risk difference and ratio in the exposed population. Epidemiology. 2015;26:873-877.
19. Pohar Perme M, Stare J, Estève J. On estimation in relative survival. Biometrics. 2012;68:113-120. https://doi.org/10.1111/j.1541-0420.2011.01640.x
20. Pavlic K, Pohar PM. Using pseudo-observations for estimation in relative survival. Biostatistics. 2019;20(3):384-399. https://doi.org/10.1093/biostatistics/kxy008
21. Lambert PC, Dickman PW, Rutherford MJ. Comparison of approaches to estimating age-standardized net survival. BMC Med Res Methodol. 2015;15:64.
22. Blakely T, Soeberg M, Carter K, Costilla R, Atkinson J, Sarfaty D. Bias in relative survival methods when using incorrect life-tables: lung and bladder cancer by smoking status and ethnicity in New Zealand. Int J Cancer. 2012;131(6):E974-E982.
23. Bower H, Andersson TM-L, Crowther MJ, Dickman PW, Lambe M, Lambert PC. Adjusting expected mortality rates using information from a control population: an example using socioeconomic status. *Am J Epidemiol*. 2017;187(4):828-836.

24. Ellis L, Coleman MP, Rachet B. The impact of life tables adjusted for smoking on the socio-economic difference in net survival for laryngeal and lung cancer. *British J Cancer*. 2014;111(1):195-202. https://doi.org/10.1038/bjc.2014.217

25. Rubio FJ, Rachet B, Giorgi R, Maringe C, Belot A. On models for the estimation of the excess mortality hazard in case of insufficiently stratified life tables. *Biostatistics*. 2021;22(1):51-67. https://doi.org/10.1093/biostatistics/kxxz017

26. Young JG, Stensrud MJ, Tchetgen Tchetgen EJ, Hernán MA. A causal framework for classical statistical estimands in failure-time settings with competing events. *Stat Med*. 2020;39:1199-1236. https://doi.org/10.1002/sim.8471

27. Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. *Stat Med*. 2004;23(1):51-64.

28. Perme MP, Henderson R, Stare J. An approach to estimation in relative survival regression. *Biostatistics*. 2009;10(1):136-146.

29. Cortese G, Scheike TH. Dynamic regression hazards models for relative survival. *Stat Med*. 2008;27(18):3563-3584.

30. Sasieni PD. Proportional excess hazards. *Biometrika*. 1996;83:127-141.

31. Nelson CP, Lambert PC, Squire IB, Jones DR. Flexible parametric models for relative survival, with application in coronary heart disease. *Stat Med*. 2007;26(30):5486-5498.

32. Sjölander A. Regression standardization with the R package stdReg. *Eur J Epidemiol*. 2016;31(6):563-574.

33. Robins JM, Hernán MA, Brumback B. Marginal structural models as an tool for standardization. *Epidemiology*. 2000;11(5):550-560. https://doi.org/10.1097/00001648-200009000-00011

34. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168(6):656-664. https://doi.org/10.1093/aje/kwn164

35. Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Stat Med*. 2016;35(30):5642-5655. https://doi.org/10.1002/sim.7084

36. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550-560. https://doi.org/10.1097/00001648-200009000-00011

37. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. *Stat Med*. 2013;32:4118-4134. https://doi.org/10.1002/sim.5823

38. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology*. 2003;14(6):680-686. https://doi.org/10.1097/01.ede.000081989.82616.7d

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Syriopoulou E, Rutherford MJ, Lambert PC. Inverse probability weighting and doubly robust standardization in the relative survival framework. *Statistics in Medicine*. 2021;40(27):6069-6092. https://doi.org/10.1002/sim.9171

**APPENDIX A. STANDARD ERRORS USING THE DELTA METHOD AND M-ESTIMATION**

**A.1 Delta Method**

Let $\hat{\beta}$ denote the parameter estimates obtained after fitting a parametric model and let $V(\hat{\beta})$ denote the variance of the estimates. Standard errors for nonlinear functions of the model parameters and covariates can be obtained by utilizing the delta method. Let $g(\hat{\beta}, \mathbf{W})$ denote a function of the model parameters and a set of covariates $\mathbf{W}$. The variance of this
function can be obtained by using the sandwich formula of the delta method:

$$V(g(\hat{\beta}, W)) = GV(\hat{\beta})G^T$$  \hspace{1cm} (A1)

with $G$ denoting the matrix of partial derivatives of the function with respect to $\hat{\beta}$:

$$G = \left. \frac{\partial g(\hat{\beta}, W)}{\partial \beta} \right|_{\beta = \hat{\beta}}. \hspace{1cm} (A2)$$

Consider, for instance, a survival model with $W$ denoting the full design matrix consisting of the baseline spline variables evaluated at time $t$, exposure, $X$, and covariates, $Z$. Assume that the function of interest is the marginal survival function, for a population of $N$ individuals. This is estimated by using regression standardization as:

$$E\left[\hat{S}(t|W)\right] = \frac{1}{N} \sum_{i=1}^{N} \hat{S}(t|W = w_i, \hat{\beta}).$$

Then, the variance of nonlinear functions of the model parameters of this model can be obtained using Equation (A1), with

$$G = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \hat{S}(t|W = w_i, \hat{\beta})}{\partial \beta}.$$

In the case that survival is modeled using a FPM, and by applying the usual transformation from the hazard function to the survival function:

$$S(t|W = w_i, \hat{\beta}) = \exp\left[-\exp\left(\eta(t|W = w_i, \hat{\beta})\right)\right]$$

with

$$\eta(t|W = w_i, \hat{\beta}) = W\hat{\beta}^T$$

and the partial derivative is:

$$\frac{\partial \hat{S}(t|W = w_i, \hat{\beta})}{\partial \hat{\beta}} = -\exp\left[-\exp\left(\eta(t|W = w_i, \hat{\beta})\right)\right] \exp\left(\eta(t|W = w_i, \hat{\beta})\right) W$$

$$= -S(t|W = w_i, \hat{\beta}) \ln\left(S(t|W = w_i, \hat{\beta})\right) W.$$

For matrix $G$, an average of the derivatives is taken over observations.

If interest is on only one standardized survival curve, for example, standardized survival of the exposed group, then $G$ is a $1 \times p$ matrix, where $p$ is the number of model parameters. More than one standardized survival curve can be estimated simultaneously through stacking rows. The advantage of this approach is that the variance-covariance matrix of the standardized survival functions is estimated and this enables to obtain this information for contrasts, such as differences, or more general functions of different standardized survival curves.

**A.2 M-Estimation**

Another method for obtaining the standard errors is the M-estimation approach, also known as estimating equations method, which is a generalized form of the delta method and takes into consideration the observed variation in the confounder distribution of a population.²⁷,²⁸

An estimator $\hat{\theta}$ is an M-estimator if it solves the vector equation:

$$\sum_{i=1}^{N} \phi(Y_i, \hat{\theta}) = 0,$$

where $\phi()$ is a $p \times 1$ function if there are $p$ parameters to be estimated.
Assume that interest is on the standard errors of the standardized survival function of the exposed and the standardized relative survival function of the unexposed, after fitting a FPM. Let \( W \) be the full design matrix consisting of the baseline spline variables evaluated at time \( t \), exposure, \( X \), and confounders, \( Z \).

Let \( v = [\theta_0(t), \theta_1(t), \beta] \), with

\[
\theta_0(t) = \frac{1}{N} \sum_{i=1}^{N} S(t | W = w_i, \hat{\beta}) \quad \text{for fixed values} \ X = 0,\\
\theta_1(t) = \frac{1}{N} \sum_{i=1}^{N} S(t | W = w_i, \hat{\beta}) \quad \text{for fixed values} \ X = 1,
\]

and \( \beta \) the model coefficients. The estimator \( \hat{v} = (\hat{\theta}_0(t), \hat{\theta}_1(t), \hat{\beta}) \) is an M-estimator if it satisfies the estimating equation:

\[
\sum_{i=1}^{N} U_{\nu,i} = \sum_{i=1}^{N} \begin{bmatrix} U_{\theta_0,i}(\beta, \theta_0(t)) \\ U_{\theta_1,i}(\beta, \theta_1(t)) \\ U_{\beta,i}(\beta) \end{bmatrix} = 0
\]

with the estimating equations of individuals \( i \) being defined as

\[
U_{\theta_0,i}(\beta, \theta_0(t)) = S(t | W = w_i, \hat{\beta}) - \theta_0(t) \quad \text{for fixed values} \ X = 0,\\
U_{\theta_1,i}(\beta, \theta_1(t)) = S(t | W = w_i, \hat{\beta}) - \theta_1(t) \quad \text{for fixed values} \ X = 1,\\
U_{\beta,i}(\beta) = \frac{dL_i(t_i, d_i, W_i, \beta)}{d\beta}.
\]

Following standard theory of M-estimators, \( n^{\frac{1}{2}}(\hat{v} - v) \) is asymptotically normal with mean 0 and variance given by the sandwich formula:

\[
\Sigma = ABA'
\]

with

\[
A = E \left[ \frac{\partial U_{\nu,i}}{\partial v} \right]^{-1} \quad \text{and} \quad B = \text{var} [U_{\nu,i}].
\]

In practice, a consistent estimate of the variance of \( \hat{v} \), is obtained by replacing estimates of \( v \) with \( \hat{v} \). Construction of \( A \) and \( B \) matrices is as follows. For matrix \( A \):

\[
A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},
\]

\( A_{11} \) is the derivative of \( U_{\theta_0,i} \) and \( U_{\theta_1,i} \) with respect to \( \theta_0(t) \) and \( \theta_1(t) \) and gives

\[
A_{11} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.
\]

\( A_{12} \) is the derivative of \( U_{\theta_0,i} \) and \( U_{\theta_1,i} \) with respect to \( \beta \) and is the same as the \( G \) matrix described in the delta method. \( A_{21} \) is the derivative of \( U_{\beta,i}(\beta) \) with respect to \( \theta_0(t) \) and \( \theta_1(t) \) and thus is a matrix of zeros. \( A_{22} \) is the derivative of \( U_{\beta,i}(\beta) \) with respect to \( \beta \) which is the Hessian matrix. This has already been estimated when fitting the model and can be derived from the variance matrix. Finally, matrix \( B \) is the variance matrix of the estimating equations for \( \theta_0(t) \), \( \theta_1(t) \), and \( \beta \) (ie, the score equations).