

RESEARCH PAPER

Understanding disparities in cancer prognosis: An extension of mediation analysis to the relative survival framework

Elisavet Syriopoulou¹  | Mark J. Rutherford¹ | Paul C. Lambert^{1,2}

¹ Biostatistics Research Group,
Department of Health Sciences,
University of Leicester, Leicester, UK

² Department of Medical Epidemiology
and Biostatistics, Karolinska Institutet,
Stockholm, Sweden

Correspondence

Elisavet Syriopoulou, Biostatistics
Research Group, George Davies Center
University Road, University of Leicester,
Leicester LE1 7RH, UK
Email: elisavet.syriopoulou@ki.se

Funding information

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



This article has earned an open data badge
“**Reproducible Research**” for mak-
ing publicly available the code necessary
to reproduce the reported results. The
results reported in this article could fully be
reproduced.

Abstract

Mediation analysis can be applied to investigate the effect of a third variable on the pathway between an exposure and the outcome. Such applications include investigating the determinants that drive differences in cancer survival across subgroups. However, cancer disparities may be the result of complex mechanisms that involve both cancer-related and other-cause mortality differences making it difficult to identify the causing factors. Relative survival, a commonly used measure in cancer epidemiology, can be used to focus on cancer-related differences. We extended mediation analysis to the relative survival framework for exploring cancer inequalities. The marginal effects were obtained using regression standardization, after fitting a relative survival model. Contrasts of interests included both marginal relative survival and marginal all-cause survival differences between exposure groups. Such contrasts include the indirect effect due to a mediator that is identifiable under certain assumptions. A separate model was fitted for the mediator and uncertainty was estimated using parametric bootstrapping. The avoidable deaths under interventions can also be estimated to quantify the impact of eliminating differences. The methods are illustrated using data for individuals diagnosed with colon cancer. Mediation analysis within relative survival allows focus on factors that account for cancer-related differences instead of all-cause differences and helps improve our understanding on cancer inequalities.

KEYWORDS

cancer inequalities, mediation analysis, natural indirect effect, regression standardization, relative survival

1 | INTRODUCTION

Survival after a cancer diagnosis varies considerably across subgroups. For instance, many studies have reported large disparities between socioeconomic groups that exist irrespective of the various approaches of defining socioeconomic groups (Danø, Andersen, Ewertz, Petersen, & Lynge, 2003; Ito et al., 2014; Jeffreys et al., 2009; Rachet et al., 2010; Rutherford, Andersson, Møller, & Lambert, 2015; Syriopoulou et al., 2017). Understanding the factors that drive survival

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

differences can be challenging due to complex mechanisms that contribute towards disparities and the methodological challenges they induce.

Mediation analysis provides a useful tool for such settings as it can be applied to explore the role of a third variable (a mediator) that may be on the pathway between an exposure and the outcome (De Stavola, Daniel, Ploubidis, & Micali, 2014; Imai, Keele, & Tingley, 2010; VanderWeele, 2011). Thus, we can investigate whether differences in the mediator distribution are partly responsible for the variation between exposure groups. Mediation analysis helps to explore potential causal mechanisms of an observed association through an effect decomposition and under certain assumptions it allows the identification of the direct effect between an exposure and an outcome and the indirect effect due to the mediator (Pearl, 2001).

Another challenge with population-based cancer data is the presence of competing events. The event of interest, which is usually death due to cancer, will never be observed for some patients due to an earlier death due to other causes. Cause of death information is often unreliable preventing a cause-specific approach. The relative survival approach can be applied instead as it utilizes the expected mortality rates of a comparable group in the general population to represent mortality due to other causes for the cancer population (Dickman & Coviello, 2015; Ederer, Axtell, & Cutler, 1961; Pohar Perme, Stare, & Estève, 2012). Under assumptions, relative survival is interpreted as net survival, that is the probability of survival in a hypothetical world where the cancer of interest is the only possible cause of death (Pavlic & Pohar Perme, 2019).

Relative survival is a useful measure for disentangling the exposure–outcome association as it enables us to focus on cancer-related differences instead of all-cause differences. All-cause survival differences are the result of complicated mechanisms that involve both cancer-related and other cause factors. Thus, it may be easier to investigate the underlying determinants of cancer-related differences and identify interventions of eliminating them. Such interventions would reduce the cancer mortality rates of those with worse prognosis but would have no impact on other cause mortality rates.

We extend mediation analysis to the relative survival framework and utilize regression standardization to obtain the marginal survival and related functions. We assess how much of the differences between exposure groups can be explained by differences in the mediator distribution, and we estimate the impact of removing differences in relative survival and the mediator distribution. For the interpretation of these measures as causal, standard mediation analysis assumptions are now extended in the relative survival framework and they need to hold both for cancer and other cause mortality.

The paper is structured as follows. First, we introduce the illustrative data and we briefly describe the main measure of interest for comparing subgroups, that is the marginal relative survival that is estimated by the standardized relative survival. Then, we investigate the role of a mediator in the exposure–outcome association and explore the impact of interventions that aim to eliminate cancer-related survival differences. Such interventions can also be quantified using the avoidable deaths measure. Finally, we conclude the paper by summarizing the methods and discussing potential limitations.

2 | INTRODUCING THE ILLUSTRATIVE EXAMPLE

We demonstrate the methods using an example on survival differences between socioeconomic groups of individuals diagnosed with colon cancer. Data include all individuals diagnosed with colon cancer in England between 2011 and 2013, made available by Public Health England. Available information includes: sex, age and stage 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 of the area of patients' residence at diagnosis (Department for Communities and Local Government, 2010; Neighbourhood Renewal Unit, 2004). Even though deprivation status is a categorical variable with five categories, as the main purpose of the analysis was to demonstrate the measures, we utilize a subset of the population, that is the least and most deprived groups. In England, completeness of stage at diagnosis has improved dramatically after 2012; however, there are a large proportion of missing data for earlier years. As a result, stage at diagnosis was missing for 33.9% of the population. Once again as these data are used only for illustration of the methods, we conducted a complete case analysis including only those with recorded stage at diagnosis. However, our approach can extend to multiple imputation approaches for missing data. The final data included 15,630 patients, 57.6% of which were in the least deprived group. More details on the study population can be found in Table 1.

3 | MARGINAL RELATIVE SURVIVAL

Let X be a binary variable with $X = 1$ for the exposed and $X = 0$ for the unexposed. The conditional all-cause mortality rate of an individual i , with exposure value x_i and confounder pattern \mathbf{z}_i , can be partitioned into the conditional expected

TABLE 1 Number of colon cancer patients (with proportions) for sex, age-groups and stage at diagnosis by deprivation group

	Deprivation group	
	Least deprived	Most deprived
Sex		
Males	4841 (53.78%)	3500 (52.81%)
Females	4161 (46.22%)	3128 (47.19%)
Age group		
18–44	233 (2.59%)	298 (4.50%)
45–54	537 (5.97%)	470 (7.09%)
55–64	1544 (17.15%)	1267 (19.12%)
65–74	2767 (30.74%)	1877 (28.32%)
75–84	2820 (31.33%)	1970 (29.72%)
85+	1101 (12.22%)	746 (11.25%)
Stage at diagnosis ^a		
I	1338 (14.86%)	912 (13.76%)
II	2644 (29.37%)	1950 (29.42%)
III	2435 (27.05%)	1716 (25.89%)
IV	2585 (28.72%)	2050 (30.93%)

^aStage I-IV for the least to the most advanced stage at diagnosis.

mortality rate had they not had the cancer, $h^*(t|X = x_i, \mathbf{Z}_1 = \mathbf{z}_{1i})$, and the conditional excess mortality rate due to cancer, $\lambda(t|X = x_i, \mathbf{Z}_2 = \mathbf{z}_{2i})$:

$$h(t|X = x_i, \mathbf{Z} = \mathbf{z}_i) = h^*(t|X = x_i, \mathbf{Z}_1 = \mathbf{z}_{1i}) + \lambda(t|X = x_i, \mathbf{Z}_2 = \mathbf{z}_{2i}).$$

\mathbf{Z} denotes the set of all confounders and consists of subsets \mathbf{Z}_1 and \mathbf{Z}_2 . Subset \mathbf{Z}_1 denotes the confounders for expected mortality and subset \mathbf{Z}_2 denotes the confounders for the excess mortality. Often \mathbf{Z}_1 will be a subset of \mathbf{Z}_2 and in that case \mathbf{Z}_2 will be the equivalent to \mathbf{Z} . For instance, sex and age affect other cause mortality but they also affect cancer mortality.

The survival analogue of excess mortality is relative survival. The conditional relative survival, $R(t|X = x_i, \mathbf{Z}_2 = \mathbf{z}_{2i})$, is a function of the conditional all-cause survival, $S(t|X = x_i, \mathbf{Z} = \mathbf{z}_i)$, and the conditional expected survival, $S^*(t|X = x_i, \mathbf{Z}_1 = \mathbf{z}_{1i})$. The conditional all-cause survival can be written as

$$S(t|X = x_i, \mathbf{Z} = \mathbf{z}_i) = S^*(t|X = x_i, \mathbf{Z}_1 = \mathbf{z}_{1i})R(t|X = x_i, \mathbf{Z}_2 = \mathbf{z}_{2i}). \quad (1)$$

It is important to point out that there is variation in expected, relative and observed survival between individuals. The expected survival (as well as the expected mortality rates) is typically obtained by available lifetables for the general population that are stratified by confounders \mathbf{Z}_1 , such as age, sex, calendar year and deprivation status. In a modelling context, relative survival is obtained from a relative survival model in which only confounders \mathbf{Z}_2 are included in the model. However, the expected survival probabilities of individuals will also be incorporated in this model and these are obtained by the population lifetables which are stratified by confounders \mathbf{Z}_1 .

Relative survival can be interpreted as net survival, that is survival in a hypothetical world where the only possible cause of death is the cancer of interest under the following assumptions: (i) the expected mortality rates that represent mortality due to other causes for the cancer population are appropriate and (ii) the potential times to death from cancer and other causes are conditionally independent (Pavlic & Pohar Perme, 2019). For the first assumption to hold, it is important to include sufficient variables, \mathbf{Z}_1 , in the lifetable to ensure comparability between the cancer and the general populations (Pohar Perme et al., 2012; Dickman & Coviello, 2015). The second assumption is the same assumption required in a cause-specific approach.

Many modelling approaches exist to estimate relative survival and our methods in principle could be applied to all, but in this paper we focus on flexible parametric survival models that use restricted cubic splines to model the baseline log cumulative excess hazard (Cortese & Scheike, 2008; Dickman, Sloggett, Hills, & Hakulinen, 2004; Estève, Benhamou, Croasdale, & Raymond, 1990; Hakulinen & Tenkanen, 1987; Nelson, Lambert, Squire, & Jones, 2007; Royston,

2001). Non-proportional excess hazards and interactions can easily be incorporated in the model. After fitting the model, a summary of the population prognosis can be obtained as marginal effects such as marginal relative survival functions. We focus on counterfactual marginal relative survival functions within subgroups of the population that are obtained through regression standardization (Sjölander, 2016; Syriopoulou, Rutherford, & Lambert, 2020). Alternative methods to estimate marginal measures like inverse probability weights could also be applied (Cole & Hernán, 2004).

The counterfactual marginal relative survival when setting $X = x$ is defined as

$$\theta(t|X = x) = E[R(t|X = x, \mathbf{Z}_2)],$$

with the expectation taken over \mathbf{Z}_2 . To relate the observed and counterfactual outcomes, the assumptions of conditional exchangeability, consistency and positivity need to hold (Hernán & Robins, 2006), both for cancer and other cause mortality. These are discussed in more detail in Syriopoulou et al. (2020). Under these assumptions, the counterfactual marginal relative survival can be estimated as the standardized relative survival. This is given as an average of the predictions of each individual in a study population with N individuals:

$$\hat{\theta}(t|X = x) = \frac{1}{N} \sum_{i=1}^N \hat{R}(t|X = x, \mathbf{Z}_2 = \mathbf{z}_{2i}).$$

For the estimation of the counterfactual marginal relative survival, the exposure is set to $X = x$ for every individual in the study population. This is different to utilizing the observed exposure value of individuals, $X = x_i$, as in Equation (1).

On the mortality scale, the marginal net probability of death can be estimated instead by $1 - \hat{\theta}(t|X = x)$.

The difference in counterfactual marginal relative survival functions between two different levels of the exposure can also be defined:

$$\theta(t|X = 1) - \theta(t|X = 0) = E[R(t|X = 1, \mathbf{Z}_2)] - E[R(t|X = 0, \mathbf{Z}_2)]. \quad (2)$$

This is a comparison of two hypothetical situations: in the first term X is set to 1 for everyone, and in the second term X is set to 0 for everyone. Under the identifiability assumptions, difference (2) can be estimated by the difference in standardized relative survival functions

$$\frac{1}{N} \sum_{i=1}^N \hat{R}(t|X = 1, \mathbf{Z}_2 = \mathbf{z}_{2i}) - \frac{1}{N} \sum_{i=1}^N \hat{R}(t|X = 0, \mathbf{Z}_2 = \mathbf{z}_{2i}).$$

The relative survival difference is a causal effect in the absence of competing risks (Young, Stensrud, Tchetgen Tchetgen, & Hernán, 2020). Difference (2) refers to the cancer-related difference if everyone was exposed versus if everyone was unexposed, in a hypothetical world where the only possible cause of death is the cancer of interest (Lambert, Dickman, & Rutherford, 2015; Pavlic & Pohar Perme, 2019). All-cause survival differences, in a setting where other causes of death are present, can also be obtained by incorporating the expected survival:

$$E[S^*(t|X = 1, \mathbf{Z}_1)R(t|X = 1, \mathbf{Z}_2)] - E[S^*(t|X = 0, \mathbf{Z}_1)R(t|X = 0, \mathbf{Z}_2)] \quad (3)$$

and is estimated by

$$\begin{aligned} & \frac{1}{N} \sum_{i=1}^N S^*(t|X = 1, \mathbf{Z}_1 = \mathbf{z}_{1i}) \hat{R}(t|X = 1, \mathbf{Z}_2 = \mathbf{z}_{2i}), \\ & - \frac{1}{N} \sum_{i=1}^N S^*(t|X = 0, \mathbf{Z}_1 = \mathbf{z}_{1i}) \hat{R}(t|X = 0, \mathbf{Z}_2 = \mathbf{z}_{2i}). \end{aligned}$$

In expression (3), survival differences may be due to differences in cancer mortality, other cause mortality or both. More details on marginal estimates and causal effects using relative survival can be found elsewhere (Syriopoulou et al., 2020).

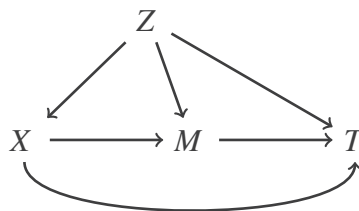


FIGURE 1 Directed acyclic graph for the relationship of the exposure X , time to a specific event T and confounding Z in the presence of a mediator M

4 | EXPLORING THE EFFECT OF A MEDIATOR

Let us assume that we are interested in the potential role of a third variable M , as a mediator, in the association of exposure X and a time-to-event outcome T . For instance, Figure 1 demonstrates a setting in which X has both a direct ($X \rightarrow T$) and an indirect ($X \rightarrow M \rightarrow T$) effect through M to T (Pearl, 2012). For simplicity, we have assumed the same set of confounders, Z , for $X - M$ and $M - T$, but this can be generalized to include a different set of confounders for one of them. The setting of Figure 1 can be extended to the relative survival framework. However, within the relative survival framework, we only utilize the information on time to death without knowing the exact cause of death. By extending mediation analysis to relative survival and under certain assumptions, identification of the natural direct and indirect effects is possible. The mediators that will be considered here are assumed to affect only cancer mortality rates and have no effect on other cause mortality rates.

Let M^x denote the counterfactual mediator distribution when intervening to set the exposure X to level x . Let $R(t|X = y, Z_2, M^x)$ be the counterfactual relative survival function when intervening to set the exposure X to level y and the mediator M to M^x . Level y is set to values 0 and 1 for the unexposed and exposed, respectively. Sometimes it is possible to have $x = y$, that is the relative survival that would have been observed if both the exposure and the counterfactual mediator distribution have been set to the same exposure level.

Within the relative survival framework, the natural direct effect (NDE_{RS}) is defined as the difference in marginal relative survival between the exposed and unexposed if both groups had the same mediator distribution as the unexposed (setting M to M^0 and thus M remains the same for each patient in both terms):

$$NDE_{RS} = E[R(t|X = 1, Z_2, M^0)] - E[R(t|X = 0, Z_2, M^0)]. \quad (4)$$

The natural indirect effect (NIE_{RS}), which gives the effect of the mediator, is defined as the difference when setting $X = 1$ and comparing the effects of having their own mediator distribution (setting M to M^1) versus if they had the same mediator distribution as the unexposed (setting M to M^0):

$$NIE_{RS} = E[R(t|X = 1, Z_2, M^1)] - E[R(t|X = 1, Z_2, M^0)]. \quad (5)$$

In (5), X is allowed to influence relative survival only through its influence on M .

Identification of NDE_{RS} and NIE_{RS} is possible under standard mediation analysis assumptions that are now extended to both outcomes, that is cancer and other causes (De Stavola, et al., 2014; Pearl, 2001). For the rest of this paragraph, referring to the outcome will imply both cancer and other causes. First, no interference assumption states that a patient's exposure has no effect on the outcome of another (both cancer and other cause) and that a patient's mediator value does not influence the outcome of another patient. In addition, an individual's exposure has no effect on the mediator of another individual. Second, consistency states that an individual's outcome under the actual values of $X = x$ and $M = m$ is equal to the outcome that would be observed under an intervention of setting the exposure $X = x$ and the mediator to $M = m$. Consistency also expands so that (i) the outcome under the actual value of $X = x$ is equal to the outcome that would be observed under an intervention of setting $X = x$ and $M = M^x$ as well as (ii) $M^x = M$ when the actual value is $X = x$. Finally, conditional exchangeability states that there is (i) no unmeasured exposure–outcome confounding conditionally on confounders, (ii) no unmeasured mediator–outcome confounding conditionally on exposure and confounders, (iii) no unmeasured exposure–mediator confounding conditional on confounders and (iv) no mediator–outcome confounder affected by exposure. Achieving conditional exchangeability for other cause mortality depends on the level of stratification

in the lifetables that are used to incorporate expected mortality rates. If the variables of the lifetables are insufficient then this assumption is violated. Various methods have been suggested to incorporate factors that are not available on population level (Bower et al., 2017; Ellis, Coleman, & Rachet, 2014; Rubio, Rachet, Giorgi, Maringe, & Belot, 2019).

The NDE_{RS} and NIE_{RS} can be estimated by (Pearl, 2001; Pearl, 2012)

$$\begin{aligned}\widehat{NDE}_{RS} &= \frac{1}{N} \sum_{i=1}^N \sum_m \hat{R}(t|X = 1, \mathbf{Z}_2 = \mathbf{z}_{2i}, M = m) \hat{P}(M = m|X = 0, \mathbf{Z}_2 = \mathbf{z}_{2i}) \\ &\quad - \frac{1}{N} \sum_{i=1}^N \sum_m \hat{R}(t|X = 0, \mathbf{Z}_2 = \mathbf{z}_{2i}, M = m) \hat{P}(M = m|X = 0, \mathbf{Z}_2 = \mathbf{z}_{2i}) \\ \widehat{NIE}_{RS} &= \frac{1}{N} \sum_{i=1}^N \sum_m \hat{R}(t|X = 1, \mathbf{Z}_2 = \mathbf{z}_{2i}, M = m) \hat{P}(M = m|X = 1, \mathbf{Z}_2 = \mathbf{z}_{2i}) \\ &\quad - \frac{1}{N} \sum_{i=1}^N \sum_m \hat{R}(t|X = 1, \mathbf{Z}_2 = \mathbf{z}_{2i}, M = m) \hat{P}(M = m|X = 0, \mathbf{Z}_2 = \mathbf{z}_{2i})\end{aligned}$$

with m taking values 0 and 1 for a binary mediator M . For a mediator with more levels, the summation is taken over all levels. Also, $\hat{P}(M = m|X = x, \mathbf{Z}_2 = \mathbf{z}_{2i})$ is the estimated probability of being in a specific level of the mediator given exposure and confounders.

Detailed steps on the estimation can be found in Box 1. To account for the uncertainty on the probabilities estimated in Step 3 and the survival functions of Step 4, bootstrap-based standard errors are obtained. By performing parametric bootstrap, the parameters are drawn repeatedly from a multivariate normal distribution and for each draw we obtain both estimates and the variance covariance matrix that are finally combined (Efron & Tibshirani, 1993). An example of Stata code for obtaining predictions can be found in the Supporting Information.

The proportion of the total causal effect (TCE_{RS}) that is due to the mediator within the net survival setting is then defined as

$$PM_{RS} = \frac{NIE_{RS}}{TCE_{RS}},$$

with the $TCE_{RS} = NDE_{RS} + NIE_{RS}$.

The NDE_{RS} and NIE_{RS} refer to differences in relative survival between exposure groups in a net-survival setting where the only possible cause of death is the cancer of interest and therefore yield differences that exist only due to the cancer of interest and not other causes.

Instead of focusing on the net survival setting as in (4) and (5), it is also possible to obtain estimates in a situation where both cancer and other causes are present. This can be done by incorporating the expected survival and forming contrasts of all-cause survival. There are many ways to incorporate the expected survival. For instance, the following contrasts can be formed:

$$NDE_{AC1} = E[S^*(t|X = 1, \mathbf{Z}_1)R(t|X = 1, \mathbf{Z}_2, M^0)] - E[S^*(t|X = 0, \mathbf{Z}_1)R(t|X = 0, \mathbf{Z}_2, M^0)]$$

$$NIE_{AC1} = E[S^*(t|X = 1, \mathbf{Z}_1)R(t|X = 1, \mathbf{Z}_2, M^1)] - E[S^*(t|X = 1, \mathbf{Z}_1)R(t|X = 1, \mathbf{Z}_2, M^0)].$$

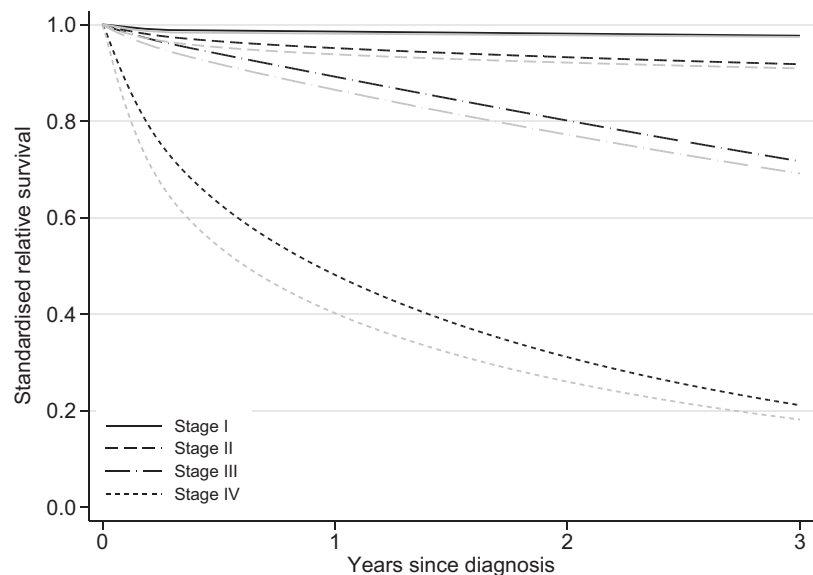
For the NDE_{AC1} , the first term includes the expected survival if exposure was set to 1 for everyone, $S^*(t|X = 1, \mathbf{Z}_1)$, and the second term includes the expected survival by setting $X = 0, S^*(t|X = 0, \mathbf{Z}_1)$. Therefore, for NDE_{AC1} , survival differences may be due to differential cancer mortality, other cause mortality or both. Understanding the mechanisms of all-cause survival differences can be complex. Thus, it might be of interest to obtain the following measures instead:

$$NDE_{AC2} = E[S^*(t|X, \mathbf{Z}_1)R(t|X = 1, \mathbf{Z}_2, M^0)] - E[S^*(t|X, \mathbf{Z}_1)R(t|X = 0, \mathbf{Z}_2, M^0)]$$

$$NIE_{AC2} = E[S^*(t|X, \mathbf{Z}_1)R(t|X = 1, \mathbf{Z}_2, M^1)] - E[S^*(t|X, \mathbf{Z}_1)R(t|X = 1, \mathbf{Z}_2, M^0)].$$

For NDE_{AC2} and NIE_{AC2} , survival differences can only be due to the cancer of interest as the expected survival, $S^*(t|X, \mathbf{Z}_1)$, is incorporated using the observed distribution of the exposure for both contrasting terms. This is the key

FIGURE 2 Standardized estimates of relative survival by years since diagnosis and stage at diagnosis. Black and grey lines refer to the relative survival of least and most deprived patients, respectively



difference with NDE_{AC1} to which the expected survival was incorporated by setting $X = 1$ or $X = 0$ for everyone in the study population.

The direct and the indirect effects can also be obtained within subsets of the whole population. For instance, the NDE among the exposed could be estimated by standardizing only to patients of the exposed group, $N_{X=1}$. Such contrasts can be useful for assessing the potential impact of interventions that aim to eliminate differences between groups by either focusing on the mediator distribution or the cancer-related survival (i.e. relative survival).

It is important to note that the interventions considered in this paper are assumed to have no impact on other cause mortality rates (i.e. other cause mortality remains unchanged after such interventions). The two competing events are assumed to be conditionally independent. Even when referring to an all-cause setting, it is assumed that any potential impact of the intervention on survival is due to changes in cancer mortality. For NDE_{AC1} , some of the survival differences will be due to differences in other cause mortality but this would be a result of differential background mortality between exposed and unexposed and not a result of the intervention. The motivation for this distinction is that the effect of certain interventions would be separable across the two competing causes; for instance, increased cancer screening engagement would likely impact only on the mortality for the cancer cause specifically, whilst only indirectly impacting on other-cause death probabilities but not the underlying other-cause mortality rates.

4.1 | Example

We explored survival differences following a diagnosis of colon cancer between socioeconomic groups by fitting a flexible parametric survival model with 5 degrees of freedom for the baseline excess hazard including sex, deprivation status, age and stage at diagnosis and allowing for time-dependent effects for deprivation, age and stage (3 degrees of freedom). Age at diagnosis was included in the model as a continuous non-linear variable using restricted cubic splines with 3 degrees of freedom. An interaction between stage and deprivation was also allowed. A multinomial regression model was fitted for stage including age as a continuous non-linear variable, deprivation status and sex. The 95% confidence intervals were obtained using the standard deviation of a parametric bootstrap sample with $k = 500$.

We found differences in standardized relative survival between the least and most deprived groups, especially for the most advanced stages (Figure 2). Differences were also observed in the stage distribution, as a higher proportion of the most deprived were diagnosed in a more advanced stage (Table 1). The role of stage as a potential mediator in the association of deprivation and survival time was further investigated by obtaining the \widehat{NDE}_{RS} and \widehat{NIE}_{RS} that are due to stage specific survival differences and due to the stage differences, respectively, in the net survival setting where the only possible cause of death is colon cancer (Figure 3). Three years after diagnosis a total difference of 3.38% (95% CI: [0.83, 5.93]) was observed in standardized net probabilities of death and 1.25% (95% CI: [0.22, 2.28]) of the difference was attributed to differences in stage at diagnosis. As a result, the proportion of total differences in standardized net probability of death that was

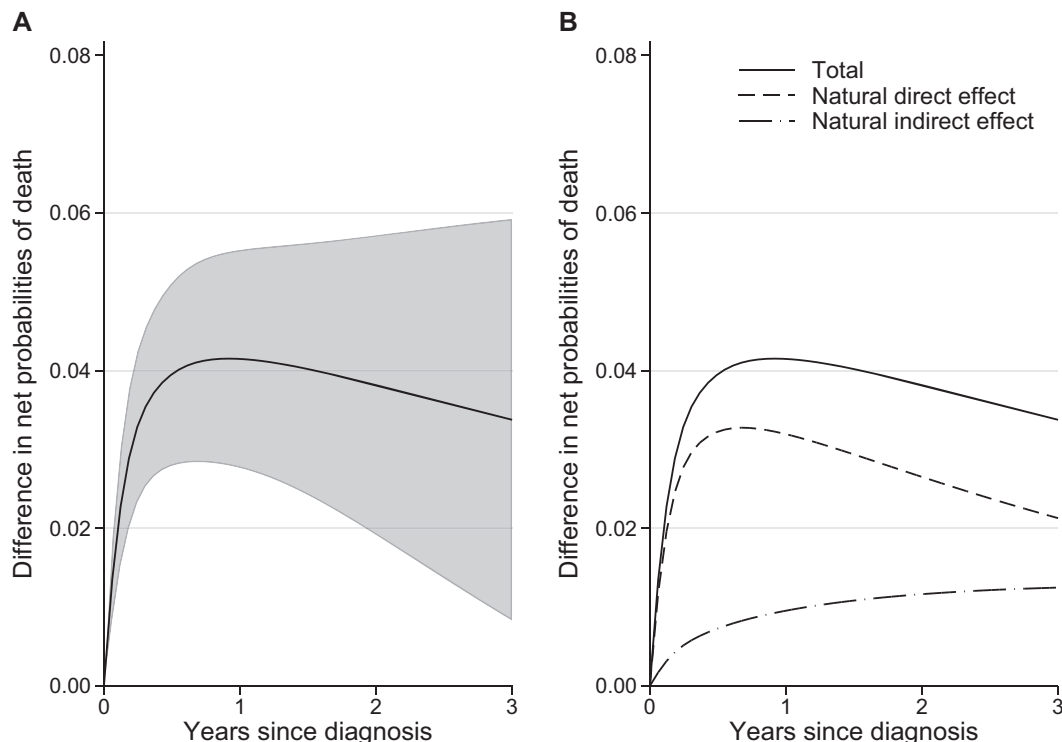


FIGURE 3 (A) Total causal effect, defined as the difference in standardized net probabilities of death, with 95% confidence intervals and (B) partitioning of the total causal effect to the natural direct and indirect effect due to stage at diagnosis

mediated through stage at 3 years was 37% (i.e. $1.25/3.38$). To obtain an estimate of cancer-related differences in the all-cause setting, where other causes of death are present, the \widehat{NDE}_{AC2} and \widehat{NIE}_{AC2} were also estimated (Figure 4). We found that the indirect effect due to stage was 1.14% (95% CI: [0.24, 2.04]) and the total difference in probabilities of death was equal to 3.01% (95% CI: [0.77, 5.26]). Thus, 38% of the difference in this all-cause setting was mediated through stage.

5 | AVOIDABLE DEATHS UNDER HYPOTHETICAL INTERVENTIONS

The impact of eliminating differences between groups, in the presence of both cancer and other cause mortality, can also be quantified as the avoidable deaths after an intervention (Syriopoulou et al., 2020). Such an intervention would be to eliminate differences in the mediator distribution. As mentioned earlier, the mediators considered in this paper are assumed to affect only the cancer mortality rates and have no effect on the rates of other cause mortality.

The avoidable deaths is a time-specific measure and has the interpretation of postponable deaths as eventually all deaths will be realized. Although the avoidable deaths can also be defined for the whole population, here we focus on the avoidable deaths among the exposed, that is obtaining marginal estimates using a subset of the population.

Assume that we are interested in the avoidable deaths under an intervention that aims to eliminate differences in the distribution of the mediator between exposed and unexposed, while keeping other cause mortality rates unchanged. First, we need the number of deaths for the exposed that is given by multiplying the number of exposed patients diagnosed in a typical calendar year, N^* with the probability of death:

$$D_1(t|X = 1, M^1) = N^*(1 - E[S^*(t|X = 1, \mathbf{Z}_1^{X=1})R(t|X = 1, \mathbf{Z}_2^{X=1}, M^1)]), \quad (6)$$

with $\mathbf{Z}_1^{X=1}$ and $\mathbf{Z}_2^{X=1}$ denoting the covariates for the exposed, for the expected and relative survival, respectively.

Then, the expected number of deaths under the intervention can be derived by shifting the mediator distribution of the exposed to the one of the unexposed (setting M to M^0):

$$D_M(t|X = 1, M^0) = N^*(1 - E[S^*(t|X = 1, \mathbf{Z}_1^{X=1})R(t|X = 1, \mathbf{Z}_2^{X=1}, M^0)]). \quad (7)$$

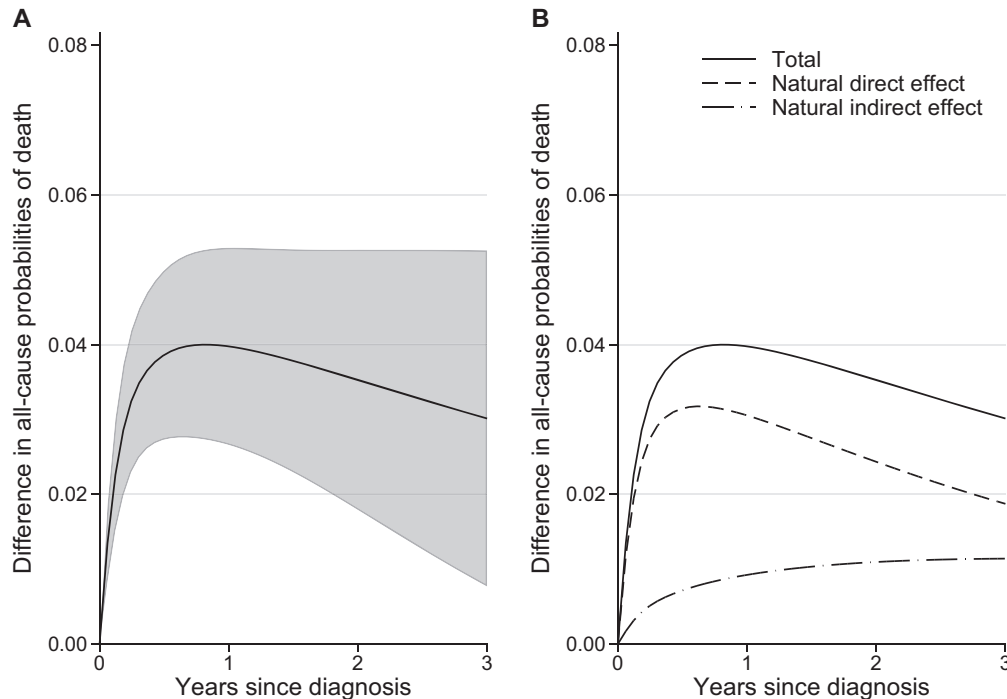


FIGURE 4 (A) Total causal effect, defined as the difference in standardized all cause probabilities of death, with 95% confidence intervals and (B) partitioning of the total causal effect to the natural direct and indirect effect due to stage at diagnosis

The avoidable deaths from eliminating differences in the mediator distribution is given by

$$AD_{RS} = D_1(t|X = 1, M^1) - D_M(t|X = 1, M^0).$$

For the identification of AD_{RS} , assumptions similar to the one described for NDE_{RS} and NIE_{RS} are required.

It is important to note that in the above scenario, we keep the expected survival of the exposed group unchanged and we assess an intervention that aims to shift the mediator distribution of the exposed to that of the unexposed with no impact on other cause mortality rates. As a result, the potential impact of the interventions considered here will only be due to changes in cancer mortality rates.

A key point for the interpretation of the avoidable deaths is the number of patients N^* applied in (6) and (7). N^* can be any number relevant to the study population. For instance, it could be the number of exposed patients diagnosed in the most recent year in our data, or it could be derived by adding all exposed patients diagnosed during the total duration of the follow-up divided by the number of years available. Some might also consider calculating the avoidable deaths per 1000 patients. When interpreting the results, it is important to keep in mind potential differences between the population being marginalized over, which in the above example is all the exposed patients of the study (over a range of calendar years), and the population used for N^* . In extreme cases, the covariate pattern might have changed over calendar time suggesting that a choice must be made over the most relevant information to present. It may be preferable to marginalize over a specific restricted population, with the appropriately calculated N^* for that population.

5.1 | Example

We estimated the avoidable deaths under two hypothetical interventions:

1. 'eliminating differences in the *stage at diagnosis distribution* as well as *relative survival differences* between the least and most deprived groups' (scenario 1).
2. 'eliminating differences in the *stage at diagnosis distribution* between the least and most deprived groups' (scenario 2).

For scenario 1, we shifted the relative survival and stage at diagnosis distribution of the most deprived patients to that of the least deprived, that is the most advantaged group. For scenario 2, we shifted the stage at diagnosis distribution of the

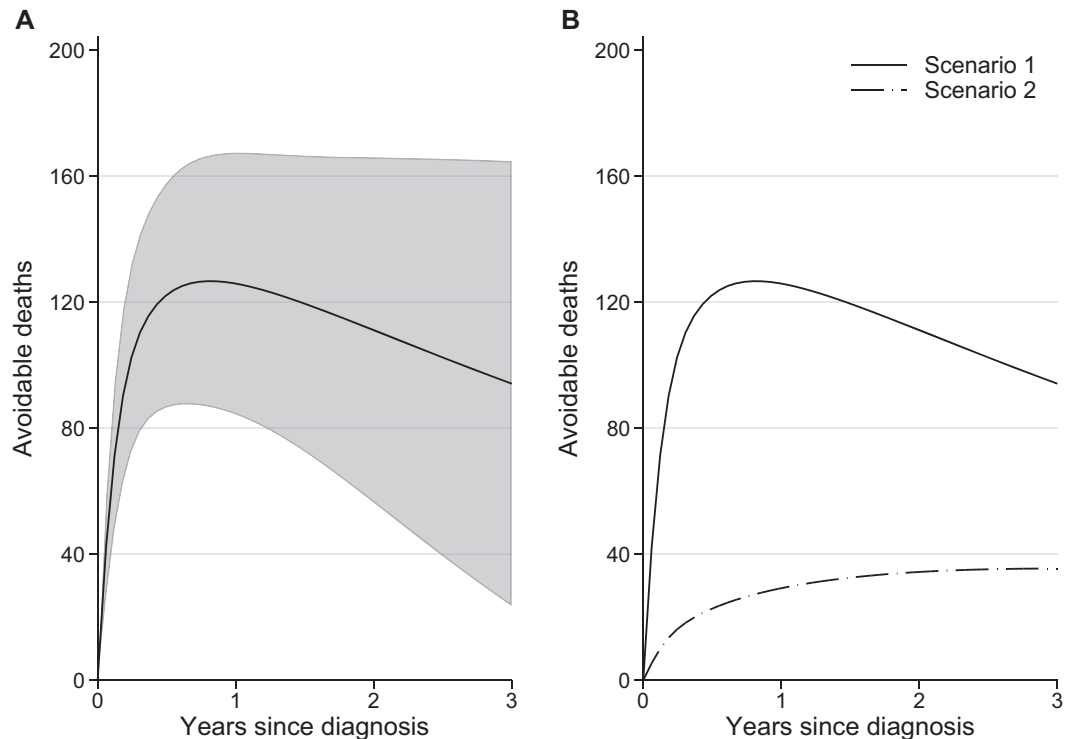


FIGURE 5 (A) Total avoidable deaths by removing relative survival and stage differences between the least and most deprived groups (Scenario 1) with 95% confidence intervals and (B) partitioning total avoidable deaths to those under an intervention of “eliminating differences in the stage at diagnosis distribution” (Scenario 2)

most deprived to that of the least deprived group. For both scenarios, we kept the expected survival of the most deprived unchanged. Three years after diagnosis 94 (95% CI: [24, 165]) avoidable deaths would be observed in total, out of 3228 (i.e. N^*) patients from the most deprived group diagnosed in 2013 the most recent year in our cohort study. Partitioning that further, we found that 35 (95% CI: [7, 64]) deaths of the total deaths would be from eliminating stage differences (scenario 2) and the remaining 59 would be from removing relative survival differences (Figure 5).

6 | DISCUSSION

We extended mediation analysis to the relative survival framework as a tool for investigating differences in cancer prognosis. Using relative survival enables forming contrasts of either relative survival in which the only possible cause of death is the cancer of interest or in all-cause survival in which both the cancer and other causes are present. We showed that, if certain assumptions hold, mediation analysis can be applied to identify the indirect effect due to a mediator both in the net survival and all-cause setting. For further exploration of the survival differences, the potential impact of interventions can be estimated as the number of deaths that could be avoided within a time frame. The interventions considered here focus on cancer-related mortality rates and are assumed to have no impact on other cause mortality rates. Because we are in a competing risks setting, changing the cancer mortality rates will increase the probability of dying from other causes, even if the other cause mortality rates remain unchanged.

To account for the uncertainty in the probability weights and the predictions of standardized survival (Step 3 and Step 4 of Box 1), we obtain bootstrap-based standard errors by performing parametric bootstrap for the parameter estimates of both models. Confidence intervals are obtained using the relevant quantiles of the bootstrap samples estimates distribution or their standard deviation. Alternative approaches could be applied such as M-estimation methods that would shorten the computational time (Stefanski & Boos, 2002).

Identifying interventions of eliminating all-cause differences can be challenging. This is because differences in all-cause survival can be the result of complex mechanisms that involve both cancer-related and other cause differences. In this paper, we utilized relative survival and focus on cancer-related differences. Quantifying differences in a real-world setting

BOX 1: Algorithm for obtaining the natural direct and indirect effects

Step 1. Fit a parametric relative survival model for the time-to-event outcome including the exposure, mediator, potential confounders and appropriate interactions and time-dependent effects.

Step 2. Fit a model for the mediator including the exposure and confounders. For example, for a binary mediator this could be a logistic regression model and for a mediator with more categories this could be a multinomial regression model.

Step 3. For each individual in the study population obtain predictions for the probability of being in a specific level of the mediator, $\hat{P}(M = m|X = x, \mathbf{Z}_2 = \mathbf{z}_{2i})$, at each level of the exposure $X = x$.

Step 4. Obtain predictions of the standardized relative survival functions at each level of $X = x$, as a weighted average of the individual relative survival functions $\hat{R}(t|X = x, \mathbf{Z}_2 = \mathbf{z}_{2i}, M = m)$, using the predictions of Step 3 as weights. Contrasts of these predictions can be formed to obtain the \widehat{NDE}_{RS} and \widehat{NIE}_{RS} .

Step 5. Repeat from Step 3 for k times while performing parametric bootstrap for the parameter estimates for both models.

Step 6. Calculate 95% confidence intervals either by taking the 2.5% and 97.5% quantiles of the \widehat{NDE}_{RS} and \widehat{NIE}_{RS} estimates across the bootstrapped samples or by using the standard deviation of the estimates obtained from the bootstrap samples.

having focused on eliminating cancer-related differences alone is also possible by incorporating the expected survival probabilities. An intervention that aims to eliminate cancer-related differences without intervening on the other-cause mortality might be easier to identify. However, one could argue that our intervention is still not well defined (Hernán & Robins, 2006). Changing the cancer mortality of one exposure group while keeping the others the same might not be straightforward in practice. For instance, an intervention that aims to increase cancer awareness in the most deprived patents will most probably increase awareness also in the least deprived group. If this is the case, our estimates will provide a lower bound of the actual population benefit of the intervention. Nevertheless, quantifying the impact of such a conceptual intervention in a formalized causal framework gives a firm basis to improve our understanding on cancer disparities even if such an intervention is difficult to identify in practice (Glymour & Spiegelman, 2017; Krieger & Davey Smith, 2016; Pearl, 2018; Vandenbroucke, 2016).

Interpretation as causal effects depends on the validity of standard mediation analysis assumptions that are now extended to the relative survival framework and therefore need to hold for both outcomes, cancer and other cause survival times. These are no interference, consistency and conditional exchangeability (De Stavola et al. 2014; Pearl, 2001). Achieving conditional exchangeability for the other cause mortality depends on the availability of relevant lifetables that are used to represent the other cause mortality of the cancer population. Causal interpretation is only appropriate when lifetables are sufficiently stratified, but in principle lifetables can be constructed for any number of factors. To deal with this issue and consider other risk factors that are not always available on a population level, adjustments at the expected mortality rates have been suggested (Bower et al., 2017; Ellis et al., 2014; Rubio et al., 2019). Finally, we have assumed no intermediate confounders, that is no mediator–outcome confounder affected by the exposure (cross-world independence assumption). Methods that do not require the cross-world assumption have been suggested before by either using a weighting-based approach with the limitation of not adding to the total effect or a Monte Carlo–based regression approach that applies also to multiple mediators (VanderWeele, Vansteelandt & Robins, 2014; Vansteelandt & Daniel, 2017). In principle, our methods can be extended to settings with intermediate confounders and this consists part of future work.

Further assumptions that relate to relative survival should also hold: appropriate expected mortality rates and conditional independence of the outcomes. The former highlights the importance of representative lifetables, and the latter requires that relative survival and expected survival are independent after adjusting for sufficient variables (Lambert et al., 2015; Seppä, Hakulinen, Läärä, & Pitkäniemi, 2016). Under these assumptions, relative survival can be interpreted as a net survival measure in a hypothetical world with cancer being the only possible cause of death. If interest is in obtaining “real”-world probabilities, we can estimate measures such as standardized crude probabilities and avoidable deaths measures, by incorporating expected mortality rates.

The exposures and confounders considered in this paper are time-fixed. However, appropriate methodology that accounts for time-varying exposures or confounders has been suggested before and this can be extended in the relative survival framework to allow the estimation of relevant causal parameters (Robins, Hernán, & Brumback, 2000).

Even though cancer inequalities had been well documented, understanding the underlying determinants of these differences is a challenging task. In this paper, we utilized mediation analysis methods and incorporated the relative survival framework to address these challenges. The proposed method has the advantage of allowing us to focus on cancer-related differences, the underlying determinants of which may be easier to identify in comparison with all-cause differences. Adjusting for sufficient confounders is essential, and caution is required when interpreting the findings.

ACKNOWLEDGEMENTS


This work was supported by a doctoral research fellowship to Elisavet Syriopoulou (reference: DRF-2017-10-116) from National Institute for Health Research. This paper 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. This work was also supported by a grant (Grant number C1483/A18262) to Paul C. Lambert from Cancer Research UK.

Supporting Information including example of Stata code for obtaining estimates for the measures of interests described in this article may be found online in the Supporting Information

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

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

REFERENCES

- Bower, H., Andersson, T. M. L., Crowther, M. J., Dickman, P. W., Lambe, M., & Lambert, P. C. (2017). Adjusting expected mortality rates using information from a control population: An example using socioeconomic status. *American Journal of Epidemiology*, *187*, 828–836.
- Cole, S. R., & Hernán, M. A. (2004). Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine*, *75*, 45–49.
- Cortese, G., & Scheike, T. H. (2008). Dynamic regression hazards models for relative survival. *Statistics in Medicine*, *27*, 3563–3584.
- Danø, H., Andersen, O., Ewertz, M., Petersen, J. H., & Lynge, E. (2003). Socioeconomic status and breast cancer in Denmark. *International Journal of Epidemiology*, *32*, 218–224.
- De Stavola, B. L., Daniel, R. M., Ploubidis, G. B., & Micali, N. (2014). Mediation analysis with intermediate confounding: Structural equation modeling viewed through the causal inference lens. *American Journal of Epidemiology*, *181*, 64–80.
- Department for Communities and Local Government. (2010). The English indices of deprivation 2010. Retrieved from <http://www.communities.gov.uk/documents/statistics/pdf/1871208.pdf>
- Dickman, P. W., Sloggett, A., Hills, M., & Hakulinen, T. (2004). Regression models for relative survival. *Statistics in Medicine*, *23*, 51–64.
- Dickman, P. W., & Coviello, E. (2015). Estimating and modelling relative survival. *The Stata Journal*, *15*, 186–215.
- Ederer, F., Axtell, L., & Cutler, S. (1961). The relative survival rate: A statistical methodology. *National Cancer Institute Monograph*, *6*. Bethesda, MD: National Cancer Institute.
- Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Ellis, L., Coleman, M. P., & Rachet, B. (2014). The impact of life tables adjusted for smoking on the socio-economic difference in net survival for laryngeal and lung cancer. *British Journal of Cancer*, *111*, 195–202.
- Estève, J., Benhamou, E., Croasdale, M., & Raymond, L. (1990). Relative survival and the estimation of net survival: Elements for further discussion. *Statistics in Medicine*, *9*, 529–538.
- Glymour, M. M., & Spiegelman, D. (2017). Evaluating public health interventions: 5. causal inference in public health research—do sex, race, and biological factors cause health outcomes? *American Journal of Public Health*, *107*, 81–85.
- Hakulinen, H., & Tenkanen, L. (1987). Regression analyses of relative survival rates. *Applied Statistics*, *36*, 309–317.

- Hernán, M. A., & Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health, 60*, 578–586.
- Imai, K., Keele, L., & Tingley, D. A. (2010). General approach to causal mediation analysis. *Psychological Methods, 15*, 309–334.
- Ito, Y., Nakaya, T., Nakayama, T., Miyashiro, I., Ioka, A., Tsukuma, H., & Rachet, B. (2014). Socioeconomic inequalities in cancer survival: A population-based study of adult patients diagnosed in Osaka, Japan, during the period 1993–2004. *Acta Oncologica, 53*, 1423–1433.
- Jeffreys, M., Sarfati, D., Stevanovic, V., Tobias, M., Lewis, C., Pearce, N., & Blakely, T. (2009). Socioeconomic inequalities in cancer survival in New Zealand: The role of extent of disease at diagnosis. *Cancer Epidemiology, Biomarkers & Prevention, 18*, 915–921.
- Krieger, N., & Davey Smith, G. (2016). The tale wagged by the dog: Broadening the scope of causal inference and explanation for epidemiology. *International Journal of Epidemiology, 45*, 1787–1808.
- Lambert, P. C., Dickman, P. W., & Rutherford, M. J. (2015). Comparison of approaches to estimating age-standardized net survival. *BMC Medical Research Methodology, 15*, 64.
- Neighbourhood Renewal Unit (2014). *The English indices of deprivation 2004 (revised)*. London: Office of the Deputy Prime Minister.
- Nelson, C. P., Lambert, P. C., Squire, I. B., & Jones, D. R. (2007). Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine, 26*, 5486–5498.
- Pavlic, K., & Pohar Perme, M. (2019). Using pseudo-observations for estimation in relative survival. *Biostatistics, 20*, 384–399.
- Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 411–420). San Francisco, CA: Morgan Kaufmann.
- Pearl, J. (2012). The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention Science, 13*, 426–436.
- Pearl, J. (2018). Does obesity shorten life? Or is it the soda? On non-manipulable causes. *Journal of Causal Inference, 6*(2), 1–6.
- Pohar Perme, M., Stare, J., & Estève, J. (2012). On estimation in relative survival. *Biometrics, 68*, 113–120.
- Rachet, B., Ellis, L., Maringe, C., Chu, T., Nur, U., Quaresma, M., & Coleman, M. P. (2010). Socioeconomic inequalities in cancer survival in England after the NHS cancer plan. *British Journal of Cancer, 103*, 446–453.
- Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology, 11*(5), 550–560. <https://doi.org/10.1097/00001648-200009000-00011>
- Royston, P. (2001). Flexible parametric alternatives to the Cox model, and more. *The Stata Journal, 1*, 1–28.
- Rubio, F. J., Rachet, B., Giorgi, R., Maringe, C., & Belot, A. (2019). On models for the estimation of the excess mortality hazard in case of insufficiently stratified life tables. *Biostatistics*. Advance online publication. <https://doi.org/10.1093/biostatistics/kxz017>
- Rutherford, M. J., Andersson, T. M. L., Møller, H., & Lambert, P. C. (2015). Understanding the impact of socioeconomic differences in breast cancer survival in England and Wales: Avoidable deaths and potential gain in expectation of life. *Cancer Epidemiology, 39*, 118–125.
- Seppä, K., Hakulinen, T., Läärä, E., & Pitkaniemi, J. (2016). Comparing net survival estimators of cancer patients. *Statistics in Medicine, 35*, 1866–1879.
- Sjölander, A. (2016). Regression standardization with the R package stdReg. *European Journal of Epidemiology, 31*, 563–574.
- Stefanski, L., & Boos, D. (2002). The calculus of M-estimation. *The American Statistician, 56*, 29–38.
- Syriopoulou, E., Bower, H., Andersson, T. M-L., Lambert, P. C., & Rutherford, M. J. (2017). Estimating the impact of a cancer diagnosis on life expectancy by socio-economic group for a range of cancer types in England. *British Journal of Cancer, 117*, 1419–1426.
- Syriopoulou, E., Rutherford, M. J., & Lambert, P. C. (2020). Marginal measures and causal effects using the relative survival framework. *International Journal of Epidemiology, 49*(2), 619–628.
- Vandenbroucke, J. P., Broadbent, A., & Pearce, N. (2016). Causality and causal inference in epidemiology: The need for a pluralistic approach. *International Journal of Epidemiology, 45*, 1776–1786.
- VanderWeele, T. J. (2011). Causal mediation analysis with survival data. *Epidemiology, 22*, 582–585.
- VanderWeele, T. J., Vansteelandt, S., & Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology, 25*(2), 300–306.
- Vansteelandt, S., & Daniel, R. M. (2017). Interventional effects for mediation analysis with multiple mediators. *Epidemiology, 28*, 258–265.
- Young, J. G., Stensrud, M. J., Tchetgen Tchetgen, E. J., & Hernán, M. A. (2020). A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in Medicine, 39*(8), 1199–1236.

SUPPORTING INFORMATION

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

How to cite this article: Syriopoulou E, Rutherford MJ, Lambert PC. Understanding disparities in cancer prognosis: An extension of mediation analysis to the relative survival framework. *Biometrical Journal*. 2021;63:341–353. <https://doi.org/10.1002/bimj.201900355>