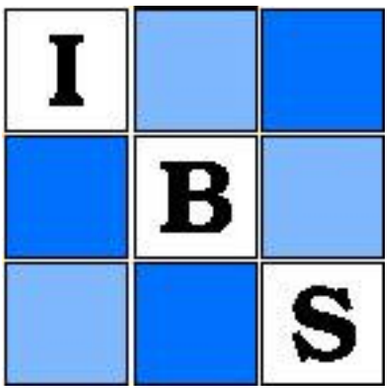

BIOMETRICS 65, 514–520
June 2009

DOI: 10.1111/j.1541-0420.2008.01108.x

Sensitivity Analysis for Principal Stratum Direct Effects, with an Application to a Study of Physical Activity and Coronary Heart Disease

Arvid Sjölander,[1,*] Keith Humphreys,[1] Stijn Vansteelandt,[2] Rino Bellocco,[1,3] and Juni Palmgren[1,4]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels väg 12, 17177 Stockholm, Sweden

[2]Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281, S9, 9000 Gent, Belgium

[3]Department of Statistics, University of Milano-Bicocca, Ed. U7, Via Bicocca degli Arcimboldi 8, 20126 Milan, Italy

[4]Department of Mathematical Statistics, Matematiska Instutionen, Stockholms Universitet, 10691 Stockholm, Sweden

*email: arvid.sjolander@meb.ki.se

SUMMARY. In many studies, the aim is to learn about the direct exposure effect, that is, the effect not mediated through an intermediate variable. For example, in circulation disease studies it may be of interest to assess whether a suitable level of physical activity can prevent disease, even if it fails to prevent obesity. It is well known that stratification on the intermediate may introduce a so-called posttreatment selection bias. To handle this problem, we use the framework of *principal stratification* (Frangakis and Rubin, 2002, *Biometrics* **58,** 21–29) to define a causally relevant estimand—the principal stratum direct effect (PSDE). The PSDE is not identified in our setting. We propose a method of sensitivity analysis that yields a range of plausible values for the causal estimand. We compare our work to similar methods proposed in the literature for handling the related problem of "truncation by death."

KEY WORDS: Direct effect; Potential outcomes; Principal stratification; Sensitivity analysis.

1. Introduction

Several cohort studies (Hu et al., 2004; Li et al., 2006; Mora et al., 2006) have recently been conducted to assess the association between physical activity (PA), obesity and circulation diseases, such as cardiovascular disease (CVD), and coronary heart disease (CHD). In these studies, self-reported PA level, body mass index (BMI), and baseline covariates have been measured for each subject at enrollment. The cohort has then been followed over time, and each disease event has been recorded. A common conclusion from these studies is that PA is negatively associated with circulation diseases, even within levels of BMI. One possible interpretation is that PA may prevent circulation diseases, even if it fails to prevent obesity. We may say that PA has a *direct* (not mediated through BMI reduction) effect on the risk for circulation diseases.

This causal interpretation may, however, not be valid. The reason is that rather than being an ordinary "baseline covariate," BMI is to a large extent affected by PA level—we may say that BMI "lies on the causal pathway" between PA and the disease outcome. It is well known that conditioning on such an intermediate variable may induce a special kind of bias known as posttreatment selection bias (Frangakis and Rubin, 2002). For example, PA may fail to prevent obesity for subjects with a genetic predisposition to being overweight. If predisposition for being overweight is correlated with predisposition to circulation diseases, then the disease prevalence among obese subjects will be shifted upward for those who are physically active, compared to those who are nonactive.

Frangakis and Rubin (2002) developed a framework for inference on causal effects adjusted for intermediate variables. In this framework, estimands are defined using *potential outcomes.* Within the present context, each subject would be considered to have a potential BMI and disease status, for each level of PA. The "intermediate-adjusted" causal effect is defined as a comparison of potential disease status across two levels of PA, for a group of subjects with the same potential BMI under both PA levels. Such a group is called a *principal stratum.* Within this principal stratum, BMI is not affected by PA, so any effect of PA on disease status must be "direct." We will thus refer to effects defined in this manner as *principal stratum direct effects* (PSDEs).

In general, principal strata are not directly observable, and PSDEs are not identified. In this article, we propose a method of sensitivity analysis for PSDEs based on pattern mixture modeling, which gives a range of plausible values. We will use data from an ongoing study, the Swedish "National March Cohort" (NMC), to illustrate the method.

Section 2 introduces basic definitions and assumptions. Section 3 presents the method of sensitivity analysis for the PSDE. Section 4 illustrates the method applied to the NMC data. Section 5 compares our work to methods proposed for handling the related problem of "truncation by death" (Zhang and Rubin, 2003; Jemiai, 2005).

Table 1
Principal stratification

	R			
	n	p	i	a
$Z(0)$	0	0	1	1
$Z(1)$	0	1	0	1

2. Definitions and Assumptions

2.1 *Potential Outcomes and Principal Stratification*

Consider a cohort study in which N subjects are enrolled. At enrollment, measurements of PA (X), BMI (Z), and important covariates (C) are collected. We assume that PA is dichotomized as "low level" ($X = 1$) versus "high level" ($X = 0$), and BMI as "obese" ($Z = 1$) versus "not obese" ($Z = 0$). Section 6 discusses possible extensions to nonbinary variables. The cohort is followed over time, and each CVD event is recorded. Let $Y = 1$ for subjects who report at least one CVD event before end of follow-up, and $Y = 0$ for subjects who remain undiagnosed through follow-up. We use $Pr(\cdot)$ generically for both probabilities and densities. We use uppercase letters for random variables, and lowercase letters for their outcomes. We assume that the data are an independent and identically distributed (i.i.d.) sample from $Pr(Y = y, Z = z, X = x, C = c)$. To define causal estimands we follow Frangakis and Rubin (2002) and use the framework of potential outcomes and principal stratification. We let $Z_i(x)$ and $Y_i(x)$ denote the potential outcome of Z and Y for subject i, at PA level $X = x$. The following consistency assumption relates the potential outcomes to the observables:

ASSUMPTION 1

$$Z_i(X_i) = Z_i; Y_i(X_i) = Y_i.$$

Assumption 1 states that the potential outcomes corresponding to the factual PA level for subject i are observed, and equal to the observed outcomes Z_i and Y_i. The potential outcomes corresponding to the other (counterfactual) activity level are unobserved, and are considered as missing.

We divide subjects into four principal strata, denoted by R. In the first stratum, $Z_i(x) = 0$, $\forall x$. This stratum contains subjects who will never become obese, regardless of whether they exercise or not. We use $R = n$ for this stratum. In the second stratum $Z_i(x) = x$ (PA prevents obesity; $R = p$). In the third stratum, $Z_i(x) = 1$, $\forall x$. This stratum consists of subjects who will always be obese, regardless of whether they exercise or not; $R = a$. In the fourth stratum $Z_i(x) = 1 - x$ (PA induces obesity; $R = i$). Table 1 summarizes the relation between x, $Z(x)$, and R. We use $R(x, z)$ to denote the set of R-values compatible with $Z(x) = z$. We define $\pi_{r\cdot c} \equiv Pr(R = r \mid C = c)$.

Although PA could hypothetically cause obesity for some subjects, this is unlikely and we therefore assume that principal stratum $R = i$ is empty:

ASSUMPTION 2

$$\pi_{i\cdot c} = 0$$

Assumption 2 implies that $R(0, 0) = (n, p)$, $R(0, 1) = a$, $R(1, 0) = n$, $R(1, 1) = (p, a)$.

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

ASSUMPTION 3

$$R \amalg X \mid C; Y(x) \amalg X \mid R, C$$

where "$\amalg$" is used to denote statistical independency.

Under Assumptions 1–3 the principal strata proportions are identified. Note that

$$\begin{aligned} Pr(Z = z \mid X = x, C = c) &= \sum_{r \in R(x,z)} Pr(R = r \mid X = x, C = c) \\ &= \sum_{r \in R(x,z)} \pi_{r\cdot c}. \end{aligned} \tag{1}$$

The first equality follows from Assumption 1 and the second equality from Assumption 3. Assumption 2 now yields:

$$\begin{aligned} \pi_{n\cdot c} &= Pr(Z = 0 \mid X = 1, C = c), \\ \pi_{a\cdot c} &= Pr(Z = 1 \mid X = 0, C = c), \\ \pi_{p\cdot c} &= 1 - \pi_{n\cdot c} - \pi_{a\cdot c} = 1 - Pr(Z = 0 \mid X = 1, C = c) \\ &\quad - Pr(Z = 1 \mid X = 0, C = c). \end{aligned} \tag{2}$$

The last row in equation (2) shows that Assumptions 1–3 can in principle be tested,[1] because it implies that $1 - Pr(Z = 0 \mid X = 1, C = c) - Pr(Z = 1 \mid X = 0, C = c) \geq 0$.

2.2 *The Principal Stratum Direct Effect*

Using the framework of principal stratification it is easy to demonstrate why stratification on Z does not yield a direct effect of X on Y. In such a "stratified analysis" we would compare $Pr(Y = y \mid Z = z, X = 0, C = c)$ with $Pr(Y = y \mid Z = z, X = 1, C = c)$. Under Assumptions 1–3 it can be shown that

$$\begin{aligned} &Pr(Y = y \mid Z = z, X = x, C = c) \\ &= \frac{\sum_{r \in R(x,z)} Pr\{Y(x) = y \mid R = r, C = c\} \pi_{r\cdot c}}{\sum_{r \in R(x,z)} \pi_{r\cdot c}}. \end{aligned} \tag{3}$$

From equation (3) we see that comparing $Pr(Y = y \mid Z = 1, X = 0, C = c)$ with $Pr(Y = y \mid Z = 1, X = 1, C = c)$, for example, is equivalent to comparing $Pr\{Y(0) = y \mid R = a, C = c\}$ with a mixture of $Pr\{Y(1) = y \mid R = a, C = c\}$ and $Pr\{Y(1) = y \mid R = p, C = c\}$. Thus, by stratifying on Z we are comparing the distribution of potential outcomes for two different groups of people. Such a comparison can only be given a causal interpretation if the two groups are "exchangeable," that is, if the distribution of potential outcomes $Y(x)$ is the same across levels of R (given C). This is, however, a very strong assumption, which is not likely to hold for our data. On the contrary, it is easy to imagine that a subject's predisposition for obesity is causally related to their predisposition for CVD.

[1] It can easily be shown, however, that there is no consistent test of Assumptions 1–3, i.e., a test for which power converges to 1 for each fixed alternative.

We define the PSDE of X on Y, for $R = r$, $r \in (n, a)$, as

$$\beta_{r\cdot c} \equiv g[E\{Y(1) \mid R = r, C = c\}] - g[E\{Y(0) \mid R = r, C = c\}], \tag{4}$$

where $g(\cdot)$ is a known, smooth, and monotonic link function. Natural link functions are the identity link that yields the risk difference, the log link that yields the log relative risk, and the logit link that yields the log odds ratio. The parameter $\beta_{r\cdot c}$ is a causal effect, because it is a comparison of potential outcomes for a single, well-defined, group of people. Moreover, $\beta_{r\cdot c}$ can be interpreted as a direct effect of X on Y within principal stratum $R = r$, because for all subjects within $R = r$, $r \in (n, a)$, Z is assured to be fixed, regardless of X.

The PSDEs are the target of our analysis. Unfortunately they are not identified without further assumptions. From equation (3), we have that $Pr\{Y(1) \mid R = n, C = c\}$ is identified and equal to $Pr(Y \mid Z = 0, X = 1, C = c)$. Similarly, $Pr\{Y(0) \mid R = a, C = c\}$ is identified and equal to $Pr(Y \mid Z = 1, X = 0, C = c)$. $Pr(Y \mid Z = 0, X = 0, C = c)$, however, is a mixture of $Pr\{Y(0) = y \mid R = n, C = c\}$ and $Pr\{Y(0) = y \mid R = p, C = c\}$. Similarly, $Pr(Y \mid Z = 1, X = 1, C = c)$ is a mixture of $Pr\{Y(1) = y \mid R = a, C = c\}$ and $Pr\{Y(1) = y \mid R = p, C = c\}$. Hence, neither $Pr\{Y(0) = y \mid R = n, C = c\}$ nor $Pr\{Y(1) = y \mid R = a, C = c\}$ are identified.

One option is to derive bounds for the PSDEs, that is, intervals which are known to contain the true values. Jemiai (2005), Zhang and Rubin (2003), and Hudgens, Hoering, and Self (2003) derived such bounds in the related problem of "truncation by death" (see Section 5). One limitation of these bounds is that they are often wide and not very informative. Furthermore, they are not practically useful in the presence of covariates, C. This is because rather than being known, the bounds have to be estimated for each level of C separately. When C is high dimensional, the variability in these estimates will be unacceptably high. Hence, a complementary sensitivity analysis that aids in discriminating between separate values of the PSDEs is often warranted. Below, we propose a novel approach that uses a pattern mixture model which facilitates such sensitivity analysis.

3. Sensitivity Analysis for the PSDE

3.1 *Pattern Mixture Formulation*

We use $\gamma_{r\cdot c}$ to quantify the following mean distances:

$$\begin{aligned}\gamma_{n\cdot c} &\equiv g[E\{Y(0) \mid R = p, C = c\}] - g[E\{Y(0) \mid R = n, C = c\}],\\ \gamma_{a\cdot c} &\equiv g[E\{Y(1) \mid R = a, C = c\}] - g[E\{Y(1) \mid R = p, C = c\}].\end{aligned} \tag{5}$$

For completeness, we define a "baseline parameter" $\alpha_{r\cdot c} \equiv g[E\{Y(0) \mid R = r, C = c\}]$, $r \in \{n, a\}$. We obtain

$$\begin{aligned}g[E\{Y(0) \mid R = n, C = c\}] &= \alpha_{n\cdot c},\\ g[E\{Y(1) \mid R = n, C = c\}] &= \alpha_{n\cdot c} + \beta_{n\cdot c},\\ g[E\{Y(0) \mid R = p, C = c\}] &= \alpha_{n\cdot c} + \gamma_{n\cdot c},\\ g[E\{Y(1) \mid R = p, C = c\}] &= \alpha_{a\cdot c} + \beta_{a\cdot c} - \gamma_{a\cdot c},\\ g[E\{Y(0) \mid R = a, C = c\}] &= \alpha_{a\cdot c},\\ g[E\{Y(1) \mid R = a, C = c\}] &= \alpha_{a\cdot c} + \beta_{a\cdot c}.\end{aligned} \tag{6}$$

Equations (3) and (6) together define the following pattern mixture model:[2]

$$\begin{aligned}&E(Y \mid Z = 0, X = 0, C = c)\\ &\quad = \frac{g^{-1}(\alpha_{n\cdot c})\pi_{n\cdot c} + g^{-1}(\alpha_{n\cdot c} + \gamma_{n\cdot c})\pi_{p\cdot c}}{\pi_{n\cdot c} + \pi_{p\cdot c}},\\ &E(Y \mid Z = 0, X = 1, C = c)\\ &\quad = g^{-1}(\alpha_{n\cdot c} + \beta_{n\cdot c}).\\ &E(Y \mid Z = 1, X = 0, C = c)\\ &\quad = g^{-1}(\alpha_{a\cdot c}),\\ &E(Y \mid Z = 1, X = 1, C = c)\\ &\quad = \frac{g^{-1}(\alpha_{a\cdot c} + \beta_{a\cdot c})\pi_{a\cdot c} + g^{-1}(\alpha_{a\cdot c} + \beta_{a\cdot c} - \gamma_{a\cdot c})\pi_{p\cdot c}}{\pi_{a\cdot c} + \pi_{p\cdot c}}.\end{aligned} \tag{7}$$

Because $(\pi_{n\cdot c}, \pi_{p\cdot c}, \pi_{a\cdot c})$ is identified under Assumptions 1–3, the right-hand side of equation (7) contains six "unknown" parameters whereas the left-hand side contains four "known" parameters. The system is thus overparameterized and cannot be solved without further assumptions.

From equation (7) we see that fixing $\gamma_{r\cdot c}$ renders $\beta_{r\cdot c}$ identified. In particular, we have that

$$\begin{aligned}\gamma_{n\cdot c} = 0 \Rightarrow \beta_{n\cdot c} &= g\{E(Y \mid Z = 0, X = 1, C = c)\}\\ &\quad - g\{E(Y \mid Z = 0, X = 0, C = c)\},\\ \gamma_{a\cdot c} = 0 \Rightarrow \beta_{a\cdot c} &= g\{E(Y \mid Z = 1, X = 1, C = c)\}\\ &\quad - g\{E(Y \mid Z = 1, X = 0, C = c)\}.\end{aligned} \tag{8}$$

That is, for $\gamma_{r\cdot c} = 0$, $\forall r$, stratification on Z yields the PSDE. The more $\gamma_{r\cdot c}$ deviates from 0, the less exchangeable are subjects across principal strata, and the more biased is such "stratified analysis." Our approach is to vary $(\gamma_{n\cdot c}, \gamma_{a\cdot c})$ over a grid of plausible values and estimate the PSDEs for each value separately.

The parameter $\gamma_{r\cdot c}$ may not be variation independent of $Pr(Y = y, Z = z, X = x, C = c)$. As a consequence, it may be possible to construct bounds for $\gamma_{r\cdot c}$. When $g(\cdot)$ is the logit-link (and Y is binary), then the $\gamma_{r\cdot c}$ and $Pr(Y = y, Z = z, X = x, C = c)$ are variation independent, and $\gamma_{r\cdot c}$ is unbounded (see the Appendix). In a practical setting, the bounds for $\gamma_{r\cdot c}$ have to be estimated. As for the bounds for $\beta_{r\cdot c}$, these estimates will be highly unstable when C is high dimensional.

3.2 *Modeling Assumptions*

If the covariates are low dimensional, the sample can be divided into strata defined by levels of the covariates, and the pattern mixture model can be applied to each stratum separately. Each stratum has its own PSDEs, and requires its own sensitivity analysis. If the covariates are high dimensional, the interpretational burden of this "nonparametric" approach will be overwhelming. In addition, most strata will contain very few, if any, observations, and estimates of the stratum-specific PSDEs will be highly unstable. Further modeling is required to deal with this "curse of dimensionality." We proceed by

[2] Note that this model is "nonparametric" in the sense that it does not impose any restrictions on $Pr[Y = y \mid Z = z, X = x, C = c]$.

assuming the following structural regression model for the conditional mean of $Y(x)$:

$$g[E\{Y(x) \mid R = r, C = c\}] = \alpha + \beta x + \zeta Z(x, r) + \gamma b(x, r) + \delta c, \tag{9}$$

in which $Z(x, r)$ is the potential outcome $Z(x)$ for a subject within principal stratum $R = r$, and $b(0, p) = -b(1, p) = 1$, $b(x, r) = 0$ for $r \in \{n, a\}$. We identify the following relationships with the previously defined parameters: $\alpha + \zeta Z(0, r) + \delta c = \alpha_{r \cdot c}$, $\beta = \beta_{n \cdot c} = \beta_{a \cdot c}$, $\gamma = \gamma_{n \cdot c} = \gamma_{a \cdot c}$. The model implies that the "effect parameters" $\beta_{r \cdot c}$ and $\gamma_{n \cdot c}$ are constant across principal strata and covariates, an assumption close in spirit to, for example, the standard "proportional hazards assumption" in survival analysis. The influence of C is captured by the baseline, which is assumed to vary linearly with C. The reason for including the term $\zeta Z(x, r)$ in the model is twofold: (i) It allows for the effect of X on Y for the "protected" ($R = p$) to differ from β. This is desirable, because for the "protected" the effect of X is both "direct" and mediated through Z. (ii) Without it, the model would imply that $E(Y \mid Z, X, C)$ is independent of Z under exchangeability ($\gamma = 0$). This is a strong restriction, which is not likely to hold in general. We model the principal strata proportions with polytomous logistic regression:

$$\log \frac{\pi_{n \cdot c}}{\pi_{p \cdot c}} = \tau_n + \omega_n c,$$
$$\log \frac{\pi_{a \cdot c}}{\pi_{p \cdot c}} = \tau_a + \omega_a c. \tag{10}$$

Under the models in equations (9) and (10), β may be identified without specifying γ (see, e.g., Teicher (1963), for a discussion on identification in pattern mixture models). We suspect, however, that the inference for β relies heavily on the particular choice of model. Hence, we nevertheless find it sensible to perform a sensitivity analysis by varying γ.[3]

3.3 *Estimation and Testing*

Define $\theta \equiv (\alpha, \beta, \delta)$, $\eta \equiv (\tau_n, \omega_n, \tau_a, \omega_a)$. For each fixed value of γ, the ML-estimate of (θ, η) can be obtained from the expectation–maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977). The observed (incomplete) data consist of n i.i.d. copies of (Y, Z, X, C). We consider the complete data to be n i.i.d. copies of (Y, Z, X, R, C). The complete data log likelihood is given by

$$\log L(\theta, \eta; y, z, x, r, c, \gamma) = \sum_i \sum_{r \in (n,p,a)} I(r_i = r) \log Pr\{Y(x_i) = y_i \mid R = r, C = c_i; \theta, \gamma\} + \sum_i \sum_{r \in (n,p,a)} I(r_i = r) \log \pi_{r \cdot c}(\eta). \tag{11}$$

The EM algorithm is an iterative procedure. Let (θ^t, η^t) denote the estimate of (θ, η) at iteration t. We initialize the algorithm with an arbitrary value (θ^0, η^0). The EM algorithm then iterates between the following two steps:

E-step: Calculate the expected complete data log likelihood given the observed data. This amounts to calculating

$$E\{I(r_i = r) \mid Y = y_i, Z = z_i, X = x_i, C = c_i; \theta^t, \eta^t, \gamma\} = \frac{I\{r_i \in R(z_i, x_i)\} Pr\{Y(x_i) = y_i \mid R = r_i, C = c_i; \theta^t, \gamma\} \pi_{r_i \cdot c}(\eta^t)}{\sum_{r \in (n,p,a)} I\{r \in R(z_i, x_i)\} Pr\{Y(x_i) = y_i \mid R = r, C = c_i; \theta^t, \gamma\} \pi_{r \cdot c}(\eta^t)}. \tag{12}$$

M-step: Maximize the complete likelihood in equation (11) with respect to (θ, η), and with $I(r_i = r)$ replaced by the (conditional) expectation counterpart. $(\theta^{t+1}, \eta^{t+1})$ is the point at which the maximum is attained. The first part of equation (11) can be maximized with standard generalized linear model (GLM; McCullagh and Nelder, 1989) software, whereas the second part requires a numerical routine for fitting polytomous logistic regression models. The user-specified parameter γ is an offset in the GLM.

The EM algorithm does not automatically provide standard errors. Those can be obtained from the Fisher information or from a bootstrap procedure.

4. Application to the National March Cohort

The NMC was established in 1997, when 300,000 Swedes participated in a national fund-raising event organized by the Swedish Cancer Society. Every participant was asked to fill in a questionnaire that included items on known or suspected risk factors for cancer and CVD. Questionnaire data were obtained on over 43,880 individuals. For further details on the NMC, see Lagerros (2006). Using the Swedish patient registry, these individuals have been followed over time, and each CVD event recorded. Based on self-reported history of PA, we classified each subject as either a "low-level exerciser" ($X = 1$) or a "high-level exerciser" ($X = 0$). If subjects had baseline (at 1997) BMI ≥ 30, they were classified as "obese" ($Z = 1$), otherwise as "not obese" ($Z = 0$). If subjects had at least one CVD event recorded during follow-up (1997–2004), they were classified as "with disease" ($Y = 1$), otherwise as "not with disease" ($Y = 0$). Because young people are both more physically active and less likely to develop CVD than old people, "age" is deemed to be a strong confounder in this setting. We use C to denote "age at baseline."

A "standard" analysis would regress Y on (X, Z, C), using, for example,

$$\text{logit} Pr(Y = 1 \mid X = x, Z = z, C = c) = \alpha^* + \beta^* x + \zeta^* z + \delta^* c, \tag{13}$$

which is equivalent to the structural model in equation (9) using $\gamma = 0$. For the NMC data, $\hat{\beta}^* = 0.26$, with 95% Wald

[3] This opinion does not seem to be universally accepted. Elliot, Joffe, and Chen (2006) and Mattei and Mealli (2006), for example, considered scenarios very similar to the one presented here. However, instead of carrying out a sensitivity analysis, they presented point estimates of the causal parameter of interest. Identification of this parameter was obtained by imposing parametric restrictions on the observed data distribution. In contrast, Rotnitzky, Robins, and Scharfstein (1998) and Vansteelandt and Goetghebeur (2005), for example, argued strongly for carrying out a sensitivity analysis whenever the parameter of interest is not nonparametrically identified.

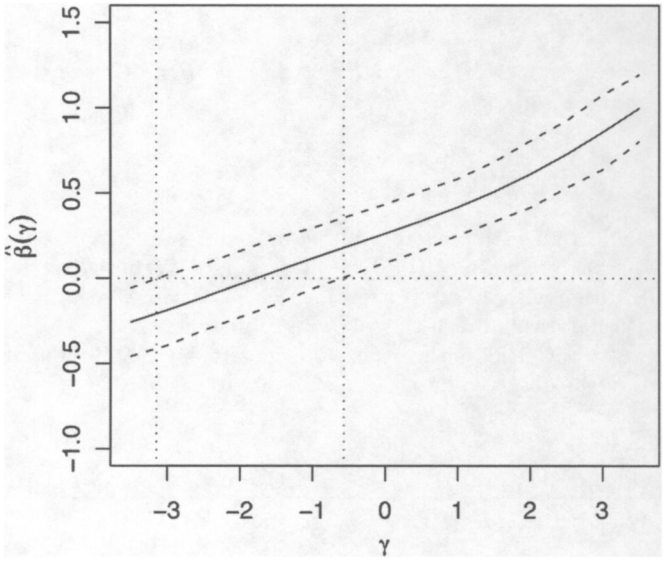

Figure 1. The ML-estimate of β in equation (9) with 95% pointwise confidence limits, as a function of γ.

CI (0.09, 0.43). This parameter value will only have the desired interpretation as a protective direct effect of PA against CVD if subjects are exchangeable across principal strata ($\gamma = 0$). To investigate the sensitivity in the NMC data to deviations from exchangeability, we fitted the structural model in equation (9), using $g(\cdot) = \text{logit}(\cdot)$. Figure 1 shows $\hat{\beta}$ as a function of γ. The dashed lines correspond to 95% pointwise Wald confidence limits, obtained from the inverse Fisher information. Figure 1 displays $\hat{\beta}$ for an extremely wide range of γ; $|\gamma| > 2.6$ corresponds to a "confounding effect" more than 10 times stronger than the direct PA effect under exchangeability. In reality we would probably expect any deviation from exchangeability to be less extreme. We observe that $\hat{\beta}$ is an approximately linear function of γ, with slope ≈ 0.15; a change in the "confounding effect" of 1 unit suggests a change in the PSDE of 0.15 units. Hence, $\hat{\beta}$ does not appear to be very sensitive to deviations from exchangeability. At the 5% significance level, the null hypothesis of no PSDE is rejected everywhere except in the region $-3.15 < \gamma < -0.57$. To further narrow the range of plausible values for β one would need to specify a range of plausible values for γ. Although we leave this as a task for subject matter experts, we note that for values of $\gamma > 0, \beta^*$ is biased downward as an estimate of the PSDE. Thus, as long we do not believe that predisposition for obesity is *negatively* correlated with predisposition for CVD (which corresponds to $\gamma < 0$), we may conclude that the direct protective effect of PA against CVD is at least as large as effects previously reported.

5. Truncation by Death

In recent years, several authors have considered the problem of "truncation by death" (Zhang and Rubin, 2003; Jemiai, 2005). In this problem, the outcome, Y, is only defined when the intermediate variable, Z, is equal to, say, 1. Y could, for example, represent "quality of life 1 year after treatment (X) assignment," and Z could represent "alive 1 year after assignment." One measure of the causal effect of X on Y is obtained by contrasting the distribution of $Y(1)$ with the distribution of $Y(0)$, for the subgroup whose members would survive under both treatment regimes. $\beta_{a\cdot c}$ is such a measure.[4] Jemiai (2005) and Zhang and Rubin (2003) derived bounds for $\beta_{a\cdot c}$. Gilbert, Bosch, and Hudgens (2003) proposed a biased selection model as a tool for conducting sensitivity analysis for $\beta_{a\cdot c}$ in the absence of covariates. Shepherd et al. (2006) extended this biased selection model to incorporate information on covariates. Jemiai et al. (2007) discussed semiparametric inference for the biased selection model. In this section, we compare our approach, based on a pattern mixture model, with the approaches of Gilbert et al./Shepherd et al./Jemiai et al.

Shepherd et al. (2006) proposed the following model

$$\text{logit}\, Pr(R = a \mid Y = y, Z = 1, X = 1, C = c) = m(c) + h(c, y), \tag{14}$$

for the mixing probabilities of the principal strata. Gilbert et al. (2003) used the same model, but without covariates. They referred to it as a *biased selection model.* Shepherd et al. (2006) showed that under Assumption 2, $\beta_{a\cdot c}$ is identified if the function $h(c, y)$ is given. Jemiai et al. (2007) showed that $h(c, y)$ is variation independent of $Pr(Y = y, Z = z, X = x, C = c)$. Gilbert et al. (2003), Shepherd et al. (2006), and Jemiai et al. (2007) performed a sensitivity analysis by modeling $h(c, y)$ and varying its model parameters over a range of values. For each value, they obtained an estimate of $\beta_{a\cdot c}$.

To show the connection to our work, we use Bayes' rule to reformulate the biased selection model as

$$Pr\{Y(1) = y \mid R = a, C = c\} = \frac{Pr\{Y(1) = y \mid R = p, C = c\} e^{h(c,y)}}{k(c)}, \tag{15}$$

where $k(c)$ is a normalizing constant. From equation (15) we see that specifying $h(c, y)$ is tantamount to specifying the "distance" (up to a normalizing constant) between the $Pr\{Y(1) = y \mid R = a, C = c\}$ and $Pr\{Y(1) = y \mid R = p, C = c\}$ for every point $Y = y$. On the contrary, by specifying $\gamma_{a\cdot c}$ in our pattern mixture model, we specify the "distance" between $E\{Y(1) \mid R = a, C = c\}$ and $E\{Y(1) \mid R = p, C = c\}$.

For binary Y, the biased selection model is equivalent to the pattern mixture model with a logit link (see the Appendix). Both models can easily be used for nonbinary Y as well (see Gilbert et al., 2003; Shepherd et al., 2006; and Section 6). For nonbinary Y, we recognize several arguments for using the pattern mixture model instead of the biased selection model. (i) To optimize the biased selection model we need to solve a set of complex integral equations (Shepherd et al., 2006). The pattern mixture model can be fitted with standard GLM software, within the EM algorithm. (ii) The dimension of $\gamma_{a\cdot c}$ is smaller than the dimension of $h(c, y)$. In this sense, specification of $\gamma_{a\cdot c}$ is less demanding than specification of $h(c, y)$. For practical use, however, both methods require additional parametric assumptions, and it is not obvious which of the methods will require the "strongest" assumptions in

[4] Note that $\beta_{n\cdot c}$ is not defined in this setting, because Y is not defined for any subject with $R = n$.

a realistic setting. Furthermore, if one is only interested in a certain parameter indexing $Pr\{Y(1) = y \mid R = a, C = c\}$ (such as $\beta_{a \cdot c}$) then the biased selection model allows for some misspecification of $h(c, y)$. In particular, Gilbert et al. (2003) demonstrated that in the absence of covariates, with the simple choice $h(c, y) = h(y) = \psi y$, every point within the bounds for $\beta_{a \cdot c} = \beta_a$ can be obtained by varying ψ. (iii) Given the simpler nature of $\gamma_{a \cdot c}$, this parameter may be easier to interpret than $h(c, y)$. This obstacle could be overcome in the biased selection model by using relation (18) in the Appendix to map values of $h(c, y)$ into values of $\gamma_{a \cdot c}$. Such a procedure is not entirely convenient though. Furthermore, because the true data generating process is not known, the average in equation (18) would have to be taken over the sample distribution, and the map would be prone to uncertainty due to sampling variability.

There are also disadvantages to the pattern mixture model. (i) It is suitable for inference on mean differences (and functions thereof), but not so convenient for other parameters. The biased selection model, on the other hand, can conveniently handle any functional of the distribution of potential outcomes. (ii) As stated earlier, $\gamma_{a \cdot c}$ is not necessarily variation independent of $Pr(Y = y, Z = z, X = x, C = c)$. This means that $\gamma_{a \cdot c}$ could potentially be taken "too far" in a sensitivity analysis. How to determine from the sample distribution when this happens is not a trivial task when data are high dimensional and sparse, and we have chosen not to elaborate on this issue. In a sense, the pattern mixture formulation thus makes it difficult for the investigator to extract all available information from the data. The biased selection model does not suffer from this problem, because $h(c, y)$ is variation independent of $Pr(Y = y, Z = z, X = x, C = c)$. We emphasize again that when Y is binary, this problem can be avoided by using the logit link, in which case $\gamma_{a \cdot c}$ is variation independent of $Pr(Y = y, Z = z, X = x, C = c)$. Furthermore, there is a natural "anchoring point" for $\gamma_{a \cdot c}$, namely "0" (exchangeability). This point is always compatible with any distribution $Pr(Y = y, Z = z, X = x, C = c)$, and moderate variations around this point are not likely to cause any conflict with $Pr(Y = y, Z = z, X = x, C = c)$.

6. Discussion

We have proposed a pattern mixture model as a tool for sensitivity analysis of the PSDE. We have shown that the parameters in this model can be estimated with the EM algorithm. When applied to NMC data, this method shows that there is likely to be a direct protective effect of PA against CVD, not mediated through a reduction in BMI. We have argued that our approach can be applied to handle "truncation by death," and we have compared our approach to earlier approaches to handle this problem.

An alternative would be to focus on the "controlled" direct effect (CDE; Pearl, 2001). The CDE is often easier to interpret than the PSDE (Robins, Rotnitzky, and Vansteelandt, 2007). On the other hand, to properly define the CDE one needs to assume that the intermediate variable, Z, can be controlled and fixed by an external intervention. This may be a reasonable assumption when Z represents a treatment, which could, at least in principle, be randomized. In our case, however, Z represents a "biomarker" (BMI), and it is not obvious how to conceptualize interventions on such variables. Hence, the PSDE may be a more suitable parameter for inference than the CDE in our setting.

Instead of focusing on the direct effect of X on Y, one may want to draw inference on the "indirect" effect, that is, the part of the effect of X on Y, which is mediated through Z. Defining the indirect effect may be complicated, if one is not willing to consider Z controllable (Pearl, 2001). In fact, we have seen no reasonable definition in the literature of indirect effects, in the context of principal stratification. Whether such a definition is possible remains an open question.

A natural extension of the pattern mixture model would be to allow for more refined measures of X, Z, and Y, i.e., for multilevel categorical variables, or continuous variables. To allow for general outcomes, Y, is straightforward. In this case, we need to model the full conditional distribution of $Y(x)$. As long as this distribution is assumed to be within the exponential family, the estimation method described in Section 3.3 is straightforward to apply. A consequence of refining X and/or Z is that the number of principal strata increases. Frangakis et al. (2004) demonstrated that when X is multilevel categorical, and Z remains binary, there is a natural extension of the monotonicity assumption (Assumption 2), which allows for nonparametric identification of the principal strata proportions. When both X and Z are multilevel categorical, it can be shown that nonparametric identification of the principal strata proportions is no longer feasible. When X and/or Z are continuous, the number of principal strata are "infinite." In this case it is not obvious whether the idea of "PSDEs" is still useful.

Another natural extension would be to introduce the notion of "time" into the modeling framework. In reality, both "PA" and "BMI" are processes in time rather than point exposures. In addition, there may be time-dependent covariates that are affected by exposure history, and which affect future exposure. Whereas a lot of work has been done for time-varying controlled direct effects (see, e.g., Robins, 1997), to the best of our knowledge no extensions to time-varying PSDEs have been considered in the literature.

Acknowledgements

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. We are grateful to Professor Rolf Sundberg (Stockholm University) for valuable discussions. RB was partially funded by Grant Agreement 2003134, EUPHORIC project, European Commission.

References

Dempster, A. P., Laird N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

Elliot, M. R., Joffe, M. M., and Chen, Z. (2006). A potential outcomes approach to developmental toxicity analyses. *Biometrics* **62,** 352–360.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58,** 21–29.

Frangakis, C. E., Brookmeyer, R. S., Varadhan, R., Safaeian, M., Vlahov, D., and Strathdee, S. A. (2004). Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a needle exchange program. *Journal of the American Statistical Association* **99,** 239–249.

Gilbert, P. B., Bosch, J. B., and Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* **59,** 531–541.

Hu, G., Tuomilehto, J., Silventoinen, K., Barengo, N., and Jousilahti, P. (2004). Joint effects of physical activity, body mass index, waist circumference and waist-to-hip ratio with the risk of cardiovascular disease among middle-aged Finnish men and women. *European Heart Journal* **25,** 2212–2219.

Hudgens, M. G., Hoering, A., and Self, S. G. (2003). On the analysis of viral load endpoints in HIV vaccine trials. *Statistics in Medicine* **22,** 2281–2298.

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

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

Lagerros, Y. T. (2006). *Physical activity from the epidemiological perspective—measurement issues and health effects.* Ph.D. thesis, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

Li, T. Y., Rana, J. S., Manson, J. E., Willet, W. C., Stampfer, M. J., Colditz, G. A., Rexrode, K. M., and Hu, F. B. (2006). Obesity as compared with physical activity in predicting risk of coronary heart disease in women. *Circulation* **113,** 499–506.

Mattei, A. and Mealli F. (2006). Application of the principal stratification approach to the Faenza randomized experiment on breast self-examination. *Biometrics* **63,** 437–446.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models,* 2nd edition. London: Chapman and Hall.

Mora, S., Lee, I., Buring, J. E., and Ridker, P. M. (2006). Association of physical activity and body mass index with novel and traditional cardiovascular biomarkers in women. *Journal of the American Medical Association* **295,** 1412–1419.

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence,* 411–420. San Francisco, CA: Morgan Kaufmann.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality,* 69–117. New York: Springer Verlag.

Robins, J. M., Rotnitzky, A., and Vansteelandt, S. (2007). Discussion of principal stratification designs to estimate input data missing due to death. *Biometrics* **63,** 650–654.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93,** 1321–1339.

Shepherd, B. E., Gilbert, P. B., Jemiai, Y., and 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.

Teicher, H. (1963). Identifiability of finite mixtures. *Annals of Mathematical Statistics* **34,** 1265–1269.

Vansteelandt, S. and Goetghebeur, E. (2005). Sense and sensitivity when correcting for observed exposures in randomized clinical trials. *Statistics in Medicine* **24,** 191–210.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death." *Journal of Educational and Behavioral Statistics* **28,** 353–368.

Received October 2007. *Revised April* 2008. *Accepted May* 2008.

Appendix

We show that (i) for nonbinary Y, $\gamma_{a\cdot c}$ is a function of $h(c, y)$ and $Pr(Y = y, Z = z, X = x, C = c)$, and (ii) for binary Y the biased selection model is equivalent to the pattern mixture model with $g(\cdot) = \text{logit}\,(\cdot)$. As a by-product of the second result we find that for binary Y and with $g(\cdot) = \text{logit}\,(\cdot)$, $\gamma_{r\cdot c}$ is variation independent of $Pr(Y = y, Z = z, X = x, C = c)$.

Shepherd et al. (2006) showed that the biased selection model in equation (14) implies that

$$Pr\{Y(1) = y \mid R = a, C = c\} = \frac{\text{expit}\{m(c) + h(c,y)\}Pr(Y = y \mid Z = 1, X = 1, C = c)}{E[\text{expit}\{m(C) + h(C,Y)\} \mid Z = 1, X = 1, C = c]},$$

$$Pr\{Y(1) = y \mid R = p, C = c\} = \frac{\{1 - \text{expit}\{m(c) + h(c,y)\}\}Pr(Y = y \mid Z = 1, X = 1, C = c)}{1 - E[\text{expit}\{m(C) + h(C,Y)\} \mid Z = 1, X = 1, C = c]}, \quad (16)$$

where $m(c)$ is the unique solution to

$$E[\text{expit}\{m(C) + h(C,Y)\} \mid Z = 1, X = 1, C = c] = \frac{Pr(Z = 1 \mid X = 0, C = c)}{Pr(Z = 1 \mid X = 1, C = c)}. \quad (17)$$

Combining equations (5) and (16) now yield

$$\gamma_{a\cdot c} = g\left(\frac{E[\text{expit}\{m(C) + h(C,Y)\}Y \mid Z = 1, X = 1, C = c]}{E[\text{expit}\{m(C) + h(C,Y)\} \mid Z = 1, X = 1, C = c]}\right) - g\left(\frac{E[\{1 - \text{expit}\{m(C) + h(C,Y)\}Y \mid Z = 1, X = 1, C = c]}{1 - E[\text{expit}\{m(C) + h(C,Y)\} \mid Z = 1, X = 1, C = c]}\right). \quad (18)$$

For binary Y and with $g(\cdot) = \text{logit}\,(\cdot)$, equation (18) simplifies to

$$\gamma_{a\cdot c} = h(c, 1) - h(c, 0). \quad (19)$$

Hence, for binary Y and with $g(\cdot) = \text{logit}\,(\cdot)$, specifying $\gamma_{a\cdot c}$ is tantamount to specifying $h(c, 1) - h(c, 0)$, and the models are thus equivalent in this case. Because $h(c, y)$ is variation independent of $Pr(Y = y, Z = z, X = x, C = c)$ (Jemiai et al., 2007), $\gamma_{a\cdot c}$ is also variation independent of $Pr(Y = y, Z = z, X = x, C = c)$ for binary Y and with $g(\cdot) = \text{logit}\,(\cdot)$. Variation independence of $\gamma_{n\cdot c}$ follows by symmetry.