The International Journal of Biostatistics

Volume 6, *Issue* 1 2010 *Article* 20

A Principal Stratification Approach to Assess the Differences in Prognosis between Cancers Caused by Hormone Replacement Therapy and by Other Factors

Arvid Sjolander, *Karolinska Institute*
Stijn Vansteelandt, *Universiteit Gent*
Keith Humphreys, *Karolinska Institute*

Recommended Citation:
Sjolander, Arvid; Vansteelandt, Stijn; and Humphreys, Keith (2010) "A Principal Stratification Approach to Assess the Differences in Prognosis between Cancers Caused by Hormone Replacement Therapy and by Other Factors," *The International Journal of Biostatistics*: Vol. 6: Iss. 1, Article 20.
DOI: 10.2202/1557-4679.1225

A Principal Stratification Approach to Assess the Differences in Prognosis between Cancers Caused by Hormone Replacement Therapy and by Other Factors

Arvid Sjolander, Stijn Vansteelandt, and Keith Humphreys

Abstract

Several recent studies have reported that women who have used hormone replacement therapy (HRT), and developed breast cancer, tend to have a better prognosis than women with breast cancer who have not used HRT. One possible explanation is that tumors caused by HRT are more benign than tumors caused by other factors. Although it is relevant to quantify differences in prognostic factors across subtypes of breast cancer, it is not obvious how to do this correctly. This is because the tumors which occur among women who are treated with HRT are a mixture of HRT-induced and other tumors. We propose a framework based on principal stratification to distinguish women with HRT-induced tumors from women with tumors caused by other factors. To estimate the difference in prognosis for these two groups, we propose two estimation methods, which can be used under both cohort and case-control sampling schemes.

KEYWORDS: counterfactuals, HRT, estimating equation, principal stratification, prognosis

Author Notes: Financial support from the Swedish Research Council (621-2004-3940 for AS and 523-2006-972 for KH) and the Swedish Foundation for Strategic Research (A3 02:129) is gratefully acknowledged. Vansteelandt acknowledges support from IAP research network grant nr. P06/03 from the Belgian government (Belgian Science Policy). We thank Prof. Juni Palmgren, Prof. Per Hall and Dr. Sara Wedren for valuable discussions.

1 Introduction

It has been established that long term use of Hormone Replacement Therapy (HRT) is a risk factor for breast cancer (Collaborative Group on Hormonal Factors in Breast Cancer, 1997). There is also clear evidence of a dosage effect of HRT on risk (See Figure 3 in Million Women Study Collaborators (2003)). Many recent studies have investigated the relationship between HRT use and prognosis in breast cancer patients; see Antoine et al (2004) for an overview. Most of these studies reported HRT to be associated with a favorable prognosis. Several authors have suggested that this association may be explained by confounding due to non-randomization of HRT. Taken at face value however the association has a plausible explanation; that HRT is associated with an increased risk of tumors which are less aggressive than other breast cancer tumors. This explanation has been suggested by, for example, Rosenberg et al (2008) in their interpretation of a case-only regression analysis of prognosis on HRT. This argument implies unobserved heterogeneity of breast cancers, in terms of prognosis. A case-only regression analysis of prognosis on HRT, treating all cancers as a homogeneous group, can therefore be considered inappropriate for assessing the HRT-prognosis relationship. We propose an alternative analysis approach which explicitly accounts for this heterogeneity. To see how HRT being associated with an increased risk of less aggressive tumors can lead to its opposite relationships with risk and prognosis, consider a simplified scenario in which each women is classified as either 'treated' (with HRT) or 'untreated' (We will later consider HRT on a continuous scale). Suppose that there exists three types of women; 1) those who do not develop cancer regardless of whether they are treated with HRT or not - we may call them 'healthy', 2) those who develop cancer if they are treated, but not otherwise - we may call them 'sensitive', and 3) those who develop cancer regardless of whether they are treated or not - we may call them 'doomed'. Table 1 illustrates this classification. We may interpret 'sensitive' (under treatment) and 'doomed' as

		subject type		
	cancer	healthy	sensitive	doomed
HRT	no	no	no	yes
	yes	no	yes	yes

Table 1: Classification of women into 'healthy', 'sensitive', and 'doomed'.

carriers of biologically different tumor types. 'Sensitive' individuals develop

tumors (under treatment) which are caused by HRT, whereas 'doomed' individuals have tumors of other causes. In realistic studies, these types of women can not be uniquely distinguished. Women who develop breast cancer that have not been treated with HRT must be 'doomed', whereas those who are treated and develop cancer are a mixture of 'sensitive" and 'doomed'. By assessing the association between HRT and prognosis in breast cancer patients, we are comparing the prognosis for untreated women with tumors not caused by HRT ('doomed') against a mixture of treated women with tumors caused by HRT ('sensitive') and tumors not caused by HRT ('doomed'). If women with tumors caused by HRT have a better prognosis than women with tumors not caused by HRT (Table 2), then it is clear that the treated patients will have a favorable prognosis compared to untreated patients. According to our discussions with subject matter experts if a woman's tumor has not been (clinically) induced by hormones then it is not likely that the amount of hormones she has been exposed to will influence her prognosis. Table 2 is, however, not the only possible explanation of HRT having opposite risk/prognosis relationships. Such relationships would also manifest if HRT improves the prognosis for those who have tumors not caused by HRT (Table 3). Another possibility leading to the same HRT-risk/prognosis relationships is that prognosis is HRT dependent within the doomed stratum, with the HRT group having a worse prognosis than the no HRT group among the doomed, but that the sensitive stratum has the most favorable prognosis. Given however that Table 2 is the most plausible and has been implied by breast cancer researchers, it is of clinical interest to quantify the association between HRT and prognosis under this condition, which is what we do in this article. To do this, however, is not a trivial task since women with different tumor subtypes can not be uniquely distinguished.

		subject type		
prognosis		healthy	sensitive	doomed
HRT	no	*	*	unfavorable
	yes	*	favorable	unfavorable

Table 2: Illustration of the hypothesis that women with tumors caused by HRT have a better prognosis than women with tumors not caused by HRT.

In this paper we formalize the relationship between HRT use and prognosis along the lines outlined above, using the framework of potential outcomes and principal stratification (Frangakis and Rubin, 2002). We consider a general

DOI: 10.2202/1557-4679.1225

		subject type		
	prognosis	healthy	sensitive	doomed
HRT	no	*	*	unfavorable
	yes	*	favorable	favorable

Table 3: Illustration of the hypothesis that HRT improves the prognosis for those who have tumors not caused by HRT.

scenario, for which HRT is defined on a continuous scale, and for which additional information on covariates is available. We show that under a set of reasonable assumptions, we can estimate the difference in prognosis for women with tumors caused by HRT, and women with tumors caused by other factors. The paper is organized as follows. In Section 2 we show how the framework of potential outcomes and principal stratification can be applied within the context of HRT, breast cancer, and prognosis. We lay out a number of fundamental assumptions and define the estimand of interest, which is a measure of the difference in prognosis for women with different tumor subtypes. In Sections 3 and 4 we discuss identifiability of the estimand, and propose two estimation methods. We consider both inference for cohort studies (Section 3) and for case-control studies (Section 4). In Section 5 we apply the proposed method to data from a population based case-control study of postmenopausal breast cancer in Swedish women. In Section 6 we carry out a small simulation study to investigate the performance of our approach. In Section 7 we discuss how our work relates to the literature on post-treatment selection bias.

2 Definitions and assumptions

We use Z to denote disease status; $Z = 1$ for cancer and $Z = 0$ for no cancer. We use X to denote duration of past HRT use. We assume that X is continuous on the range $[0, \infty)$, where '0' corresponds to 'no past HRT use'. We use C to denote measured baseline covariates. C is allowed to contain covariates measured on any mixture of scales. We assume that for each woman diagnosed with cancer, a particular prognostic factor Y is measured. Y is allowed to be measured on any scale. We assume that higher levels of Y indicate a worse prognosis. To define the target estimand we follow Frangakis and Rubin (2002) and use the framework of potential outcomes and principal stratification. We let $Z(x)$ and $Y(x)$ denote the potential outcomes of Z and Y, at HRT level

$X = x$. The following consistency assumption relates the potential outcomes to the observables:

A 1 $Z(X) = Z;\ Y(X) = Y.$

A1 states that the potential outcomes corresponding to the factual HRT level are equal to the observed outcomes Z and Y. The potential outcomes corresponding to the other (counterfactual) levels are unobserved, and are considered as missing. For mathematical convenience we define $Y(x) = *$ ('not defined') when $Z(x) = 0$ ($\Rightarrow Y = *$ when $Z = 0$). Let $Z(\cdot)$ denote the entire potential outcome function $\{Z(x)\ \forall x\}$. Figure 1 shows three examples of potential outcome functions. The solid line represents women who develop cancer if and only if they are treated with HRT for 3 or more years. The dashed line represents women who develop cancer if and only if they are treated with HRT for less than 4 years. The dotted line represents women who develop cancer if they are treated with HRT for 1-2 years or 5-6 years. We say that two women belong to the same principal stratum if they have the same potential outcome functions $Z(\cdot)$.

Although HRT could hypothetically prevent breast cancer for some women, this is unlikely. We therefore assume that a woman who develops cancer at one level of HRT, would also have developed cancer if she would have been exposed to a higher level of HRT:

A 2 $Z(x) \geq Z(x')$ *if* $x \geq x'$.

Under A2, $Z(\cdot)$ is a step function such as the solid line in Figure 1, and functions such as the dashed and dotted line are assumed not to exist. Hence, under A2, we can, without loss of information, summarize $Z(\cdot)$ into a scalar, R, defined as the minimum level x for which $Z(x) = 1$. For a woman with $Z(\cdot)$ represented by the solid line in Figure 1, $R = 3$. R is continuous on the range $(0, \infty)$, and its distribution has a point mass at 0, where the stratum whose women develop cancer regardless of whether they are treated or not ($Z(x) = 1\ \forall x$), is represented. Previous research has established a dose-respone relationship in line with A2 (Million Women Study Collaborators, 2003). Furthermore, all subject matter experts that we have discussed the issue with, have confirmed that A2 is reasonable from a biological perspective.

We will assume that X can be considered randomized within levels of C. In terms of potential outcomes we formulate this assumption as

A 3 $\{Y(x), R\} \amalg X | C \quad \forall x,$

DOI: 10.2202/1557-4679.1225

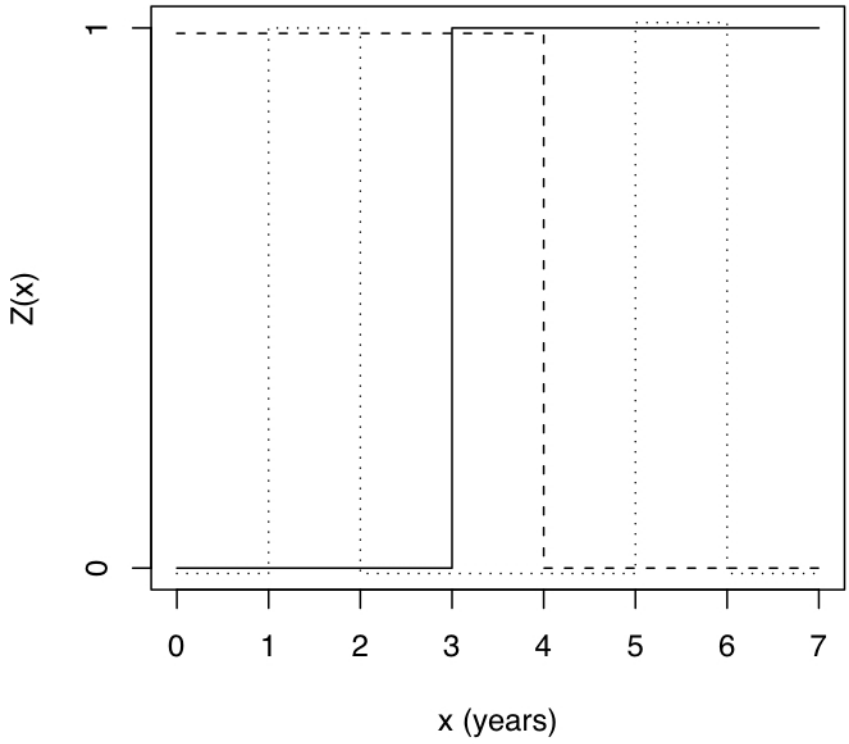

Figure 1: Examples of potential outcome functions.

where II is used to denote statistical independency. We remind the reader that $Y(x) = *$ when $x < R$. Whether A3 is plausible or not depends, of course, on which covariates are included in C. Women use HRT for several reasons, but mainly because of menopausal symptoms, to prevent osteoperosis or to prevent heart disease. Thus, relevant covariates, which we have used in our analysis of data (Section 5), are age and BMI. If all relevant covariates are included in C, A3 follows naturally.

For our purpose it is useful to think about the principal strata as representing women with different types of cancers. We distinguish between two types of cancers, those which are HRT-induced and those which are not. We call a specific tumor HRT-induced if it occurs in the presence of HRT use, but would

not have occurred in the absence of HRT use. A woman with $R > 0$ who develops cancer ($Z = 1$), would not have done so had she not been treated. Thus, for a woman with $(R > 0, Z = 1)$, the cancer is HRT-induced. A woman with $R = 0$ develops cancer regardless of whether she is treated or not. If she is not treated, then we can with certainty say that the cancer is non-HRT-induced. If she is treated, then the situation is more complex. Hypothetically, she may develop two types of cancer under HRT use, one which is HRT-induced and one which is non-HRT-induced. In this case, the observed cancer may be either exactly one of these types (if one of them occurred and was diagnosed before the other one occurred), or a combination of both. For most women, however, (moderate levels of) HRT does not cause cancer. Hence, most of the observed cancers within $R = 0$ are bound to be non-HRT-induced. We therefore proceed by assuming that women with cancer and $R > 0$ have HRT-induced tumors, and women with cancer and $R = 0$ have non-HRT-induced tumors. We note that this assumption is crucial in our analysis as it entails identification by eliminating the 'competing risk problem' with one tumor type preempting the other.

By contrasting the distribution of potential outcomes $Y(x)$ for the stratum $0 < R \leq x$ with the stratum $R = 0$, we measure the discrepancy in prognosis between women with HRT-induced tumors vs women with non-HRT-induced tumors. More specifically, our target estimand is

$$m(x, C) \equiv g[E\{Y(x)|0 < R \leq x, C\}] - g[E\{Y(x)|R = 0, C\}],\ x > 0,$$

where $g(\cdot)$ is a known, smooth, monotone link function. We define $m(0, C) \equiv 0$. We note that $m(x, C)$ does not measure the causal effect of HRT on prognosis *per se*. This is because it compares the potential outcome $Y(x)$ for two different groups of people (Frangakis and Rubin, 2002). In particular, suppose that there exists a mutation in a gene which causes non-HRT-induced tumors and also affects general health status through other pathways. In such case, women with HRT-induced tumors and women with non-HRT-induced tumors will have different prognosis ($m(x, C) \neq 0$), even if the tumors themselves are completely identical in every relevant aspect. To separate such a 'confounding effect' from a 'tumor effect', however, would require randomization of tumors, which is, of course, not possible.

To better understand what $m(x, C)$ does measure, consider the following simple numerical example. Assume that X is categorical with levels $\{0, 1, 2\}$, and that the population can be divided into three principal strata $R = 0$, $R = 1$, and $R = 2$, with equal population proportions, given covariates C; $Pr(R = r|C) = 1/3$ for $r \in \{0, 1, 2\}$. Women with $R = 0$ develop cancer regardless of treatment level. These cancers are non-HRT-induced. Women with $R = 1$

DOI: 10.2202/1557-4679.1225

develop cancer if they are treated at level $X = 1$ or $X = 2$, and women with $R = 2$ develop cancer only if they are treated at level $X = 2$. These cancers are HRT-induced. Assume that $E\{Y(x)|R = 0, C\} = 3$ for $x \in \{0, 1, 2\}$, $E\{Y(x)|R = 1, C\} = 2$ for $x \in \{1, 2\}$, and $E\{Y(2)|R = 2, C\} = 0$. The situation is represented in Table 4. $m(1, C)$ compares the mean prognosis for

		R		
$E\{Y(x)\|R, C\}$		2	1	0
	0	*	*	3
x	1	*	2	3
	2	0	2	3

Table 4: Numerical example 1. $Y(x) \equiv *$ when $R > x \Leftrightarrow Z(x) = 0$.

women with HRT-induced tumors vs women with non-HRT-induced tumors, at treatment level $X = 1$. Assuming an identity link, $g(\cdot) = \cdot$, we have that $m(1, C) = E\{Y(1)|R = 1, C\} - E\{Y(1)|R = 0, C\} = 2 - 3 = -1$. $m(2, C)$ compares the mean prognosis for women with HRT-induced tumors vs women with non-HRT-induced tumors, at treatment level $X = 2$. We have that $m(2, C) = E\{Y(2)|R = 2, C\}\frac{Pr(R=2|C)}{Pr(R=2|C)+Pr(R=1|C)} + E\{Y(2)|R = 1, C\}\frac{Pr(R=1|C)}{Pr(R=2|C)+Pr(R=1|C)} - E\{Y(2)|R = 0, C\} = 0 \times \frac{1/3}{1/3+1/3} + 2 \times \frac{1/3}{1/3+1/3} - 3 = -2$. In this numerical example, women with HRT-induced tumors have different a prognosis than women with non-HRT-induced tumors, at both levels $X = 1$ and $X = 2$, since neither $m(1, C)$, nor $m(2, C)$ equals 0. Furthermore, this difference becomes more accentuated at higher levels of HRT use, since $m(2, C) < m(1, C)$.

The example further clarifies that $m(x, C)$ is not a measure of the HRT effect on prognosis *per se*. Indeed, within each stratum R, the mean prognosis is the same for all HRT-levels at which $Z(x) = 1$. Hence, although $m(x, C)$ depends on treatment level in this example, HRT does not have any effect on prognosis. The reason why $m(x, C)$ depends on treatment level in the example, is because women with HRT-induced tumors occurring at level $X = 2$ have different prognosis from women with HRT-induced tumors occurring at level $X = 1$. That $m(x, C)$ is not a measure of the HRT effect *per se*, however, does not mean that it is not influenced by an existing HRT effect. To see this, consider the numerical example in Table 5. In this second example, $m(1, C) = -1$ and $m(2, C) = -2$, as in the first example, but HRT now has an effect on prognosis for women within $R = 1$, since $E\{Y(2)|R = 1, C\} \neq E\{Y(1)|R = 1, C\}$. The two examples tell us that an x-gradient in $m(x, C)$

		R		
$E\{Y(x)\|R,C\}$		2	1	0
	0	*	*	3
x	1	*	2	3
	2	1	1	3

Table 5: Numerical example 2. $Y(x) \equiv *$ when $R > x \Leftrightarrow Z(x) = 0$.

may be explained either by heterogeneity in prognosis for women with HRT-induced tumors, or by an HRT effect on prognosis, or a combination of both. To disentangle one phenomenon from the other would require additional strong assumptions.

If a particular woman has a tumor which is HRT-induced, then it is quite possible that her prognosis depends on the dosage of HRT. A larger dosage may for example cause a more aggressive cancer. According to our discussions with subject matter experts, if a woman's tumor is non-HRT-induced, then there is no reason to believe that her prognosis is affected by her dosage of HRT. In particular we will assume

A 4 $Pr\{Y(x)|R=0,C\} = Pr\{Y(0)|R=0,C\}\ \forall x.$

3 Inference from cohort studies

A cohort study generates an iid sample from $Pr(Y,Z,X,C)$.

3.1 Identification

Define $\pi(X,C) \equiv Pr(Z=1|X,C)$. Under A1-A3 it can be shown (see Appendix A) that

$$Pr(R \leq x|C) = \pi(x,C) \tag{1}$$

In particular we have that $Pr(R=0|C) = \pi(0,C)$. The relation in (1) implies that under cohort sampling, the conditional distribution of principal strata, $Pr(R|C)$, is identified. Equation (1) also implies the following testable restriction

$$\pi(x,C) \geq \pi(x',C) \quad \text{if } x \geq x'. \tag{2}$$

Thus, under cohort sampling, A1-A3 are partially testable, since violations of the inequality in (2) falsify at least one of the assumptions.

DOI: 10.2202/1557-4679.1225

Under A1-A3 the conditional distribution of Y, given $(Z=1,X,C)$, is a mixture of potential outcomes for women with HRT-induced and non-HRT-induced cancers (see Appendix A):

$$\begin{aligned} &Pr(Y|Z=1,X=x,C) \\ &= Pr\{Y(x)|R=0,C\}\frac{\pi(0,C)}{\pi(x,C)} + Pr\{Y(x)|0<R\leq x,C\}\left\{1-\frac{\pi(0,C)}{\pi(x,C)}\right\}. \end{aligned} \tag{3}$$

The key to identifiability of the mixture components is A4. Combining (3) with A4 gives that

$$\begin{aligned} &m(x,C) \\ &= g\left\{\frac{E(Y|Z=1,X=x,C) - E(Y|Z=1,X=0,C)\frac{\pi(0,C)}{\pi(x,C)}}{1-\frac{\pi(0,C)}{\pi(x,C)}}\right\} \\ &- g\{E(Y|Z=1,X=0,C)\}. \end{aligned} \tag{4}$$

The expression in (4) shows that $m(x,C)$ is identified if $\pi(x,C) \neq \pi(0,C)$. We will throughout assume that this is the case.

In order for $Pr\{Y(x)|0<R\leq x,C\}$ to be a proper density/distribution, it has to be non-negative. This requirement, combined with (3) and A4, shows that A4 implies the following additional testable restriction

$$Pr(Y,Z=1|X=0,C) \leq Pr(Y,Z=1|X,C). \tag{5}$$

It can be shown (see Appendix B) that A1-A4 do not imply any restrictions on the joint distribution for (Y,Z,X,C) other than those stated by (2) and (5). We note that although A1-A4 imply (2) and (5), the reverse does not hold. Thus, A1-A4 can be faslified but not verified from data.

3.2 Estimation

Modeling of $m(x,C)$ is required to deal with its high dimensionality. We propose two approaches; one implicit (Section 3.2.1) and one explicit (Section 3.2.2).

3.2.1 Implicit method

One possible approach is to specify models for $E(Y|Z=1,X,C)$ and $\pi(X,C)$ indexed by ξ and α, respectively, and use the relation in (4) to translate these

models into a model for $m(x, C)$. For example, using the standard models

$$E(Y|Z = 1, X, C; \xi) = \xi_0 + \xi_1 X + \xi_2 C \quad (6)$$

$$\text{logit}\pi(X, C; \alpha) = \alpha_0 + \alpha_1 X + \alpha_2 C, \quad (7)$$

and the identity link $g(\cdot) = \cdot$, yields

$$m(x, C; \xi, \alpha) = \frac{\xi_1 x}{1 - \frac{\text{expit}(\alpha_0 + \alpha_2 C)}{\text{expit}(\alpha_0 + \alpha_1 x + \alpha_2 C)}}. \quad (8)$$

One can view the denominator in (8) as a 'bias correction', required to give the regression parameter ξ_1 the desired interpretation as a difference in prognosis between women with different tumor subtypes. Replacing ξ and α with their (ML-)estimates, $\hat{\xi}$ and $\hat{\alpha}$, yields an estimate of $m(x, C; \xi, \alpha)$. The standard error of $m(x, C; \hat{\xi}, \hat{\alpha})$, as a function of (x, C), can be obtained from the delta method. We will refer to this method as 'implicit'. The advantage of the implicit method is that it is computationally straightforward. A disadvantage is that the implied model for $m(x, C)$ may be quite complex and hard to interpret. In general, $m(x, C; \xi, \alpha)$ depends on x through the full parameter vector (ξ, α), and through the particular value of the covariate vector C. Hence, it may be hard to formulate scientifically relevant questions about $m(x, C)$ using the implicit method. An important exception is the null hypothesis $m(x, C) = 0$. From (4), it follows that $m(x, C) = 0 \Leftrightarrow E(Y|Z = 1, X, C) = E(Y|Z = 1, C)$. Hence, under (6), a test for $m(x, C) = 0$, $\forall x$, is equivalent to a test for $\xi_1 = 0$. The relation $E(Y|Z = 1, X, C) = E(Y|Z = 1, C)$ can of course be tested in other parametric or semiparametric models as well, or even nonparametrically.

3.2.2 Explicit method

In this section we propose an alternative method in which $m(x, C)$ is modeled directly, and works when $g(\cdot)$ is the identity link or the log-link. We will refer to this method as 'explicit'. The explicit method involves fitting one main model together with three 'nuisance' models.

Main model. We assume a model for $m(x, C)$ indexed by a parameter ψ:

$$m(x, C) = m(x, C; \psi). \quad (9)$$

An example is the linear model

$$m(x, C; \psi) = \psi_0 + \psi_1 x. \quad (10)$$

DOI: 10.2202/1557-4679.1225

In (10), $\psi_0 = \lim_{x \to 0} m(x, C)$ represents an intrinsic, non-HRT-dependent, difference in prognosis between women with different tumor subtypes, whereas $_1 = m(x+1, C) - m(x, C)$ represents a x-gradient in $m(x, C)$. If, for example, people with HRT induced tumors have (on average) a better prognosis than people with non-HRT-induced tumors, but the particular level of HRT is not associated with the prognosis, then $(\psi_0 \neq 0, \psi_1 = 0)$. If the level of HRT has an influence on the prognosis, then we would expect that $\psi_1 \neq 0$ as well. Note $_0 \neq 0$ implies a discontinuity for $m(x, C)$ in 0, i.e. $\lim_{x \to 0} m(x, C) \neq m(0, C)$. In a comparison $m(x+1, C) - m(x, C)$ this discontinuity cancels. In (10), we make the assumption that $m(x, C)$ does not depend on C. This assumption does not imply that $E\{Y(x)|R, C\}$ does not depend on C. Since $m(x, C)$ is a mean difference, it implies that there is no interaction between x and C on the particular scale defined by $g(\cdot)$. For the identity link, we define

$$H_i(\psi) = Y_i - m(X_i, C_i; \psi) \left\{ 1 - \frac{\pi(0, C_i)}{\pi(X_i, C_i)} \right\}. \tag{11}$$

For the log-link, we define

$$H_i(\psi) = \frac{Y_i}{\frac{\pi(0,C_i)}{\pi(X_i,C_i)} + \exp\{m(X_i, C_i; \psi)\} \left\{ 1 - \frac{\pi(0,C_i)}{\pi(X_i,C_i)} \right\}}. \tag{12}$$

We estimate ψ as the solution to the estimating equation

$$\sum_{i=1}^{n} d(X_i, C_i) Z_i \{ H_i(\psi) - q(C_i) \} = 0, \tag{13}$$

where $d(X_i, C_i) \in \mathbb{R}^p$ is an arbitrary function satisfying $E\{d(X_i, C_i)|Z_i = 1, C_i\} = 0$, and $q(C_i)$ is an arbitrary function. The estimating equation in (13) is unbiased (see Appendix C) which assures that the solution, $\hat{\psi}$, is consistent, and asymptotically normal (under standard regularity conditions). We define

$$G_i = E \left\{ \frac{\partial H_i(\psi)}{\partial \psi} | Z_i = 1, X_i, C_i \right\},$$

We will use the index functions

$$d^{opt}(X_i, C_i) = \sigma^{-2}(X_i, C_i) \left[G_i - \frac{E\left\{ G_i \sigma^{-2}(X_i, C_i) | Z_i = 1, C_i \right\}}{E\left\{ \sigma^{-2}(X_i, C_i) | Z_i = 1, C_i \right\}} \right],$$

where $\sigma^2(X_i, C_i) \equiv \text{Var}\left\{ H_i(\psi) | Z_i = 1, X_i, C_i \right\}$, and

$$q^{opt}(C_i) = E(Y_i | Z_i = 1, X_i = 0, C_i).$$

It can be shown that these index functions maximize the efficiency of $\hat{\psi}$ (see Appendix D). We will throughout assume homoscedasticity; $\sigma^2(X_i, C_i) = \sigma^2$. This assumption does not induce bias if it is violated, but at worst a loss of efficiency. Under homoscedasticity, $d^{opt}(X_i, C_i)$ simplifies to

$$d^{opt}(X_i, C_i) = \sigma^{-2}\{G_i - E(G_i|Z_i = 1, C_i)\}.$$

When solving the equation in (13), σ^2 cancels. In practice, both $\pi(X_i, C_i)$ and the index functions $d^{opt}(X_i, C_i)$ and $q^{opt}(C_i)$ are unknown, and must be estimated. Estimation of these quantities is described below.

As shown in Appendix C, the estimating equation in (13) has the attractive feature of being unbiased when *either* $E\{d(X_i, C_i)|Z_i = 1, C_i\} = 0$, *or* $q(C_i) = q^{opt}(C_i)$. This implies that it is doubly robust, in the sense that $\hat{\psi}$ is consistent as long as, in addition to the models for $m(x, C)$ and $\pi(X, C)$, at least one of the index functions is correctly specified. The proposed estimator is somewhat related to doubly robust estimators found in the context of semiparametric regrssion models (Robins and Rotnitzky, 2001). For an overview of the use of doubly robust estimators, see Bang and Robins (2005).

Nuisance model 1. To estimate $\pi(X_i, C_i)$ we specify a model

$$\pi(X_i, C_i) = \pi(X_i, C_i; \alpha). \tag{14}$$

We obtain a consistent estimate of α by solving the maximum likelihood score equation

$$\sum_{i=1}^{n} S_i(\alpha) = \sum_{i=1}^{n} \frac{\partial}{\partial \alpha} \log[\pi(X_i, C_i; \alpha)^{Z_i}\{1 - \pi(X_i, C_i; \alpha)\}^{1-Z_i}] = 0. \tag{15}$$

After having solved (15), we replace the true value of $\pi(X_i, C_i)$ in (13) with its estimate, $\pi(X_i, C_i; \hat{\alpha})$. In Appendix D, we provide a more efficient estimation strategy based on joint estimation of ψ and α.

Nuisance model 2. To estimate $q^{opt}(C_i)$ we specify a model

$$E(Y_i|Z_i = 1, X_i = 0, C_i) = \mu(C_i; \beta). \tag{16}$$

We obtain a consistent (and efficient) estimate of β by solving the unbiased equation

$$\sum_{i=1}^{n} \frac{\partial \mu(C_i; \beta)/\partial \beta}{\text{var}(Y_i|Z_i = 1, X_i = 0, C_i)} I(X_i = 0) Z_i\{Y_i - \mu(C_i; \beta)\} = 0, \tag{17}$$

When $H_i(\psi)$ is defined as in (11), the assumption that $\sigma^2(X_i, C_i) = \sigma^2$ implies that $\text{var}(Y_i|Z_i = 1, X_i = 0, C_i) = \text{const}$. For $H_i(\psi)$ defined as in (12), we will

DOI: 10.2202/1557-4679.1225

also assume that $\text{var}(Y_i|Z_i = 1, X_i = 0, C_i) = \text{const}$. After having solved (17) we replace the true value of $q^{opt}(C_i)$ in (13) with its estimate, $\mu(C_i; \hat{\beta})$.

Nuisance model 3. To estimate $d^{opt}(X_i, C_i)$ we need to estimate the conditional mean $E\{G_i|Z_i = 1, C_i\}$. One option is to specify an explicit model for $E\{G_i|Z_i = 1, C_i\}$. This approach has three disadvantages. 1) $E\{G_i|Z_i = 1, C_i\}$ is determined by the law $Pr(X|Z = 1, C)$. This law is not variation independent of the law $\pi(X, C)$, and as a consequence an explicit model for $E\{G_i|Z_i = 1, C_i\}$ may not be logically compatible with the model $\pi(X, C; \alpha)$. 2) Fitting the model for $E\{G_i|Z_i = 1, C_i\}$ requires knowing the true value of G_i for each woman. This value, however, depends on the true value of ψ. Thus, an estimating equation for the parameter indexing the model for $E\{G_i|Z_i = 1, C_i\}$ has to be solved simultaneously with the equation for ψ, which increases the computational complexity. 3) G_i is a complicated function of the observables, and it may be hard, even for a subject matter expert, to well specify a model for its conditional mean. We therefore propose an alternative approach. We use Bayes rule to rewrite $E\{G_i|Z_i = 1, C_i\}$ as

$$E\{G_i|Z_i = 1, C_i\} = \frac{E\{G_i\pi(X_i, C_i)|C_i\}}{E\{\pi(X_i, C_i)|C_i\}}. \tag{18}$$

If we model and estimate the law $Pr(X_i|C_i)$ we can calculate $E\{G_i|Z_i = 1, C_i\}$ by averaging over $G_i\pi(X_i, C_i)$ and $\pi(X_i, C_i)$ conditional on C_i in (18). Thus, an explicit model for $Pr(X_i|C_i)$ translates into an implicit model for $E\{G_i|Z_i = 1, C_i\}$. $Pr(X_i|C_i)$ is clearly variation independent of $\pi(X, C)$, which rules out any model incompatibilities. Furthermore, the parameters of the model for $Pr(X_i|C_i)$ can be estimated without knowledge of ψ. Finally, when X represents a treatment ordinated by a physician, a subject matter expert may well be able to roughly specify the treatment distribution across levels of covariates. Note also that the model for $Pr(X_i|C_i)$ does not need to be entirely correct in order for $\hat{\psi}$ to be consistent so long as it yields the correct ratio of conditional expectations in (18). Furthermore, it follows from the double robustness of the estimating equation in (13) that $\hat{\psi}$ will be consistent even if this ratio is misspecified, as long as the model for $E(Y_i|Z_i = 1, X_i = 0, C_i)$ in (16) is correct. We specify a model

$$Pr(X_i|C_i) = Pr(X_i|C_i; \gamma). \tag{19}$$

We obtain a consistent (and efficient) estimate of γ by solving the maximum likelihood score equation

$$\sum_{i=1}^{n} S_i(\gamma) = \sum_{i=1}^{n} \frac{\partial}{\partial\gamma} \log Pr(X_i|C_i; \gamma) = 0. \tag{20}$$

3.2.3 Asymptotic properties of the explicit method

Define $\theta = (\alpha', \beta', \gamma', \psi')'$. Estimating θ following the procedure described above is equivalent to solving the joint equation

$$\sum_i^n U_i(\theta) = \left\{ \begin{array}{c} S_i(\alpha) \\ \frac{\partial \mu(C_i;\beta)/\partial \beta}{\text{var}(Y_i|Z_i=1,X_i=0,C_i)} I(X_i=0) Z_i \{Y_i - \mu(C_i;\beta)\} \\ S_i(\gamma) \\ d^{opt}(X_i, C_i; \alpha, \gamma) Z_i \{H_i(\psi, \alpha) - q^{opt}(C_i; \beta)\} \end{array} \right\} = 0,$$

where we have highlighted the dependency on (α, β, γ) in the last row of the equation system. Under regularity conditions, the combined estimating equation has a unique solution $\hat{\theta}$, where $n^{1/2}(\hat{\theta} - \theta)$ has an asymptotically normal distribution with mean 0 and variance

$$E' \left\{ \frac{\partial U_i(\theta)}{\partial \theta} \right\}^{-1} \text{var}\{U_i(\theta)\} E \left\{ \frac{\partial U_i(\theta)}{\partial \theta} \right\}^{-1} \tag{21}$$

Replacing θ in (21) by $\hat{\theta}$ and the population moments by their sample counterparts yields a sandwich estimator for the variance of $\hat{\theta}$.

4 Inference from case-control studies

A case-control study generates two samples; one iid sample of size n_1 from $Pr(Y, X, C|Z = 1)$, and one iid sample of size n_0 from $Pr(X, C|Z = 0)$. We define $n \equiv n_0 + n_1$. We develop inference for $m(x, C)$ assuming a low prevalence of breast cancer, i.e.

A 5 $\pi(x, c) \approx 0 \ \forall x, c.$

This assumption is often reasonable as it forms the basis for choosing case-control designs in practice.

4.1 Identification

Under a case-control sampling scheme, $\pi(X, C)$ is not identified. Hence, the conditional distribution of principal strata, $Pr(R|C)$, is not identified. When the disease is rare (A5), this poses no major identification problems for $m(x, C)$. To see this, note that the expression in (4) only contains $E(Y|Z =$

DOI: 10.2202/1557-4679.1225

$1, X, C)$, which is trivially identified under case-control sampling, and the relative risk $\frac{\pi(0,C)}{\pi(x,C)}$. Define the odds ratio

$$\eta(X,C) \equiv \frac{\pi(X,C)\{1-\pi(0,C)\}}{\pi(0,C)\{1-\pi(X,C)\}}.$$

It is well known that when $\pi(X,C) \approx 0$, the relative risk $\frac{\pi(0,C)}{\pi(X,C)}$ approximates the odds ratio $\eta^{-1}(X,C)$. From Bayes rule, $\eta(X,C)$ is trivially identified from case-control sampling, and equal to $\frac{Pr(X|Z=1,C)Pr(X=0|Z=0,C)}{Pr(X|Z=0,C)Pr(X=0|Z=1,C)}$. Hence, under A1-A5, $m(x,C)$ is 'approximately' identified under case-control sampling.

We note that A1-A3 are partially testable under case-control sampling, even when A5 is violated. This is because the restriction in (2) implies the testable odds ratio restriction

$$\eta(x,C) \geq \eta(x',C) \quad \text{if } x \geq x'. \tag{22}$$

4.2 Estimation

Under A5, both estimation procedures for cohort studies proposed in Section 3.2 can be used for case-control studies, with the following minor modifications.

4.2.1 Modified implicit method

Under A5, the implicit estimation method described in Section 3.2.1 can easily be adapted to case-control studies by replacing the relative risk $\frac{\pi(0,C)}{\pi(x,C)}$ in (4) with the odds ratio $\eta^{-1}(x,C)$. As an example, consider the implied model in (8). Under the rare-disease approximation, this model simplifies to

$$m(x,C;\xi,\alpha) = \frac{\xi_1 x}{1-\exp(-\alpha_1 x)}. \tag{23}$$

Hence, under A1-A5, $g(\cdot) = \cdot$, and the standard models in (6) and (7), $m(x,C)$ only depends on x through ξ_1 and α_1. It is instructive to consider the asymptotic (in x) behavior of the implied model in (23). We have that

$$\begin{aligned} \lim_{x\to 0} m(x,C;\xi,\alpha) &= \xi_1/\alpha_1, \\ \lim_{x\to\infty} m(x,C;\xi,\alpha)/x &= \xi_1. \end{aligned}$$

The limit value of ξ_1/α_1 can be interpreted as an intrinsic, non-HRT-dependent, difference in prognosis between women with different tumor subtypes, similar to ψ_0 in (10). The limit value of ξ_1 reflects the fact that for large values of x we expect that most tumors are HRT-induced. Hence, for large values of x, $m(x,C)/x \approx \{E(Y|Z=1,X=x,C) - E(Y|Z=1,X=0,C)\}/x = \xi_1$.

4.2.2 Modified explicit method

To adapt the explicit method described in Section 3.2.2 to case-control studies, we make the following approximations.

1. $H_i(\psi)$ contains the relative risk $\frac{\pi(0,C_i)}{\pi(X_i,C_i)}$. We replace the relative risk with the odds ratio, $\eta^{-1}(X_i, C_i)$, and use $\tilde{H}_i(\psi)$ to denote the resulting approximation to $H_i(\psi)$.

2. To calculate the efficient[1] index function, $d^{opt}(X_i, C_i)$, we need to calculate G_i and $E\{G_i|Z_i = 1, C_i\}$. G_i contains the relative risk $\frac{\pi(0,C_i)}{\pi(X_i,C_i)}$. We replace the relative risk with the odds ratio, $\eta^{-1}(X_i, C_i)$, and use $\tilde{G}_i$ to denote the resulting approximation to G_i. Fitting an implicit model for $E\{G_i|Z_i = 1, C_i\}$ (see Section 3.2.2) involves fitting a model for $Pr(X_i|C_i)$ and averaging over $G_i\pi(X_i, C_i)$ and $\pi(X_i, C_i)$, conditional on C_i. Under case-control sampling this is not straightforward, because neither $\pi(X_i, C_i)$, nor $Pr(X_i|C_i)$ is identified. Under A5, however, we have that $Pr(X_i|C_i) \approx Pr(X_i|Z_i = 0, C_i)$. Hence, we may fit the model for $Pr(X_i|C_i)$ to the controls separately. We thus replace the estimating equation in (20) by

$$\sum_{i=1}^{n}(1 - Z_i)S_i(\gamma) = 0,$$

which yields an approximately consistent estimate of γ under A5. We further divide the numerator and denominator in (18) by $\pi(0, C_i)$ to obtain

$$E\{G_i|Z_i = 1, C_i\} = \frac{E\left\{G_i\frac{\pi(X_i,C_i)}{\pi(0,C_i)}|C_i\right\}}{E\left\{\frac{\pi(X_i,C_i)}{\pi(0,C_i)}|C_i\right\}} \approx \frac{E\left\{\tilde{G}_i\eta(X_i, C_i)|C_i\right\}}{E\left\{\eta(X_i, C_i)|C_i\right\}}, \quad (24)$$

where the approximation is valid under A5. We use $\tilde{d}^{opt}(X_i, C_i)$ to denote the estimate of $d^{opt}(X_i, C_i)$ obtained by replacing G_i by $\tilde{G}_i$, and $E\{G_i|Z_i = 1, C_i\}$ by the right-hand-side of (24).

[1]Note that the $d^{opt}(X_i, C_i)$ is efficient under cohort sampling, but may not be efficient under case-control sampling. We conjecture though, that this is the case. This conjecture is supported by van der Laan (2008), who showed that the efficient score under cohort sampling stays efficient if it is inversely weighted by the known disease prevalence under case-control sampling.

DOI: 10.2202/1557-4679.1225

Under A5, $\tilde{H}_i(\psi) \approx H_i(\psi)$ and $\tilde{d}^{opt}(X_i, C_i) \approx d^{opt}(X_i, C_i)$, and the modified estimation method gives an approximately consistent estimate of ψ. When A5 is violated, then $E\{\tilde{d}^{opt}(X_i, C_i)|Z_i = 1, C_i\}$ may differ significantly from 0. Hence, the modified procedure may be more robust to deviations from A5 if $E\{G_i|Z_i = 1, C_i\}$ is modeled explicitly. Note also, that one of the arguments that we gave in Section 3.2.2 for not modeling $E\{G_i|Z_i = 1, C_i\}$ explicitly, is invalid under case-control sampling. More specifically, when $\pi(X_i, C_i)$ is only modeled up to the odds ratio $\eta(X_i, C_i)$, then there is no risk for incompatibility between the models for $E\{G_i|Z_i = 1, C_i\}$ and $\eta(X_i, C_i)$, since the law $Pr(X|Z = 1, C)$ is variation independent of $\eta(X_i, C_i)$ (see Appendix F). To estimate the odds ratio $\eta(X_i, C_i)$ it is natural to assume that $\pi(X_i, C_i)$ follows a logistic regression model. If so, a consistent estimate of $\eta(X_i, C_i)$ is obtained as the solution to the score equation (15) (Prentice and Pyke, 1979).

The variance expression in (21) is valid under cohort sampling. When the study is subject to case-control sampling, the variance expression must acknowledge this. Assume that $n_z n^{-1} = \rho_z$. Using a standard Taylor expansion argument, it can be shown (see Appendix E) that $n^{1/2}(\hat{\theta} - \theta)$ obtained from the estimation procedure described in this section, has an asymptotically normal distribution with mean 0 and variance $\mathcal{I}'^{-1}\mathcal{J}\mathcal{I}^{-1}$, where

$$\mathcal{I} = \sum_{z \in \{0,1\}} \rho_z E\left\{ \frac{\partial U_i(\theta)}{\partial \theta} \Big| Z = z \right\}$$

and

$$\mathcal{J} = \sum_{z \in \{0,1\}} \rho_z \text{var}\{U_i(\theta)|Z = z\}.$$

Replacing θ in (21) by $\hat{\theta}$ and the population moments by their sample counterparts yields a sandwich estimator for the variance of $\hat{\theta}$.

5 Application

We used the methods described above to analyze data from a large population-based case-control study of breast cancer in Swedish post-menopausal women, aged 50 to 74 years in 1993-1995. The study, known as CAHRES (Cancer And Hormone REplacementS), compares 3000 cases, and a similar number of healthy controls. Several papers have been published based on this data (e.g. Magnusson (1999)). Recently, Rosenberg (2006) analyzed breast cancer prognosis data obtained from the Swedish death registry, for cases in this

study, for association with HRT use. In line with the observational studies reviewed by Antoine et al (2004), Rosenberg (2006) reported a positive association between HRT use and favorable prognostic factors. In particular, associations were found between tumor size and grade. For our analysis we have combined these two prognostic factors into a single prognostic index (PI): PI = log{size(cm) + grade(1-3)}. A small value is interpreted as carrying a favorable prognosis. Equal weights for size and grade are used, as is also the case in the Nottingham Prognostic Index (NPI; Galea et al (1992)), a breast cancer prognosis index developed in the 1990's. Due to missingness in size and/or grade, PI could only be calculated for 986 cases. In what follows we use Y to denote the PI and X to denote duration of HRT use (in 1000 days using estrogen or progestin). We include age and body mass index (BMI) in the covariate vector C.

Before carrying out an analysis using the methods described in this paper, we assessed the appropriateness of the underlying assumptions 1-3. To do this we constructed a coarse version of X defined as

$$X^*(X) = p \text{ if } q_{p-0.1} < X \leq q_p,\ p \in \{0.1, 0.2, ..., 1\},$$

where q_p is the $100p$-percentile of the empirical distribution of X. Treating X^* as an ordinal variable we fitted the model

$$\text{logit}Pr(Z = 1|X^* = p, C) = \nu + \delta_p + \omega^T C.$$

If we assume that $\eta(x', C) = \eta(x, C)$ if $X^*(x') = X^*(x)$, then the testable criterion in (22) translates to $\delta_p \geq \delta_{p'}$ if $p \geq p'$. Figure 2 displays $\hat{\delta}_p$, $p \in \{0.1, 0.2, ..., 1\}$, together with the corresponding 95% confidence intervals. There is a strong trend, which is consistent with the assumptions. There are a few exceptions to monotonically increasing values of δ_p ($\hat{\delta}_{0.3}, \hat{\delta}_{0.5}, \hat{\delta}_1$), which may indicate a violation of the assumptions, but may also be explained by sampling variability.

In a preliminary analysis, we fitted the following standard model

$$E(Y|Z = 1, X, \text{Age, BMI}; \xi) = \xi_0 + \xi_1 X + \xi_2 \text{Age} + \xi_3 \text{BMI}.$$

We obtained $\hat{\xi}_1 = -0.018$, with 95% Wald confidence interval (-0.031,-0.004). To obtain an estimate of $m(x, C)$, using the implicit method of Section 4.2.1, we additionally fitted the model

$$\text{logit}Pr(Z = 1|X, \text{Age, BMI}; \alpha) = \alpha_0 + \alpha_1 X + \alpha_2 \text{Age} + \alpha_3 \text{BMI}. \quad (25)$$

We obtained $\hat{\alpha}_1 = 0.21$. Using an identity link, $g(\cdot)$, we replaced ξ_1 and α_1 in (23) with their estimates, $\hat{\xi}_1$ and $\hat{\alpha}_1$. We used the delta method to obtain 95%

DOI: 10.2202/1557-4679.1225

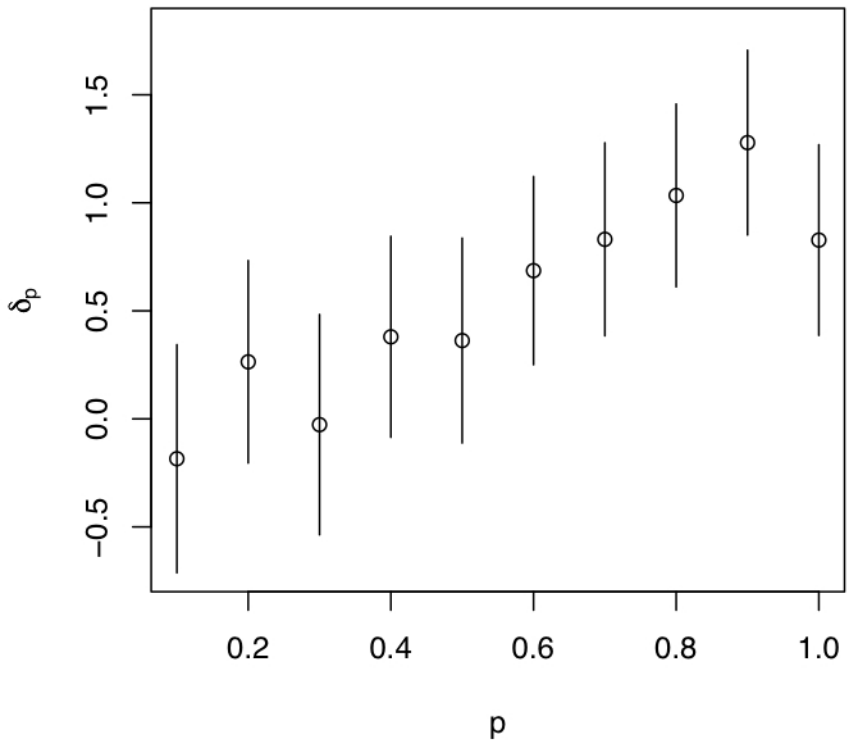

Figure 2: $\hat{\delta}_p$, $p \in \{0.1, 0.2, ..., 1\}$, together with 95% Wald confidence intervals.

point-wise Wald confidence limits for $m(x, C; \hat{\xi}, \hat{\alpha})$, for $x \in (0, 10)$. Figure 3 displays the result. As predicted by the theory, $\lim_{x \to 0} m(x, C; \hat{\xi}, \hat{\alpha}) = \hat{\xi}_1 / \hat{\alpha}_1 = -0.086$, and $\lim_{x \to \infty} m(x, C; \hat{\xi}, \hat{\alpha})/x = \hat{\xi}_1 = -0.018$.

We next fitted a model for $m(x, C)$ directly, using the explicit method of Section 4.2.2. We used an identity link, $g(\cdot) = \cdot$, and assumed a linear main model

$$m(x, \text{Age, BMI}; \psi) = \psi_0 + \psi_1 x. \tag{26}$$

We assumed the logistic model for $\pi(X, C)$ in (25), and the additional nuisance models

$$E(Y|Z = 1, X = 0, \text{Age, BMI}; \beta) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{BMI}, \tag{27}$$

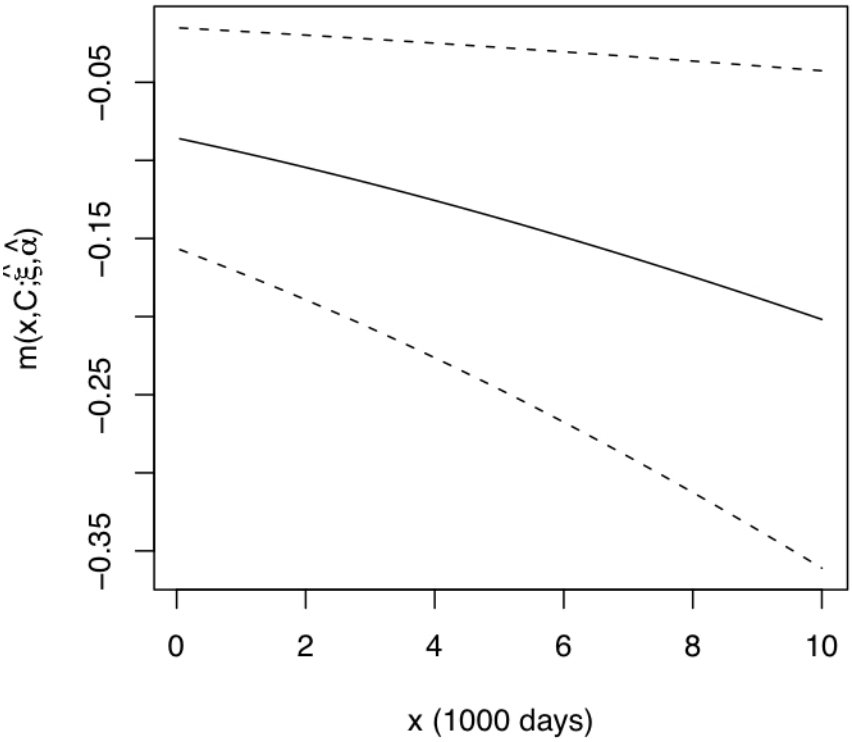

Figure 3: $m(x, C; \hat{\xi}, \hat{\alpha})$ with 95% point-wise Wald confidence limits.

$$\text{logit}Pr(X = 0|\text{Age, BMI}; \gamma) = \gamma_0^0 + \gamma_1^0\text{Age} + \gamma_2^0\text{BMI}, \tag{28}$$

$$Pr(X = x|X \neq 0, \text{Age, BMI}; \gamma) = \tau(\text{Age, BMI}; \gamma)\exp\{-\tau(\text{Age, BMI}; \gamma)x\}, \tag{29}$$

where

$$\log\tau(\text{Age, BMI}; \gamma) = \gamma_0^1 + \gamma_1^1\text{Age} + \gamma_2^1\text{BMI}. \tag{30}$$

The model for $Pr(X|\text{Age, BMI})$ is a mixture of a point mass in $X = 0$, and an exponential decay model on the range $(0, \infty)$, which appeared to be consistent with the data. Fitting the main and the nuisance models to the CAHRES data, using the method proposed in Section 4.2.2 (with $d(X_i, C_i) = \tilde{d}^{opt}(X_i, C_i)$), gave the following estimate of ψ with 95% Wald confidence interval:

$$\hat{\psi}_0 = -0.197(-0.483, 0.088), \;\; \hat{\psi}_1 = 0.015(-0.048, 0.079). \tag{31}$$

DOI: 10.2202/1557-4679.1225

Comparing the implicit model and the explicit model shows that $\hat{\psi}_0 < \hat{\xi}_1/\hat{\alpha}_1$. Thus, the explicit model indicates a larger non-HRT-dependent difference in prognosis between women with different tumor subtypes, than the implicit model. The explicit model, however, indicates a positive x-gradient in $m(x, C)$ ($\hat{\psi}_1 > 0$), whereas the implicit model indicates a negative x-gradient ($\hat{\xi}_1 < 0$). The estimates obtained from the explicit model are more uncertain than the estimates obtained from the implicit model. In particular, the confidence intervals for (ψ_0, ψ_1) both contain 0 (no difference), whereas the confidence intervals for $(\xi_1/\alpha_1, \xi_1)$ do not contain 0. The larger uncertainty in $(\hat{\psi}_0, \hat{\psi}_1)$ is the price we pay for using an approximately doubly robustness estimation method of the parameters indexing the explicit model. To summarize, our analysis indicate that women with HRT-induced tumors have a better prognosis than women with non-HRT-induced tumors. It is not clear, however, whether this discrepancy becomes more or less attenuated at higher levels of HRT.

6 Simulation study

The modified explicit estimation method described in Section 4.2.2 relies heavily on the rare disease assumption, A5. To investigate the sensitivity to this assumption, we performed a small simulation study. We let C consist of two components, following the joint empirical distribution of age and BMI for the controls in the CAHRES study. We let $Pr(X|C)$ follow the model in (28) and (29), and $Pr(Z|X, C)$ follow the model in (25). We let $Pr(Y|Z = 1, X, C)$ follow a normal distribution with standard deviation equal to 0.1. $E(Y|Z = 1, X, C)$ was defined implicitly as follows. We let $m(x, C)$ and $E(Y|Z = 1, X = 0, C)$ follow the models in (26) and (27), respectively. Assuming that A1-A4 holds, we used the relation in (4) to translate the models for $m(x, C)$, $Pr(Z|X, C)$, and $E(Y|Z = 1, X = 0, C)$ into a model for $E(Y|Z = 1, X, C)$. For this purpose, we used an identity link, $g(\cdot) = \cdot$. The value of $(\alpha, \beta, \gamma, \psi)$ was set to the corresponding ML-estimate from the CAHRES data, with exception for α_0 as described below. To simulate deviations from A5, we calculated the values for α_0 yielding a marginal disease prevalence, $Pr(Z = 1)$, of 0.01, 0.05, and 0.1, respectively. For each value of α_0, 1000 samples, each with $n_1 = 1000$ cases and $n_0 = 3000$ controls, were drawn from the joint model. For each sample, we applied the estimation procedure described in Section 4.2.2 (with $d(X_i, C_i) = \tilde{d}^{opt}(X_i, C_i)$), fitting the models (26)-(30) to the data. Table 6A displays the mean (over the 1000 samples) point estimate, mean standard error (as calculated from the sandwhich formula), and the coverage probability of the

corresponding 95% Wald confidence interval, for $Pr(Z = 1) \in \{0.01, 0.05, 0.1\}$. We observe that the method works well for $Pr(Z = 1) = 0.01$ (low bias and coverage probability close to the nominal level). When the disease prevalence increases, the performance decreases.

Theory states that $\hat{\psi}$ is approximately consistent when either $d(X_i, C_i) = \tilde{d}^{opt}(X_i, C_i)$, so that $E\{d(X_i, C_i)|Z_i = 1, C_i\} \approx 0$, or $q(C_i) = q^{opt}(C_i)$. To verify this property we repeated the simulation procedure described above with $d(X_i, C_i) = \tilde{G}_i$ and $q(C_i) = q^{opt}(C_i)$. For this choice, $E\{d(X_i, C_i)|Z_i = 1, C_i\} \neq 0$. We then repeated the procedure with $d(X_i, C_i) = \tilde{d}^{opt}(X_i, C_i)$ and $q(C_i) = 0 \neq q^{opt}(C_i)$. The results are displayed in Table 6B and 6C, respectively. We observe that the method works well even though $E\{d(X_i, C_i)|Z_i = 1, C_i\} \neq 0$; the results in Table 6B are similar to those in Table 6A. When $q(C_i) = 0$, however, the performance decreases dramatically. In particular, $\hat{\psi}_0$ is biased upwards and $\hat{\psi}_1$ is biased downwards. Since the method does not work well even at the lowest level of prevalence, we did not carry out the simulation for higher levels. Additional simulations showed, that when $q(C_i) = 0$ and $d(X_i, C_i) = \tilde{d}^{opt}(X_i, C_i)$ the method does only perform acceptably when the sample size is very large. Table 6D shows the result for $n_1 = n_0 = 10000$.

7 Discussion

In this paper we have used the framework of principal stratification to assess the difference in prognosis between women with cancers caused by HRT and women with cancers caused by other factors. We have proposed two estimation methods for the parameters indexing a model for this difference in prognosis.

One possible extension of our approach would be to include information on relevant tumor characteristics, such as estrogen receptor (ER) status (i.e. to define Z/cancer more finely) in order to better identify the HRT-induced and non-HRT-induced subtypes. An association between HRT and prognosis is likely to be, in part, but not completely, explained by measurable tumor characteristics. Estrogen usage is, for example, associated with the hormone receptor status of primary breast cancer (Collins et al, 2005; Lower et al, 1999) - breast cancers in HRT users are significantly more likely to be estrogen receptor (ER) positive than they are in non-users. Moreover, ER-positive breast cancer is known to have a significantly better prognosis than its ER-negative counterpart (in part due to the fact that ER-positive cancers respond to anti-estrogen therapies). Although extending our approach to incorporate ER status might lead to improvements in parameter estimation efficiency, even without such an extension, the principal stratification approach described in

DOI: 10.2202/1557-4679.1225

		$_0$ (= -0.197)			$_1$ (= 0.011)		
	$Pr(Z=1)$	mean($\hat{\psi}_0$)	mean(se)	cp	mean($\hat{\psi}_1$)	se	cp
A							
	0.01	-0.194	0.043	0.943	0.011	0.006	0.945
	0.05	-0.189	0.044	0.937	0.010	0.007	0.951
	0.1	-0.178	0.045	0.893	0.010	0.007	0.933
B							
	0.01	-0.196	0.044	0.951	0.011	0.007	0.956
	0.05	-0.192	0.045	0.915	0.011	0.007	0.932
	0.1	-0.175	0.045	0.889	0.010	0.007	0.917
C							
	0.01	-0.141	0.379	0.956	-0.003	0.070	0.952
D							
	0.01	-0.180	0.136	0.964	0.008	0.026	0.961

Table 6: Mean estimate, mean standard error (se), and coverage probability (cp) of the corresponding 95% Wald confidence interval. A: $n_1 = 1000$, $n_0 = 3000$, $d(X_i, C_i) = \tilde{d}^{opt}(X_i, C_i)$, $q(C_i) = q^{opt}(C_i)$; B: $n_1 = 1000$, $n_0 = 3000$, $d(X_i, C_i) = \tilde{G}_i$, $q(C_i) = q^{opt}(C_i)$; C: $n_1 = 1000$, $n_0 = 3000$, $d(X_i, C_i) = \tilde{d}^{opt}(X_i, C_i)$, $q(C_i) = 0$; D: $n_1 = n_0 = 10000$, $d(X_i, C_i) = \tilde{d}^{opt}(X_i, C_i)$, $q(C_i) = 0$

this article, for describing the HRT-prognosis association is still interesting, essentially because there is not a 1:1 mapping of ER status and hormone/non-hormone induced cancer. Moreover, it is not immediately obvious how to extend the semi-parametric approach described herein, in these terms.

The approach has other possible uses. Consider for example the new vaccines against human papillomavirus (HPV). It is well established that HPV is a necessary (but not sufficient) cause of cervical cancer (Walboomers et al, 1999). Thus, a vaccination against HPV may also prevent cervical cancer. Some women, however, may get infected even though they are vaccinated (we call them 'doomed'). For ethiological reasons we may want to compare the prognosis in cervical cancer for these women against the prognosis for those who get infected only if they are not vaccinated (we call them 'sensitive'). Our method can be used to estimate the mean difference in prognosis for these types of women, provided that the vaccine has no effect on the cancer prognosis for the doomed. Given that the vaccine is target against HPV and not cervical cancer *per se*, this assumption may be reasonable to make.

Our work is closely related to a number of recent publications on post-

treatment selection bias (Zhang and Rubin, 2003; Gilbert et al, 2003; Jemiai, 2005; Jemiai et al, 2007; Shepherd et al, 2006, 2007; Sjolander et al, 2009). These papers have focused on a scenario where each study participant receives one level of a binary treatment. After treatment, some subjects experience a particular event and, for this subgroup, an outcome is measured. The estimand of interest is the treatment effect for those subjects who would experience the event, regardless of whether they are treated or not. Using the terminology from Section 1, we may call those subjects 'doomed'. Since the 'doomed' are only observable in the untreated arm (see Section 1), this principal stratum effect is not in general identified. Zhang and Rubin (2003) derived bounds for the effect, and Jemiai (2005) proposed a sensitivity analysis. In this sensitivity analysis, the difference in outcome distributions for doomed and harmed is quantified in a selection bias parameter. This parameter is varied over a range of plausible values, and each value is mapped into one value for the principal stratum effect. The connection to our work is clear if we think about the treatment as HRT, the post-treatment event as 'cancer', and the outcome as 'prognosis'. Our work is different, however, in that we a priori assume the treatment effect for the doomed to be zero (A4). Our scientific interest lies instead in the selection bias parameter, $m(x, C)$, which, under the assumption of no effect for the doomed (A4), is identified.

A Proof of (1) and (3).

$$\begin{aligned}\pi(x,C) &= Pr\{Z(x)=1|X=x,C\} \\ &= Pr(R\leq x|X=x,C) \\ &= Pr(R\leq x|C).\end{aligned}$$

The first equality follows from A1, the second from A2, and the third from A3.

$$\begin{aligned}&Pr(Y|Z=1,X=x,C) \\ &= Pr\{Y(x)|Z(x)=1,X=x,C\} \\ &= Pr\{Y(x)|R\leq x,X=x,C\} \\ &= Pr\{Y(x)|R\leq x,C\} \\ &= Pr\{Y(x)|R=0,C\}\frac{\pi(0,C)}{\pi(x,C)} + Pr\{Y(x)|0<R\leq x,C\}\left\{1-\frac{\pi(0,C)}{\pi(x,C)}\right\}\end{aligned} \tag{32}$$

DOI: 10.2202/1557-4679.1225

The first equality follows from A1, the second from A2, the third from A3, and the fourth from Bayes rule together with (1).

B Proof that A1-A4 do not imply any other restrictions on $Pr(Y, Z, X, C)$ than (2) and (5).

The proof follows the structure of Appendix A in Jemiai et al (2007). Consider an arbitrary joint distribution $Pr^*(Y, Z, X, C)$ satisfying the restrictions in (2) and (5). To prove that (2) and (5) are the only restrictions implied by A1-A4, we must be able to construct a joint distribution $Pr(Y(\cdot), Z(\cdot), Y, Z, X, C)$ which satisfies A1-A4 and marginalizes to $Pr^*(Y, Z, X, C)$. We do this in the absence of covariates C since the construction can be repeated within levels of C. We construct the candidate distribution as follows:

1. $Pr(Y, Z, X) \equiv Pr^*(Y, Z, X)$.
2. Given $(Y, Z, X = x)$, $\{Y(x), Z(x)\} \equiv (Y, Z)$.
3. We impose $Z(x) \geq Z(x')$ if $x \geq x'$, and define R as the minimum level x for which $Z(x) = 1$.
4. From step 2 and step 3 it follows that
$$Pr(R \leq r|Y = y, Z = z, X = x) = \begin{cases} 0 & \text{if} \quad z = 0 \text{ and } r \leq x \\ 1 & \text{if} \quad z = 1 \text{ and } r > x \end{cases}$$
5.
$$\begin{aligned} &Pr(R \leq r|Y = y, Z = z, X = x) \\ &\equiv \begin{cases} \frac{a(r,x)}{a(\infty,x)} & \text{if} \quad z = 0 \text{ and } r > x \\ b(y,x) + \frac{a(r,0)}{a(x,0)}\{1 - b(y,x)\} & \text{if} \quad z = 1 \text{ and } r \leq x \end{cases} \end{aligned}$$

where

$$a(r,x) \equiv \begin{cases} Pr^*(Z = 1|X = r) - Pr^*(Z = 1|X = x) & \text{if} \quad r \in [0, \infty) \\ 1 - Pr^*(Z = 1|X = x) & \text{if} \quad r = \infty \end{cases}$$

and

$$b(y,x) \equiv \frac{Pr^*(Y = y, Z = 1|X = 0)}{Pr^*(Y = y, Z = 1|X = x)}.$$

The cumulative distribution defined in this manner is valid (i.e. monotonically increasing with r, and attaining 1 at $r = \infty$) due to (2) and (5).

6. Given $R = 0$, $Y(x) \equiv Y(x')$, $\forall x, x'$.

7. From step 2 and 6 it follows that

$$Pr\{Y(x') = y' | R = r, Y = y, Z, X = x\} = \begin{cases} 1 & \text{if } (x' = x \text{ or } r = 0) \text{ and } y' = y \\ 0 & \text{if } (x' = x \text{ or } r = 0) \text{ and } y' \neq y \end{cases}$$

8.

$$Pr\{Y(x') = y' | R = r, Y = y, Z, X = x\} \equiv Pr\{Y(x') = y' | R = r, X = x'\} \quad \text{if } x' \neq x \text{ and } r > 0. \tag{33}$$

Note that the right-hand-side of (33) is defined implicitly by step 1, 4, 5, and 7.

Let $Y_{-(x,x')}(\cdot)$ denote the function $Y(\cdot)$, except the points $Y(x)$ and $Y(x')$. Step 1 through 8 defines $Pr(R, Y, Z, X)$, $Pr\{Y(\cdot)|R = 0, Y, Z, X\}$, and $Pr\{Y(x')|R = r, Y, Z, X = x\}$, for all x, x' and $r > 0$. We finally allow $Pr\{Y_{-(x,x')}(\cdot)|Y(x'), R = r, Y, Z, X = x\}$ to be any proper distribution for all x, x' and $r > 0$. That the distribution $Pr(Y(\cdot), Z(\cdot), Y, Z, X, C)$ marginalizes to $Pr^*(Y, Z, X)$ follows from step 1. That A1 is satisfied follows from step 2. That A2 is satisfied follows from step 3. That A4 is satisfied follows from step 6. It remains to show that A3 is satisfied. We first show that $R \amalg X$. We have that

$$\begin{aligned} & Pr(R \leq r | X = x) \qquad (34) \\ &= \int_{y,z} Pr(R \leq r | Y = y, Z = z, X = x) Pr(Y = y, Z = z | X = x) dy dz \\ &= \int_y I(r \leq x) \times 0 \times Pr^*(Y = y, Z = 0 | X = x) dy \\ &+ \int_y I(r > x) \frac{a(r, x)}{a(\infty, x)} Pr^*(Y = y, Z = 0 | X = x) dy \\ &+ \int_y I(r \leq x) \left[b(y, x) + \frac{a(r, 0)}{a(x, 0)} \{1 - b(y, x)\} \right] Pr^*(Y = y, Z = 1 | X = x) dy \\ &+ \int_y I(r > x) \times 1 \times Pr^*(Y = y, Z = 1 | X = x) dy \\ &= Pr^*(Z = 1 | X = r). \qquad (35) \end{aligned}$$

DOI: 10.2202/1557-4679.1225

Hence, $R \amalg X$.

We now show that $Y(x) \amalg X|R$. We first consider the case when $R = 0$. In this case we have that

$$\begin{aligned}
&Pr\{Y(x') = y'|R = 0, X = x\} \\
&= \int_{y,z} \Big[Pr\{Y(x') = y'|R = 0, Y = y, Z = z, X = x\} \\
&\qquad \times Pr(R = 0, Y = y, Z = z|X = x) dy dz \Big] / Pr(R = 0|X = x) \\
&= \Big\{ Pr(R = 0|Y = y', Z = 0, X = x) Pr(Y = y', Z = 0|X = x) \\
&\quad + Pr(R = 0|Y = y', Z = 1, X = x) Pr(Y = y', Z = 1|X = x) \Big\} \\
&\quad / Pr(R = 0|X = x) \\
&= \frac{0 \times Pr^*(Y = y', Z = 0|X = x) + b(y', x) Pr^*(Y = y', Z = 1|X = x)}{Pr(R = 0|X = x)} \\
&= Pr^*\{Y = y'|Z = 1, X = 0\},
\end{aligned}$$

where the last equality follows, since from (34), $Pr(R = 0|X = x) = Pr^*(Z = 1|X = 0)$. Hence, $Y(x) \amalg X|R = 0$. When $R > 0$ it follows immediately from step 8 that $Y(x) \amalg (Y, Z, X)|R$. Hence, $Y(x) \amalg X|R = r$, $\forall r$, which concludes the proof.

C Proof that the estimating equation in (13) is unbiased and doubly robust.

When $g(\cdot)$ is the identity link and $H(\psi)$ is defined as in (11), we have that

$$\begin{aligned}
E(Y|Z = 1, X = 0, C) &= E(Y|Z = 1, X, C) - m(X, C; \psi) \left\{ 1 - \frac{\pi(0, C)}{\pi(X, C)} \right\} \\
&= E\{H(\psi)|Z = 1, X, C\}, \qquad (36)
\end{aligned}$$

where the first equality follows from (4), and the second from the definition of $H(\psi)$. When $g(\cdot)$ is the log link, equality between $E(Y|Z = 1, X = 0, C = c)$ and $E\{H(\psi)|Z = 1, X, C\}$ holds when $H(\psi)$ is defined as in (12). Since $E(Y|Z = 1, X = 0, C)$ is not a function of x we have that

$$E\{H(\psi)|Z = 1, X, C\} = E\{H(\psi)|Z = 1, C\}. \qquad (37)$$

An immediate consequence is that

$$\begin{aligned} E\{d(X,C)H(\psi)|Z=1,C\} &= E[E\{d(X,C)H(\psi)|Z=1,X,C\}|Z=1,C] \\ &= E\{d(X,C)|Z=1,C\}E\{H(\psi)|Z=1,C\}, \end{aligned} \tag{38}$$

where the second equality follows from (37). Now, rewrite

$$\begin{aligned} &E[d(X,C)Z\{H(\psi)-q(C)\}] \\ &\quad = E\Big[E[d(X,C)\{H(\psi)-q(C)\}|Z=1,C]Pr(Z=1|C)\Big] \\ &\quad = E\Big[E\{d(X,C)|Z=1,C\}[E\{H(\psi)|Z=1,C\}-q(C)]Pr(Z=1|C)\Big], \end{aligned} \tag{39}$$

where the second equality follows from (38). The right handside of (39) equals 0 when either $E\{d(X,C)|Z=1,C\}=0$ or $q(C)=E\{H(\psi)|Z=1,C\}$. Combining (36) with (37) shows that $E\{H(\psi)|Z=1,C\}=E(Y|Z=1,X=0,C)$. Finally, since the estimation procedure that we propose for $d(X,C)$ does not rely on the model for $q(C)$ and vice versa (see Section 3.2.2), the doubly robustness follows.

D Derivation of the efficient index functions for the estimating equation in (13).

Let $\mathcal{M}$ be the model defined by A2-A4, models (9) and (14). Then it follows from Appendix B that, under assumption A1, the observed data laws allowed by model $\mathcal{M}$ are those satisfying

$$\begin{aligned} E(Z|X,C) &= \pi(X,C;\alpha) \\ E\{H(\psi,\alpha)|Z=1,X,C\} &= E\{H(\psi,\alpha)|Z=1,C\} \end{aligned} \tag{40}$$

where $H(\psi,\alpha)$ is defined like $H(\psi)$ but with $\pi(0,C)$ and $\pi(X,C)$ substituted with $\pi(X,C;\alpha)$ and $\pi(0,C;\alpha)$, respectively. The model can therefore be parameterized with the known function $\pi(X,C;\alpha)$, the unknown finite-dimensional parameters ψ and α, and infinite-dimensional parameters η_1 and η_2 indexing $f(Y|Z,X,C)$, which satisfies (40), and $f(X,C)$, respectively. Specifically, the likelihood of the observed data can be written as

$$f(Y|Z,X,C;\eta_1,\psi,\alpha)f(Z|X,C;\alpha)f(X,C;\eta_2)$$

DOI: 10.2202/1557-4679.1225

where the dependence of $f(Y|Z,X,C)$ on (ψ,α) is implied by (40). The nuisance tangent space for this model is $\Lambda_{nuis}=\Lambda_{1,nuis}+\Lambda_{2,nuis}$ where $\Lambda_{2,nuis}=\{a\left(X,C\right):E\left[a\left(X,C\right)\right]=0\}\cap L_2\left(P\right)$ is the closed linear span of all scores for parametric submodels for the joint law of (X,C) and

$$\begin{aligned}\Lambda_{1,nuis}=\{&a(H,Z,X,C):E[a(H,Z,X,C)|Z,X,C]=0,\\ &E[Ha(H,Z,X,C)|Z=1,X,C]=E[Hat(H,Z,X,C)|Z=1,C]\}\cap L_2(P)\end{aligned}$$

is the closed linear span of all scores for parametric submodels for the joint conditional law of $H\equiv H(\psi,\alpha)$, given (Z,X,C).

Denote $\sigma^2\left(X,C\right)\equiv\mathrm{Var}\left(H|Z=1,X,C\right)$. The orthocomplement to $\Lambda_{2,nuis}$ in the Hilbert space $L_2^0\left(P\right)$ (with covariance inner product) of functions in $L_2\left(P\right)$ with mean zero is $\Lambda_{2,nuis}^{\perp}=\{d(H,Z,X,C):E\left[d(H,Z,X,C)|X,C\right]=0\}\cap L_2^0\left(P\right)$ because the orthogonal projection of an arbitrary function $e(H,Z,X,C)$ in $L_2^0\left(P\right)$ onto $\Lambda_{2,nuis}$ is $E\left\{e(H,Z,X,C)|X,C\right\}$. Let $d(H,Z,X,C)$ be an arbitrary function, and define

$$\begin{aligned}K\left(H,d\right) &= \mathrm{Cov}\left(H,d\left(H,Z,X,C\right)|Z=1,X,C\right),\\ J\left(H,d\right) &= K\left(H,d\right)\\ &\quad -E\left[\sigma^{-2}\left(X,C\right)|Z=1,C\right]^{-1}E\left[\sigma^{-2}\left(X,C\right)K\left(H,d\right)|Z=1,C\right].\end{aligned}$$

The orthocomplement to $\Lambda_{1,nuis}$ in $L_2^0\left(P\right)$ is then

$$\begin{aligned}\Lambda_{1,nuis}^{\perp}=\{&E[d(H,Z,X,C)|Z,X,C]\\ &+\sigma^{-2}(X,C)J(H,d)Z[H-E(H|Z=1,C)]\}\cap L_2^0\left(P\right),\end{aligned}$$

since the orthogonal projection of an arbitrary function $e(H,Z,X,C)$ in $L_2^0\left(P\right)$ onto $\Lambda_{1,nuis}$ is

$$\begin{aligned}&e(H,Z,X,C)-E\left[e(H,Z,X,C)|Z,X,C\right]\\ &\quad-\sigma^{-2}\left(X,C\right)J\left(H,e\right)Z\left[H-E(H|Z=1,C)\right].\end{aligned}$$

Because $\Lambda_{1,nuis}$ and $\Lambda_{2,nuis}$ are mutually orthogonal, the orthogonal projection of an arbitrary function $e(H,Z,X,C)$ in $L_2^0\left(P\right)$ onto $\Lambda_{nuis}=\Lambda_{1,nuis}+\Lambda_{2,nuis}$ as the sum of the separate projections onto $\Lambda_{1,nuis}$ and $\Lambda_{2,nuis}$, so that

$$\begin{aligned}\Lambda_{nuis}^{\perp} &= \{E\left[d(H,Z,X,C)|Z,X,C\right]-E\left[d(H,Z,X,C)|X,C\right]\\ &\quad+\sigma^{-2}\left(X,C\right)J\left(H,d\right)Z\left[H-E(H|Z=1,C)\right]\}\cap L_2^0\left(P\right)\end{aligned}$$

To find the efficient scores for ψ and α, note from the model formulation that the scores S_ψ for ψ and S_α for α under parametric submodels satisfy the following restrictions

$$E\left(S_\psi|Z=1,X,C\right)=0$$

$$E\left(HS_{\psi}|Z=1,X,C\right) = E\left(HS_{\psi}|Z=1,C\right)+E\left(\frac{\partial H}{\partial\psi}|Z=1,C\right)-E\left(\frac{\partial H}{\partial\psi}|Z=1,X,C\right) \tag{41}$$

$$\begin{aligned} S_{\alpha} &= S_{\alpha}^{*}+S_{\alpha}^{**} \\ S_{\alpha}^{**} &= \pi(X,C;\alpha)^{-1}\left\{1-\pi(X,C;\alpha)\right\}^{-1}\left\{Z-\pi(X,C;\alpha)\right\}\partial\pi(X,C;\alpha)/\partial\alpha \end{aligned}$$

$$\begin{aligned} E\left(HS_{\alpha}^{*}|Z=1,X,C\right) &= E\left(HS_{\alpha}^{*}|Z=1,C\right)+E\left(\frac{\partial H}{\partial\alpha}|Z=1,C\right) \\ &- E\left(\frac{\partial H}{\partial\alpha}|Z=1,X,C\right)+\mathrm{Cov}(H,S_{\alpha}^{**}|Z=1,C) \end{aligned}$$

where $S_{\alpha}^{*}\equiv\partial\log f(Y|Z,X,C;\eta_1,\psi,\alpha)/\partial\alpha$, $S_{\alpha}^{**}=\partial\log f(Z|X,C;\alpha)/\partial\alpha$ and

$$\begin{aligned} \mathrm{Cov}(H,S_{\alpha}^{**}|Z=1,C) &= \mathrm{Cov}\{E(H|Z=1,C),S_{\alpha}^{**}|Z=1,C\} \\ &+E\{\mathrm{Cov}(H,S_{\alpha}^{**}|Z=1,X,C)\}=0 \end{aligned}$$

The efficient score for ψ is now obtained as the orthogonal projection of S onto $\Lambda_{nuis}^{\perp}$ which, from previous results, equals

$$\sigma^{-2}\left(X,C\right)J_{\mathrm{eff},\psi}\left(H,d\right)Z\left[H-E(H|Z=1,C)\right]$$

where $J_{\mathrm{eff},\psi}\left(H,d\right)$ is defined like $J\left(H,d\right)$, but with $K\left(H,d\right)$ replaced with $-E\left[\partial H(\psi,\alpha)/\partial\psi|Z=1,X,C\right]$. This can be seen because $\mathrm{Cov}(H,S_{\psi}\ |Z=1,X,C)=E\left(HS_{\psi}|Z=1,X,C\right)$, which is given by (41), where the first two terms in the righthand side of (41) do not contribute to the expression for $J(H,S_{\psi})$ because they are functions of only C. Likewise, the efficient score for α is obtained as the orthogonal projection of S_{α} onto $\Lambda_{nuis}^{\perp}$ which, from previous results, equals

$$\begin{aligned} &E(S_{\alpha}|Z,X,C)-E(S_{\alpha}|X,C) \\ &\quad+\sigma^{-2}\left(X,C\right)J_{\mathrm{eff},\alpha}\left(H,d\right)Z\left[H-E(H|Z=1,C\right] \\ &=S_{\alpha}^{**}+\sigma^{-2}\left(X,C\right)J_{\mathrm{eff},\alpha}\left(H,d\right)Z\left[H-E(H|Z=1,C)\right] \end{aligned}$$

where $J_{\mathrm{eff},\alpha}\left(H,d\right)$ is defined like $J\left(H,d\right)$, but with $K\left(H,d\right)$ replaced with $-E\left[\partial H(\psi,\alpha)/\partial\alpha|Z=1,X,C\right]$.

Consider now the model $\mathcal{M}^{*}$ defined by A2-A4, (9), (14) and (19). Then a similar development shows that the efficient scores for ψ and α are as given above, and that the efficient score for γ equals $\partial\log f(X|C;\gamma)/\partial\gamma$.

DOI: 10.2202/1557-4679.1225

E Derivation of the asymptotic variance for $\hat{\theta}$ under case-control sampling.

The proof follows closely the proof in Prentice and Pyke (1979). A first-order Taylor expansion of $\sum_{i=1}^{n} U_i(\hat{\theta}) = 0$ about the true value θ gives

$$\sum_{i=1}^{n} U_i(\hat{\theta}) = \sum_{i=1}^{n} U_i(\theta) + \frac{\partial \sum_{i=1}^{n} U_i(\theta)}{\partial \theta}(\hat{\theta} - \theta) = 0, \tag{42}$$

or

$$n^{1/2}(\hat{\theta} - \theta) = \left\{ -n^{-1} \sum_{i=1}^{n} \frac{\partial U_i(\theta)}{\partial \theta} \right\}^{-1} n^{-1/2} \sum_{i=1}^{n} U_i(\theta). \tag{43}$$

According to the law of large numbers, $n^{-1} \sum_{i=1}^{n} \frac{\partial U_i(\theta)}{\partial \theta}$ converges almost surely to $\mathcal{I}$. We can rewrite $n^{-1/2} \sum_{i=1}^{n} U_i(\theta)$ as

$$\begin{aligned} & n^{-1/2} \sum_{i=1}^{n} U_i(\theta) \\ & = \sum_{z \in \{0,1\}} (n_z/n)^{1/2} \left[n_z^{-1/2} \sum_{i=1}^{n} I(Z_i = z)\{U_i(\theta) - \mu_z\} \right] + n^{-1/2} \sum_{z \in \{0,1\}} n_z \mu_z, \end{aligned}$$

where $\mu_z = E\{U_i(\theta)|Z = z\}$. The central limit theorem can be applied to the terms in square brackets. Also, it follows from results in Prentice and Pyke (1979), that $\sum_{z \in \{0,1\}} n_z \mu_z = 0$. Hence, $n^{-1/2} \sum_{i=1}^{n} U_i(\theta)$ is asymptotically normal with mean 0 and variance $\mathcal{J}$.

F Proof that $Pr(X|Z = 1, C)$ is variation independent of $\eta(X, C)$.

We carry out the proof in the absence of covariates, for brevity.

Define $Pr(Z = z, X = x) \equiv p(z, x)$. The function $p(z, x)$ is restricted by

$$0 \leq p(z, x) \ \ \forall z, x, \tag{44}$$

$$1 = \int p(1, x)dx + \int p(0, x)dx. \tag{45}$$

Thus, at a given value $z \in \{0, 1\}$, $p(z, x)$ is restricted by

$$0 \leq p(z, x) \ \ \forall x, \tag{46}$$

$$\int p(z,x)dx \leq 1. \tag{47}$$

From Bayes rule we have that

$$\eta(x) \equiv \frac{\pi(x)\{1-\pi(0)\}}{\pi(0)\{1-\pi(x)\}} = \frac{p(1,x)p(0,0)}{p(1,0)p(0,x)}. \tag{48}$$

$\eta(x)$ is restricted by

$$\eta(0) = 1, \tag{49}$$

$$0 \leq \eta(x) < \infty \ \ \forall x. \tag{50}$$

$Pr(X|Z=1)$ is determined by the function $p(1,x)$. Thus, $Pr(X|Z=1)$ and $\eta(X)$ are variation independent if $p(1,X)$ and $\eta(X)$ are variation independent. Consider an arbitrary function $p^*(1,x)$ satisfying (46) and (47) with $z=1$, and an arbitrary function $\eta^*(x)$ satisfying (49) and (50). To prove that $p(1,X)$ and $\eta(X)$ are variation independent we must be able to construct a full law $p(z,x)$ which a) marginalizes to $p^*(1,z)$, b) is indexed by $\eta(x) = \eta^*(x)$, c) satisfies (44) and (45). We first define

$$p(1,x) \equiv p^*(1,x), \tag{51}$$

$$\eta(x) \equiv \eta^*(x). \tag{52}$$

From (48) we have that

$$p(0,x) = \frac{p(0,0)}{p(1,0)} p(1,x)\eta^{-1}(x). \tag{53}$$

Substituting into (45) and solving for $p(0,0)$ gives

$$p(0,0) = \frac{p(1,0)\{1-\int p(1,x)dx\}}{\int p(1,x)\eta^{-1}(x)dx}.$$

Substituting back into (53) gives

$$p(0,x) = \left\{1 - \int p(1,x)dx\right\} \frac{p(1,x)\eta^{-1}(x)}{\int p(1,x)\eta^{-1}(x)dx}. \tag{54}$$

The law defined by (51), (52) and (54) satisfies a), b), and c) listed above. Thus, $p(1,X)$ and $\eta(X)$ are variation independent, which implies that $Pr(X|Z=1)$ and $\eta(X)$ are variation independent. In addition, we observe that the law defined by (51), (52) and (54) is the only possible candidate for $p(z,x)$. Thus, as a byproduct of the proof we get that $p(1,X)$ and $\eta(X)$ together fully specify the joint distribution $p(Z,X)$.

DOI: 10.2202/1557-4679.1225

References

Antoine C, Liebens F, Carly B, Pastijn A, Rozenberg S. (2004). Influence of HRT on prognostic factors for breast cancer: a systematic review after the Women's Health Initiative trial. *Human Reproduction* **19**(3), 741-756.

Bang H and Robins JM. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **58**, 962-973.

Collaborative Group on Hormonal Factors in Breast Cancer. (1997). Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. *Lancet* **350**, 1047-1059.

Collins JA, Blake JM, Crosignani PG. (2005). Breast cancer risk with postmenopausal hormonal treatment. *Human Reproduction Update* **11**, 545-560.

Easton DF, Pooley KA, Dunning AM, Pharoah DP et al. (2007). A genome-wide association study identifies multiple breast cancer susceptibility loci. *Nature* **447**, 1087-1093.

Frangakis CE, Rubin DB. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21-29.

Galea MH, Blamey RW, Elston CE, Ellis IO. (1992). The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Research and Treatment* **22**, 207-219.

Gilbert PB, Bosch JB, Hudgens MG. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* **59**, 531-541.

Jemiai Y. *Semiparametric methods for the effect of treatment on an outcome existing only in a post-randomization selected subpopulation.* (2005). Ph.D. Thesis, Harvard University, Cambridge, Massachusetts.

Jemiai Y, Rotnitzky A, Shepherd BE, Gilbert PB. (2007). Semiparametric estimation of treatment effects on an outcome measured after a post-randomization event occurs. *Journal of the Royal Statistical Society, Series B* **69**, 879-901.

Lower EE, Blau R, Gazder P, Stahl DL. (1999). The effect of estrogen usage on the subsequent homone receptor status of primary breast cancer. *Breast Cancer Res Treat* **58**, 205-211.

Magnusson C, Baron JA, Correia N, Bergström R, Adami HO, Persson I. (1999). Breast cancer risk following long-term oestrogen- and oestrogen-progestin-replacement therapy. *Int. J. Cancer* **81**, 339-344.

Million Women Study Collaborators. (2003). Breast cancer and hormone-replacement therapy in the Million Women Study. *The Lancet* **362**(9382) 419-427.

Nanda K, Basitan LA, Schulz K. (2002). Hormone replacement therapy and the risk of death from breast cancer: a systematic review. *American Journal of Obstetrics and Gynecology* **186**, 325-334.

Prentice RL, Pyke R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403-411.

Robins JM, Rotnitzky A. (2001). Comment on 'Inference for semiparametric models: Some questions and an answer'. *Statistica Sinica* **11**(4), 920-936.

Rosenberg L. (2006). *Hormone-related factors and breast cancer - studies of risk and prognosis*. PhD thesis, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm.

Rosenberg LU, Granath F, Dickman PW, Einarsdottir K, Wedren S, Persson I, Hall P. (2008). Menopausal hormone therapy in relation to breast cancer characteristics and prognosis: a cohort study. *Breast Cancer Research* **10**:r78.

Shepherd BE, Gilbert PB, Jemiai Y, Rotnitzky A. (2006). Sensitivity analysis comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics* **62**, 332-342.

Shepherd BE, Gilbert PB, Lumley T. (2007). Sensitivity analysis comparing time-to-event outcomes existing only in a subset selected postrandomization. *Journal of the American Statistical Association* **102**, 573-582.

Sjölander A, Humphreys K, Vansteelandt S, Bellocco R, Palmgren J. (2009). Sensitivity analysis for principal stratum direct effects, with an application to a study of physical activity and coronary heart disease. *Biometrics* **65**, 514-520.

van der Laan MJ. (2008). Estimation based on case-control designs with known incidence probability. *Working Paper 234*. Berkley Electronic Press, Berkley.

DOI: 10.2202/1557-4679.1225

Walboomers JM, Jacobs MV, Manos MM, et al (1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *The Journal of Pathology* **189**(1): 129.

Zhang JL, Rubin DB. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics* **28**(4), 353-368.