JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Johns Hopkins University, Dept. of Biostatistics Working Papers

7-12-2006

INVESTIGATING MEDIATION WHEN COUNTERFACTUALS ARE NOT METAPHYSICAL: DOES SUNLIGHT UVB EXPOSURE MEDIATE THE EFFECT OF EYEGLASSES ON CATARACTS?

Brian Egleston
Biostatistics Facility, Fox Chase Cancer Center, brian.egleston@fccc.edu

Daniel O. Scharfstein
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Beatriz Munoz
Department of Ophthalmology, Johns Hopkins University

Sheila West
Department of Ophthalmology, Johns Hopkins University

Suggested Citation

Egleston, Brian; Scharfstein, Daniel O.; Munoz, Beatriz; and West, Sheila, "INVESTIGATING MEDIATION WHEN COUNTERFACTUALS ARE NOT METAPHYSICAL: DOES SUNLIGHT UVB EXPOSURE MEDIATE THE EFFECT OF EYEGLASSES ON CATARACTS?" (July 2006). *Johns Hopkins University, Dept. of Biostatistics Working Papers.* Working Paper 113.
http://biostats.bepress.com/jhubiostat/paper113

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.
Copyright © 2011 by the authors

Investigating mediation when counterfactuals are not metaphysical: Does sunlight UVB exposure mediate the effect of eyeglasses on cataracts?

Brian Egleston[1], Daniel Scharfstein[2], Beatriz Munoz[3], Sheila West[3]

[1] Biostatistics Facility
Fox Chase Cancer Center
333 Cottman Avenue
Philadelphia, PA 19111
e-mail: Brian.Egleston@fccc.edu
Phone: 215-214-3917

[2] Department of Biostatistics
[3] Department of Ophthalmology
Johns Hopkins University
Baltimore, MD 21205

July, 2006

Abstract

We investigate the degree to which a reduction in ocular sunlight ultra-violet B (UVB) exposure mediates a relationship between wearing eyeglasses and a decreased risk of cataracts. An estimand is proposed in which causal effects are estimated locally within strata based on potential UVB exposure without glasses and the degree to which glasses use reduces UVB exposure. We take advantage of the structure of the data in which the counterfactual UVB exposures if the participants in the study who wore glasses had not worn glasses are considered observable.

Keywords: mediation, causal inference, ophthalmology

This research was supported by the National Institute on Aging, Grant T32 AG00247.

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

1 Introduction

We review traditional methods and describe identification and estimation of a causal estimand for investigating mediation. We provide an example of a potential outcomes analysis in which a counterfactual outcome can be thought of as observable. This is an example of a potential outcomes analysis in which the counterfactuals are not inherently metaphysical.

The motivating example for this work is the degree to which a reduction in sunlight exposure mediates a relationship between eye glass use and a reduced risk of cataracts. Cataracts are a major source of vision loss in older persons, and billions are spent each year on corrective surgery (West *et al.*, 1998). Many authors have found that lifelong sunlight or ultra-violet B (UVB) exposure is a risk factor for the development of cortical cataracts (West *et al.* 1998, Cruickshanks, Klein, and Klein 1992, Delcourt *et al.* 2000). While there is evidence in the literature that glasses are effective in reducing cortical cataracts in older age, (Cruickshanks *et al.* 1992) to our knowledge no study has tried to quantify the degree to which the effect of glasses on preventing cataracts is related to their effectiveness at reducing sunlight UVB exposure.

1.1 Salisbury Eye Evaluation

Data came from the Salisbury Eye Evaluation (SEE) project, a population-based cohort study of 2,520 older adults in Salisbury, MD (West *et al.*, 1997). At baseline enrollment, which occurred during 1993-1995, participants were asked about their past use of glasses and sun exposure. In addition, photographs of the eye were taken and the percent of the cortical opacification was later assessed by 2 trained graders. The participants were coded as having developed clinically significant cortical cataracts in at least one eye if more than 3/16 sectors were affected, as described previously (West, *et al.* 1998).

Our aim was to estimate the degree to which corrective- and sun-glasses protected the wearer from cortical cataract development as a function of the amount of reduction in sunlight UVB afforded by glasses. In addressing this aim, we defined the use of glasses as reporting any daytime outside use of corrective- or sun-glasses at the age of 31. The amount of sunlight UVB exposure to the eyes, measured in Maryland Sun Years (MSY) (West *et al.* 1998, Duncan, Munoz, Bandeen-Roche, West 1997), was defined for the four-year age period 31-34. We used 34 as the cut-off for exposure to reduce problems

http://biostats.bepress.com/jhubiostat/paper113

with eyeglass cross-over; many of those who did not wear glasses at age 31 began to wear glasses later in life. In the data, recalled UVB exposure in the early thirties had a 0.79 correlation with total recalled UVB exposure suggesting that early exposure is a reasonable proxy measure for recalled total exposure.

2 Data Structure and Notation

We define the data structure for a random individual. Let Z denote the observable indicator of use of eyeglasses (1 if use, 0 otherwise). Let $M(1)$ and $M(0)$ denote the UVB exposure under use and non-use of glasses, respectively. The observable UVB exposure is $M = M(Z)$. Let $Y(1, m)$ and $Y(0, m)$ denote the indicators of clinically significant cataracts under use and non-use of glasses and UVB-exposure level m, respectively (1 if cataracts, 0 otherwise). We define $Y(z) = Y(z, M(z))$ to be the indicator of clinically significant cataracts under glasses use level z. The observable outcome is $Y = Y(Z) = Y(Z, M(Z))$. Let X denote a vector of covariates.

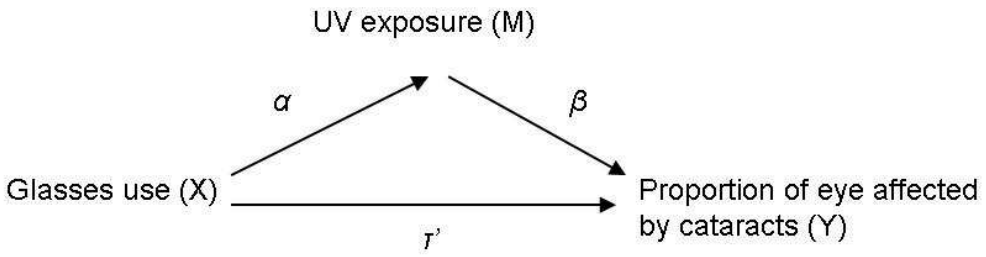
As one cannot observe an individual in multiple states of nature at the same point in time, the canonical observed data structure is $O = (Z, M, Y, X)$. An interesting feature of our study is that eye researchers have previously defined a mathematical formula, based on survey responses and optical physics regarding ambient UVB, that maps, for an individual subject who reported wearing corrective- or sun-glasses at ages 31, the UVB exposure she would have received if she had not worn glasses. With this substantive knowledge, we can think of $M(0)$ as observable. We could not reconstruct $M(1)$ for the whole sample because we did not know how much people in the study observed not to wear glasses would have worn glasses if they were glasses users. In our investigation, the observed data structure for an individual will be $O^\dagger = (Z, M(0), M, Y, X)$. We assume that we observe n i.i.d. copies of O or $O^\dagger$.

3 Mediation

Our causal hypothesis is that the effect of glasses on cortical cataracts is mediated by the effect of glasses on reducing sunlight UVB exposure. To our knowledge, there is no scientifically plausible mechanism by which eyeglasses would be causally linked to cortical cataracts beyond their impact on UVB exposure; any other association would likely be

Hosted by The Berkeley Electronic Press

Figure 1: Mediational Model Discussed by Baron and Kenny (1986)



through non-causal confounding, such as by sociodemographic status. In this paper, we will not examine other types of cataracts, such as nuclear, with which glasses use might be associated through their relationship with other risk factors, such as myopia (Chang, Congdon, Bykhovskaya, Munoz, West 2005). To place our work in a broader context, we use this section to review and comment on traditional methods for investigating mediation. The best known method for investigating mediation in the social sciences is the path analytic model described by Baron and Kenny (1986) and represented in the directed acyclic graph in Figure 1. This model is expressed in terms of the canonical observed data random vector O. Their model can be summarized as follows:

$$\begin{aligned} E[Y|Z=z] &= \gamma_1 + \tau z \\ E[M|z=z] &= \gamma_2 + \alpha z \\ E[Y|Z=z, M=m] &= \gamma_3 + \tau' z + \beta m \end{aligned}$$

The common way in which this model is used to investigate mediation involves the evaluation of three conditions: (1) the total effect (τ) is clinically relevant and statistically significant; (2) the associations between exposure and the mediator (α) and between mediator and outcome after controlling for exposure (β) are statistically significant; and (3) the direct effect (τ') is smaller in absolute value, from both a clinical and statistical perspective, than the total effect τ. The influence of Baron and Kenny's paper cannot be understated. A social-science citation index reveals that their paper has been cited in over 6,200 papers.

Using the potential outcomes framework and proper assumptions, we can assign causal meaning to Baron and Kenny's direct and indirect measures. Before turning to these assumptions, we discuss two types of causal direct and indirect measures.

http://biostats.bepress.com/jhubiostat/paper113

3.1 Prescriptive and Descriptive Effects

Direct and indirect effects in the causal inference literature are generally defined as either prescriptive or descriptive (Pearl 2001). Prescriptive effects are defined as a contrast between a functional of the distribution of $Y(z, m)$ and a functional of the distribution of $Y(z^*, m^*)$. When $z = 1$, $z^* = 0$, and $m = m^*$, the prescriptive effect contrasts the distribution of cataract outcomes in a world in which all subjects wear eyeglasses and have UVB exposure m and a world in which all subjects do not wear eyeglasses and have UVB exposure m.

Descriptive effects are defined as a contrast between a functional of the distribution of $Y(z, M(z))$ and a functional of the distribution of $Y(z^*, M(z^{**}))$, where either $z^* = z, z^{**} \neq z$ or $z^* \neq z, z^{**} = z$. When $z = 1$, $z^* = 0$, and $z^{**} = 1$, the descriptive effect contrasts the distribution of cataract outcomes in a world in which all subjects wear eyeglasses and a world in which all subjects do not wear eyeglasses and have UVB exposure equal to what they would have had, had they worn glasses.

In the next subsection, we describe sufficient conditions for identification of prescriptive and descriptive effects and make the connection to the parameters of the Baron and Kenny model.

3.2 Identification of Prescriptive and Descriptive Effects

To start, we assume that we can randomize Z; we will later generalize the results to the observational study setting. For the moment, we consider the cataract outcome to be continuous, for example, the maximum of the proportions of each eye that are covered with cataracts. We examine identification in the context of the canonical data, O.

Under randomization of Z, we have that

$$Z \perp \{M(z), Y(z, m) : \text{ for all } m, z\} \tag{3.1}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Under the Baron and Kenny model and (3.1),

$$\begin{aligned} E[M(z)] &= E[M(z)|Z=z] \\ &= \gamma_2 + \alpha z \\ E[Y(z)] &= \int E[Y(z,m)|Z=z, M(z)=m] dF_{M(z)}(m) \\ &= (\gamma_3 + \beta\gamma_2) + (\tau' + \beta\alpha)z \\ E[Y(z)] &= E[Y(z, M(z))|Z=z] \\ &= \gamma_1 + \tau z \end{aligned}$$

This implies that $\gamma_1 = \gamma_3 + \beta\gamma_2$ and $\tau = \tau' + \beta\alpha$. Furthermore, note that Baron and Kenny's total effect can be written as:

$$\tau = E[Y(1) - Y(0)]$$

Hence, the total effect is the expected difference between the cataract outcome under glasses and the cataract outcome under no glasses.

To attach causal meaning to τ', we need additional assumptions. In the next two subsections, we introduce two sets of assumptions. First we discuss the assumptions of Robins (2003) and next we discuss the assumptions of Pearl (2001).

3.2.1 Assumptions of Robins (2003)

Robins (2003) assumed that

$$Y(z,m) \perp M(z) \mid Z = z \text{ for all } z \text{ and } m \tag{3.2}$$

In words, Assumption (3.2) states that, among subjects with observed glasses use level z, the potential cataract outcomes under glasses level z and UVB exposure level m is independent of UVB exposure under glasses use z. Robins (2003) refers to this assumption as the fully randomized structured tree graph (FRCISTG) model.

In this section, we show how this assumption can be used to give causal meaning to Baron and Kenny's estimand; we later discuss the limitations of this assumption. Under the Baron and Kenny model and Assumptions (3.1,3.2),

$$\begin{aligned} E[Y(z,m)] &= E[Y(z,m)|Z=z, M(z)=m] \\ &= E[Y|Z=z, M=m] \\ &= \gamma_3 + \tau' z + \beta m \end{aligned}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper113

Thus,

$$\tau' = E[Y(1,m) - Y(0,m)]$$

Baron and Kenny's direct effect is then the expected difference between the cataract outcome if an individual is forced to wear glasses and have UVB exposure m and the cataract outcome if she retains the same UVB exposure level but is forced not to wear glasses. Here, the use of glasses varies, while the UVB exposure level is kept fixed at m. Thus, τ' measures the causal effect of glasses on cataracts holding UVB exposure fixed at m. Pearl (2001) refers to this effect as the controlled directed effect (CDE). Baron and Kenny's model assumes that the CDE is constant with respect to m.

Baron and Kenny's measure of the indirect effect, $\alpha \times \beta$, is equal to $\tau - \tau'$. From above, we see that

$$\begin{aligned} \tau - \tau' &= \{E[Y(1)] - E[Y(0)]\} - \{E[Y(1,m)] - E[Y(0,m)]\} \\ &= \alpha \times \beta \end{aligned}$$

Thus, $\alpha \times \beta$ represents the expected difference *between* the treatment effect for an individual when she wears and does not wear glasses *and* the treatment effect for an individual when she wears and does not wear glasses but is forced to have UVB exposure m. We refer to this expected difference as the controlled indirect effect (CIE). Baron and Kenny's model assumes that the CIE does not depend on m. The CDE and CIE are prescriptive effects.

Robins' (2003) introduced the no-interaction assumption that allows τ' and $\alpha \times \beta$ to be interpreted as descriptive direct and indirect effect measures. His no-interaction assumption states that

$$Y(1,m) - Y(0,m) = B \tag{3.3}$$

where B is a random variable that does not depend on m. This assumption says that, when each individual is forced to have UVB exposure m, the difference between their cataract outcome under glasses and their cataract outcome under no glasses does not vary with m. Under (3.1,3.2) and Baron and Kenny's model, we know that $E[B] = \tau'$.

Under (3.1,3.2,3.3), we see that Robins' (2003) pure direct effect (PDE) and

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Pearl's (2001) natural direct effect (NDE) is

$$\begin{aligned} PDE(NDE) &= E[Y(1, M(0))] - E[Y(0, M(0))] \\ &= \int E[Y(1, m) - Y(0, m)|M(0) = m]dF_{M(0)}(m) \\ &= \int E[B|M(0) = m]dF_{M(0)}(m) \\ &= E[B] = \tau' \end{aligned}$$

Similarly, Robins' (2003) total direct effect (TDE) and Pearl's (2001) natural direct effect (NDE) is τ'. We can also show that Robins' (2003) pure indirect effect (PIE), total indirect effect (TIE), and Pearl's (2001) natural indirect effect (NIE) is $\alpha \times \beta$.

3.2.2 Pearl's Identifying Assumption

Pearl (2001) presents alternative identifying assumptions that do not require the no-interaction assumption and can be used for binary outcomes. Specifically, he assumes that Z and M are randomized and that

$$Y(z, m) \perp M(1 - z) \text{ for all } z, m$$

(Pearl (2001) actually assumes that this independence holds within levels of confounding covariates; for didactic purposes, we have excluded such covariates.) Since we view the data temporally, i.e., Z precedes M and M precedes Y, we find it natural to adapt Pearl's assumptions as follows. We assume (3.1,3.2) hold and that

$$Y(z, m) \perp M(1 - z)|Z = z \text{ for all } z, m \tag{3.4}$$

Under (3.1,3.2,3.4) and Baron and Kenny's model, the pure and natural direct effects under Pearl's identification assumption equal the same values as shown in section 3.2.1.

The key difference between Robins' assumptions and our adaptation of Pearl's assumptions is that under the latter assumptions $E[Y(z, M(1 - z))]$ is identifiable since

$$\begin{aligned} E[Y(z, M(1 - z))] &= \int E[Y(z, m)|M(1 - z) = m, Z = z]dF_{M(1-z)}(m) \\ &= \int E[Y(z, m)|M(z) = m, Z = z]dF_{M(1-z)}(m) \\ &= \gamma_3 + \tau' z + \beta E[M(1 - z)] \end{aligned}$$

Thus, one can estimate non-linear measures of direct and indirect effects, e.g., $E[Y(0, M(1))]/E[Y(1, M(1))]$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper113

3.2.3 Binary Outcomes

For binary outcomes, such as the presence or absence of cataracts, practitioners of statistics have commonly sought to assess mediation by extrapolating Baron and Kenny's framework to the generalized linear model domain (see, for example, Bridge, Day, Richardson, Birmaher, and Brent 2003, O'Leary *et al.* 2005, Ayalon, Arean, and Alvidrez 2005). In general, they fit models of the form:

$$\begin{aligned} g(E[Y|Z=z]) &= \gamma_1 + \tau z \\ h(E[M|z=z]) &= \gamma_2 + \alpha z \\ g(E[Y|Z=z, M=m]) &= \gamma_3 + \tau' z + \beta m \end{aligned}$$

where $g(\cdot)$ and $h(\cdot)$ are specified link functions. It is important to note that, due to the non-linearity of the link functions, there is, in general, a complicated relationship between the model parameters. In using this model, practitioners use the same procedure to infer mediation as in Baron and Kenny's linear model.

Can we use Assumption (3.4) to interpret τ' as the PDE or NDE? In other words, is τ' a contrast between a functional of the distribution of $Y(z, M(z))$ and a functional of the distribution of $Y(1-z, M(z))$. Under Assumptions (3.1,3.2,3.4),

$$\begin{aligned} P[Y(z, M(z))=1] &= \text{logit}^{-1}(\gamma_1 + \tau z) \\ P[Y(1-z, M(z))=1] &= \int \text{logit}^{-1}(\gamma_3 + \tau'(1-z) + \beta m) dF_{M(z)}(m) \end{aligned}$$

Due to the complicated nature of the expression for $P[Y(1-z, M(z))=1]$, it is not possible to, in general, interpret τ' as PDE or NDE. By similar arguments, it is not possible to interpret a function of τ and τ' as a PIE or NIE. Furthermore, CDE (CIE) is not equal to PDE (PIE) or NDE (NIE).

3.2.4 Observational Studies

So far, we have assumed that Z is randomized. The results above can be extended to the setting where we further condition (3.1), (3.2), and (3.4), on a subset of observed baseline covariates, X. The assumptions are, respectively,

$$Z \perp \{M(z), Y(z,m) : \text{ for all } m, z\} \mid X = x \text{ for all } x \tag{3.5}$$

$$Y(z,m) \perp M(z) \mid Z = z, X = x \text{ for all } z, x \text{ and } m \tag{3.6}$$

$$Y(z,m) \perp M(1-z) | Z = z, X = x \text{ for all } z, x \text{ and } m \tag{3.7}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

3.2.5 Identification with $O^\dagger$

We have discussed identification specifically with the canonical data structure, O, not $O^\dagger$ in which $M(0)$ is considered observable. Knowing $M(0)$ on those in whom $Z = 1$ does allow us to weaken the above assumptions slightly. To identify $E[Y(1, M(0))]$, we would just need assumption (3.1) and the following assumption under randomization.

$$Y(1, m) \perp M(1) \mid M(0) = m, Z = 1 \text{ for all } m \tag{3.8}$$

Then, the following holds where the terms in the final equality are identifiable from the observed data.

$$\begin{aligned} E[Y(1, M(0))] &= \int E[Y(1, m)|M(1) = m, M(0) = m, Z = 1] dF_{M(0)}(m) \\ &= \int E[Y|M = m, M(0) = m, Z = 1] dF_{M(0)}(m) \end{aligned}$$

Also, we could incorporate $M(0)$ into X in assumption (3.6) for identification of $E[Y(1, m)]$.

3.3 Limitations of the Direct and Indirect Paradigm

As shown above, Baron and Kenny's path analytic conceptualization of mediation as the linear decomposition of effects into direct and indirect effects has heavily dominated the literature. For example, Holland (1988), Pearl (2000, 2001), Robins and Greenland (1992), Robins (2003), and van der Laan and Peterson (2004) have written extensively on assumptions necessary to assign causal interpretations to varying estimates of direct and indirect effects in linear models.

One problem with Robins' and Pearl's causal presentation of direct and indirect effects is that the assumptions necessary for identification are quite strong. For example, imagine a study assessing whether high triglyceride levels partially mediate a relationship between Highly Active Antiretroviral Therapy (HAART) for HIV and mortality that is independent of HIV's effect on mortality. We could imagine that the development of abnormally high triglyceride levels might be indicative of general system dysregulation or frailty (Fried, Ferrucci, Darer, Williamson, and Anderson *et al.* 2004) which is hard to measure. It might be difficult to find a set of covariates, X, comprehensive enough for assumptions (3.5),(3.6), and (3.7) to hold.

Another problem with Robins' and Pearl's descriptive effects is that it is difficult to imagine an experiment which could directly identify them. By definition, mediators are in the causal pathway between an exposure and an outcome. Temporally, an

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper113

investigator assigns treatment and then the treatment affects the mediator. Hence, it would not be possible to construct any experiment in which we could directly observe $Y(1, M(0))$.

A limitation of Baron and Kenny's path analytic approach in general is that the path models are only valid for continuous mediators and outcomes. Robins' no-interaction assumption might identify path analytic models, but it is not directly generalizable to a probability scale.

The limitations of Baron and Kenny's estimands, along with the limitations of the direct and indirect estimands as they have been conceptualized in the literature in general, lead us to break with the strict direct and indirect paradigm. Instead, we develop a strata specific estimand based on $M(0)$ and $M(1)$ that allows for investigations of mediation. In this regard, we are following the advice given by Rubin (2004) who advocates that researchers look for more creative tools of investigating mediation, rather than strictly defining mediated effects as indirect effects.

4 Causal Estimand

For this study, our estimand of interest is the relative risk,

$$\nu(p, m) = \frac{P[Y(1) = 1 | P = p, M(0) = m]}{P[Y(0) = 1 | P = p, M(0) = m]} \tag{4.1}$$

where $P = M(1)/M(0)$ represents the proportion of baseline (no glasses) exposure that a person receives if he wears glasses. This is the relative risk of developing cataracts for an individual under glasses versus no glasses conditional upon being in the principal stratum (Frangakis and Rubin 2002) in which $P = p$ and $M(0) = m$.

We chose such an estimand because it is interpretable and relevant to the question at hand. If the effect of glasses on reducing the risk of cataracts is mediated by their effectiveness at reducing UVB exposure, we would expect that $\nu(1, m) = \nu(1, m') = 1$ for all m and m' and $\nu(p, m) > \nu(p', m)$ if $p > p'$. If there is no mediational effect, then we would expect that $\nu(p, m) = \nu(p', m)$ for all p, p', and m.

The monotonicity of $\nu(p, m) > \nu(p', m)$ if $p > p'$ might be violated if the individuals in the principal strata defined by $\{P = p, M(0) = m\}$ and $\{P = p', M(0) = m\}$ differ substantially by confounding covariates such as diabetes. However, we believe that within neighborhoods of $P = p$ where $M(0) = m$, the principal strata should be

COBRA Collection of Biostatistics Research Archive

Hosted by The Berkeley Electronic Press

comparable enough that we would generally expect such monotonicity. Outdoor workers had high baseline UVB exposure, for example, while housewives did not. We are hence making local inferences about causality within similar subpopulations, not global inferences across the whole population.

5 Identification and Models

Here we discuss the assumptions necessary to identify our estimand and the regression models used for estimation. In order to identify the causal estimand, we need to estimate $P[Y(0) = 1|P = p, M(0) = m]$ and $P[Y(1) = 1|P = p, M(0) = m]$. The structure of the data necessitates that we identify and estimate them in different ways.

5.1 A Non-metaphysical Counterfactual

As mentioned in Section 2, there is a unique aspect of our data structure, $O^\dagger$. Using an empirical model that has been previously been developed (West *et al.* 1998), we were able to determine, for those who wore eyeglasses, the UVB exposure they would have had, had they not worn eyeglasses. In particular, the empirical model states,

$$M = \sum_{s=1}^{12} G(s)R(s) \sum_{t=5}^{18} F(t,s)H(t,s)T_{hats}(t,s)T_{eye}(t,s) \tag{5.1}$$

where

M	=	Total UVB exposure
s	=	Month
t	=	Hour of day
$G(s)$	=	Geographic correction factor
$R(s)$	=	Ocular ambient exposure ratio
$F(t,s)$	=	Fraction of time spent outdoors
$H(t,s)$	=	Global ambient exposure
$T_{hats}(t,s)$	=	Percent of possible UVB exposure penetrating hats
$T_{eye}(t,s)$	=	Percent of possible UVB exposure penetrating glasses

As shown in the equation, the model included terms for season, "the fraction of time spent outdoors for each period of the day, the global ambient exposure during this day, fixed factors that reflect diminutions conferred by the use of hats and eyewear, [and]

http://biostats.bepress.com/jhubiostat/paper113

a geographic correction factor."(West *et al.*, 1998) The effect of glasses on exposure reduction ($T_{eye}(t,s)$) was calculated using two pieces of information. First, participant responses were used to determine how frequently and during which activities participants wore glasses. Next, the fraction of sunlight UVB by which a pair of glasses reduced exposure was adjusted for the predominant type of eyeglasses, such as plastic or glass, in use during that historical time period.

By setting $T_{eye}(t,s) = 1$ for all s and t, we are able to determine the counterfactual $M(0)$ for subjects who wore glasses. Here we are implicitly assuming that an individual would still have the same sunlight-exposure related activities if they did not wear glasses. This assumption is reasonable since partaking in exposure related-activities drives eyeglass use, and not vice versa.

The ability to determine UVB exposure under both states of glasses use for participants who wore glasses allows us to make less stringent assumptions for identification than those of Dominici, Zeger, Parmigiani, Katz, and Christian (2004) who proposed a conceptually similar estimand. Dominici *et al.* (2004) had to make untestable assumptions about the joint distribution of the potential outcomes; in particular the joint distribution of $M(0)$ and $M(1)$. In our work, we can identify the joint distribution of $M(1)$ and $M(0)$ in the group who wore glasses since we observe both variables in this group.

5.2 Assumptions

Here we define the assumptions necessary to identify our causal estimand.

Assumption 1: Stable Unit Treatment Value Assumption (SUTVA)

Our first assumption is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1980) which states that an individual's potential outcomes are unrelated to glasses use of other study participants and the mechanism by which participants come to wear glasses. In a randomized trial setting, the second component of SUTVA implies that there are no more than two well-defined treatment arms of a study. We expand this assumption by assuming that participants are consistent in their pattern of glasses use in their early thirties. Such an assumption of "perfect compliance" is reasonable given that we defined eyeglass use in such a way as to minimize eyeglass use cross-over.

Assumption 2: $Z \perp \{Y(0), Y(1), M(1)\} \mid M(0), X$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

This states that eyeglass use is independent of the potential outcomes given the confounding covariates, X, and baseline exposure level $M(0)$. This is an observational study equivalent of the randomization assumption in randomized trials. In this study we define X to be job status in the participants' thirties (outdoor work over water, outdoor work over land, inside work, and homemaker), age modeled as a natural cubic spline with 3 knots, sex, race (black vs. white), high school graduate, and diabetic status. The few individuals in our study who were students, or disabled in their thirties were grouped with the inside workers.

Assumption 3: $Y(0) \perp M(1) \mid Z, M(0), X$

Assumption 3 assumes that within levels of covariates, X, and baseline UVB exposure, $M(0)$, glasses-wearers' observed UVB exposures are not associated with their counterfactual cataract outcomes had the wearers not worn glasses. The assumption similarly holds for those who do not wear glasses. This is similar to assumption (3.7) but weaker because we condition on baseline UVB exposure and do not require the independence to be for $Y(0, m)$ for all m where both glasses use and UVB exposure are considered fixed for an individual.

This is similar to an assumptions made explicitly by Dominici *et al.* (2004) and implicitly by Taylor, Wang, and Thiebaut (2005) in their work. However, Dominici *et al.* had to make additional assumptions, as stated above, and Taylor *et al.* (2005) discussed identification in a hypothetical example where the number of principal strata are finite.

5.3 Identification

Here we discuss identification of $P[Y(1) = 1 | P = p, M(0) = m]$. For ease of presentation we demonstrate how we can identify $E[Y(1) | P_{dp} = 1, M(0) = m]$ where P_{dp} is a random variable indicating whether an individual's value of P falls within a neighborhood dp of p; that is $P_{dp} = 1$ when $P = p$ falls within dp, 0 otherwise. The following equality holds by assumption 2.

$$
\begin{aligned}
& E[Y(1) | P_{dp} = 1, M(0)] \\
&= E\left[\frac{P[Y = 1 | Z = 1, P_{dp} = 1, M(0), X] P[P_{dp} = 1 \mid Z = 1, M(0), X]}{E[P[P_{dp} = 1 \mid Z = 1, M(0), X]]} \,\middle|\, M(0) \right]
\end{aligned}
$$

Each of the terms on the right hand side of the equality is identifiable from the observed data because we observe $M(0)$ on all individuals and we observe P on individuals in whom $Z = 1$.

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper113

Similarly, we can use the following equality to demonstrate the identifiability of $P[Y(0) = 1|P = p, M(0) = m]$. As above, for ease of presentation we demonstrate how we can identify $E[Y(0)|P_{dp} = 1, M(0) = m]$ using Assumptions 2 and 3.

$$\begin{aligned} & E[Y(0)|P_{dp} = 1, M(0)] \\ &= E\left[\frac{P[Y = 1|Z = 0, M(0), X]P[P_{dp} = 1\ |Z = 1, M(0), X]}{E[P[P_{dp} = 1\ |Z = 1, M(0), X]]}\ \middle|\ M(0)\right] \end{aligned}$$

5.4 Models

Here we define a set of models we need for estimation of the causal estimand. Define

$$\boldsymbol{\psi}^* = (\boldsymbol{\beta_0}^{*'}, \boldsymbol{\beta_1}^{*'}, \boldsymbol{\gamma}^{*'}, \boldsymbol{\eta}^{*'}, \phi^*)$$

to represent the true model parameters. We assume that $\boldsymbol{\psi}^* \in \boldsymbol{\psi}$, where

$$\boldsymbol{\psi} \ = \ \{\boldsymbol{\psi} = (\boldsymbol{\beta_0}', \boldsymbol{\beta_1}', \boldsymbol{\gamma}', \boldsymbol{\eta}', \phi)' : \boldsymbol{\beta_z} \in R^{j_z}, \boldsymbol{\gamma} \in R^k, \boldsymbol{\eta}, \in R^l, \phi \in R, z = 0, 1\}$$

where j_z, k, and l are equal to the number of parameters indexing the respective models.

Let $g_z(P, M(0); \boldsymbol{\beta}_z)$, $h(M(0), X; \boldsymbol{\gamma})$, $k(M(0), X; \boldsymbol{\eta})$ represent smooth functions (in this study P and $M(0)$ are modeled using natural cubic splines with four knots) indexed by $\boldsymbol{\beta}_z$, $\boldsymbol{\gamma}$, and $\boldsymbol{\eta}$ respectively. We are assuming the following regression models.

$$\text{logit } P[Y(0) = 1|P, M(0)] = g_0(P, M(0); \boldsymbol{\beta}_0^*) \tag{5.2}$$

$$\text{logit } P[Y(1) = 1|P, M(0)] = g_1(P, M(0); \boldsymbol{\beta}_1^*) \tag{5.3}$$

$$\text{logit } P[Z = 1|M(0), X] = h(M(0), X; \boldsymbol{\gamma}^*) \tag{5.4}$$

$$\text{logit } E[P|M(0), X] = k(M(0), X; \boldsymbol{\eta}^*) \tag{5.5}$$

Since $1 > P > 0$, we model P as a Beta random variable since the Beta distribution is a flexible distribution for continuous variables bounded between 0 and 1. We estimate $f(p \mid M(0), X, Z = 1; \boldsymbol{\eta}^*, \phi^*)$ using a Beta regression as proposed by Ferrari and Cribari-Neta (2004). A Beta regression is a generalized linear model using a logit link and a Beta family. Under our parameterization, we assume that P has a conditional mean and variance equal to,

$$\begin{aligned} E[P|M(0), X] &= \mu(M(0), X; \boldsymbol{\eta}^*) \\ Var[P|M(0), X] &= \frac{\mu(M(0), X; \boldsymbol{\eta}^*)(1 - \mu(M(0), X; \boldsymbol{\eta}^*))}{1 + \phi^*} \end{aligned}$$

where $\mu(M(0), X; \boldsymbol{\eta}^*) = \text{expit}\{k(M(0), X; \boldsymbol{\eta}^*)\}$, $\text{expit}\{U\} = \exp\{U\}/(1 + \exp\{U\})$ and ϕ^* is a scalar.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

5.5 Estimation

We can estimate $\boldsymbol{\gamma}^*$, $\boldsymbol{\eta}^*$, and ϕ^* by maximum likelihood via logistic and Beta regressions, respectively. The unbiased estimating equations for these parameters are as follows.

$$\begin{aligned}
\boldsymbol{U}_{\boldsymbol{\gamma}^*}(O^\dagger;\boldsymbol{\psi}^*) &= h'(M(0),X;\boldsymbol{\gamma}^*)(Z-\text{expit }\{h(M(0),X;\boldsymbol{\gamma}^*\})) \\
\boldsymbol{U}_{\boldsymbol{\eta}^*}(O^\dagger;\boldsymbol{\psi}^*) &= \mu'(M(0),X,\boldsymbol{\eta}^*)\frac{\partial}{\partial\mu(X;\boldsymbol{\eta}^*)}l(O^\dagger,\phi^*,\mu(M(0),X,\boldsymbol{\eta}^*)) \\
\boldsymbol{U}_{\phi^*}(O^\dagger;\boldsymbol{\psi}^*) &= \frac{\partial}{\partial\phi^*}l(O^\dagger,\phi^*,\mu(M(0),X,\boldsymbol{\eta}^*))
\end{aligned}$$

where,

$$\begin{aligned}
&l(O^\dagger,\phi^*,\mu(M(0),X,\boldsymbol{\eta}^*)) \\
&= Z(\log\Gamma(\phi^*)-\log\Gamma(\mu(X;\boldsymbol{\eta}^*)\phi^*)-\log\Gamma((1-\mu(X;\boldsymbol{\eta}^*))\phi^*)+ \\
&\quad (\mu(X;\boldsymbol{\eta}^*)\phi^*-1)\log P+((1-\mu(X;\boldsymbol{\eta}^*))\phi^*-1)\log(1-P)),
\end{aligned}$$

$\Gamma(\cdot)$ is the gamma function, and $h'(M(0),X;\boldsymbol{\gamma}^*)$ and $\mu'(M(0),X,\boldsymbol{\eta}^*)$ represent the partial derivatives of the functions with respect to the parameters but evaluated at the truth.

An unbiased estimating equation for $\boldsymbol{\beta}_1$ is

$$\boldsymbol{U}_{\boldsymbol{\beta}_1}(O^\dagger;\boldsymbol{\psi}^*) = \frac{Zg_1'(P,M(0);\boldsymbol{\beta}_1^*)\,(Y-\text{ expit }\{g_1(P,M(0);\boldsymbol{\beta}_1^*)\})}{\text{expit }\{h(M(0),X;\boldsymbol{\gamma}^*)\}}$$

Estimation of $\boldsymbol{\beta}_0^*$ is similar but more complex since we do not observe $M(1)$ on those who did not wear glasses. We can estimate $\boldsymbol{\beta}_0$ by the unbiased estimating equation,

$$\begin{aligned}
&\boldsymbol{U}_{\boldsymbol{\beta}_0}(O^\dagger;\boldsymbol{\psi}^*) \\
&= E\left[\frac{(1-Z)g_0'(P,M(0);\boldsymbol{\beta}_0^*)\,(Y-\text{ expit }\{g_0(P,M(0);\boldsymbol{\beta}_0^*)\})}{(1-\text{expit }\{h(M(0),X;\boldsymbol{\gamma}^*))\}}\,\middle|\,O^\dagger\right] \\
&= \frac{\int_0^1(1-Z)g_0'(p,M(0);\boldsymbol{\beta}_0^*)Y(0)f(p|M(0),Z=1,X;\boldsymbol{\eta}^*,\phi^*)dp-}{(1-\text{expit }\{h(M(0),X;\boldsymbol{\gamma}^*))\}} \\
&\quad \frac{\int_0^1(1-Z)g_0'(p,M(0);\boldsymbol{\beta}_0^*)\text{expit }\{g_0(p,M(0);\boldsymbol{\beta}_0^*)\}f(p|M(0),Z=1,X;\boldsymbol{\eta}^*,\phi^*)dp}{(1-\text{expit }\{h(M(0),X;\boldsymbol{\gamma}^*))\}}
\end{aligned}$$

5.6 Large Sample Theory

To estimate $\boldsymbol{\psi}^*$, denoted $\widehat{\boldsymbol{\psi}}=\{\widehat{\boldsymbol{\beta}}_0',\widehat{\boldsymbol{\beta}}_1',\widehat{\boldsymbol{\gamma}},,\widehat{\boldsymbol{\eta}}',\widehat{\phi}\}$, we can solve the following estimating equation.

$$\sum_{i=1}^{n}\mathbf{U}(O_i^\dagger;\boldsymbol{\psi})=0$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper113

where

$$\boldsymbol{U}(O_i^\dagger;\boldsymbol{\psi}) =$$

$$\left[\boldsymbol{U}_{\boldsymbol{\beta}_0}(O_i^\dagger;\boldsymbol{\psi})', \boldsymbol{U}_{\boldsymbol{\beta}_1}(O_i^\dagger;\boldsymbol{\psi})', \boldsymbol{U}_{\boldsymbol{\gamma}}(O_i^\dagger;\boldsymbol{\psi})', \boldsymbol{U}_{\boldsymbol{\eta}}(O_i^\dagger;\boldsymbol{\psi})', U_{\phi}(O_i^\dagger;\boldsymbol{\psi})\right]'$$

Where $i = 1, ..., n$ indexes the study participants and $E[\boldsymbol{U}(O_i^\dagger;\boldsymbol{\psi}^*)] = 0$ as demonstrated above. Solving the estimating equation is equivalent to estimating $\boldsymbol{\gamma}^*$, $\boldsymbol{\eta}^*$, and ϕ^* via maximum likelihood, estimating $\boldsymbol{\beta}_1^*$ by a weighted logistic regression, and estimating $\boldsymbol{\beta}_0^*$ by numerical optimization. Under mild regularity conditions (Huber 1964), by the theory of M-estimation, it can be shown that for $\boldsymbol{\psi}^* \in \boldsymbol{\psi}$,

$$\sqrt{n}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*) \xrightarrow{D} Normal(0, \Sigma^*)$$

where

$$\Sigma^* = E\left[\frac{\partial \boldsymbol{U}(O_i^\dagger;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right]^{-1} E\left[\boldsymbol{U}(O_i^\dagger;\boldsymbol{\psi}^*)\boldsymbol{U}(O_i^\dagger;\boldsymbol{\psi}^*)'\right] E\left[\frac{\partial \boldsymbol{U}(O_i^\dagger;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right]^{-1'}$$

and $\frac{\partial \boldsymbol{U}(O_i^\dagger;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}$ is evaluated at $\boldsymbol{\psi}^*$. The variance-covariance Σ^* can be estimated by

$$\widehat{\Sigma} = E_n\left[\frac{\partial \boldsymbol{U}(O_i^\dagger;\widehat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}}\right]^{-1} E_n\left[\boldsymbol{U}(O_i^\dagger;\widehat{\boldsymbol{\psi}})\boldsymbol{U}(O_i^\dagger;\widehat{\boldsymbol{\psi}})'\right] E_n\left[\frac{\partial \boldsymbol{U}(O_i^\dagger;\widehat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}}\right]^{-1'}$$

where $E_n[U_i] = \frac{1}{n}\sum_{i=1}^n U_i$. Natural estimates of $\nu(p,m)$, $\pi_0(p,m) = P[Y(0) = 1|P = p, M(0) = m]$ and $\pi_1(p,m) = P[Y(1) = 1|P = p, M(0) = m]$ would be $\pi_0(p,m;\widehat{\boldsymbol{\beta}}_0) = \text{expit}\{g_0(p,m;\widehat{\boldsymbol{\beta}}_0)\}$, $\pi_1(p,m;\widehat{\boldsymbol{\beta}}_1) = \text{expit}\{g_1(p,m;\widehat{\boldsymbol{\beta}}_1)\}$, $\nu(p,m;\widehat{\boldsymbol{\beta}}_0,\widehat{\boldsymbol{\beta}}_1) = \widehat{\pi}_1(p,m)/\widehat{\pi}_0(p,m)$. Since the distribution of estimators of relative risk are better approximated by the normal distribution on the log scale, let $\xi = \log \nu(p,m;\boldsymbol{\beta}_0^*,\boldsymbol{\beta}_1^*)$. By the δ-method, we know that,

$$\sqrt{n}(\widehat{\xi} - \xi^*) \xrightarrow{D} Normal\left(0, \frac{\partial \xi}{\partial \boldsymbol{\psi}}\Sigma^*\frac{\partial \xi}{\partial \boldsymbol{\psi}}'\right)$$

where $\partial\xi/\partial\boldsymbol{\psi}$ is evaluated at ξ^*. We can exponentiate the bounds of the confidence interval derived for ξ^* to obtain the confidence interval for $\nu(p,m;\boldsymbol{\beta}_0^*,\boldsymbol{\beta}_1^*)$.

6 Analysis

The SEE study had 2,520 participants. Of these, 170 did not report any significant outside sun exposure in their thirties. Since we defined eyeglass use based on outside use, this group was not eligible to have the exposure of interest and was excluded from the

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 1: Baseline characteristics of sample

Variable	No Eyeglass Use	Eyeglass Use
Number of participants	830 (42%)	1125 (58%)
Sun exposure if glasses never worn, M(0)	.17 (.11)	.16 (.11)
Age	73.5 (5.0)	72.7 (4.8)
Diabetic	17.4%	17.2%
Male	54.6%	39.9%
Black	30.7%	22.1%
<12 years education	58.0%	45.6%
Job characteristics		
Worked over water	1.7%	1.2%
Worked outside on land	41.1%	28.5%
Worked inside	38.9%	44.2%
Worked as homemaker	18.3%	26.1%

analysis. A further 171 participants were excluded due to missing covariate or outcome data not related to prior eye-surgery. Finally, 221 participants were excluded based on having prior surgery that precluded assessment of cataracts. Three participants with extreme values of baseline UVB exposure ($M(0) > .54$) were excluded from the study to avoid leverage problems in estimation; our results cannot be extrapolated to those with extreme baseline UVB exposure. This left 1,955 in our study; 1,125 (58%) wore glasses outside in their early thirties, 830 (42%) did not.

Table 1 notes the baseline characteristics of the sample. There were not substantial differences between those in the dataset who did and did not wear glasses based on age or diabetic status, although the difference in ages was statistically significant (p-value<0.05). However, those who did wear glasses outside were more likely to be women, white, and more likely to have a high school education. The distribution of jobs between the two groups also differed.

Table 1 also reveals a statistically significant (p-value<0.05) baseline difference in UVB exposure that would have occurred if no one had worn glasses ($M(0)$) across the glass wearing groups, but the magnitude of the difference was small. The difference was likely related to the different job characteristics of those who wore glasses in the study. The difference in baseline exposure demonstrates the importance of including baseline

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper113

UVB exposure under no glasses in our propensity model to account for confounding. After weighting by the estimated propensity of having the eyeglasses use actually observed, the means and proportions of the covariates were almost identical across glass wearing groups, thus indicating that our propensity score models were self-consistent (Tan 2006). Further, the range of the estimated probabilities of wearing eyeglasses was similar between the two groups suggesting that there was sufficient overlap in the covariate space. None of the propensity weights were near 0 or 1.

We first assessed the evidence for mediation using a method similar to that advocated by Baron and Kenny; the major difference being that here we use a logistic rather than linear model. As noted before, Baron and Kenny's method is not directly generalizable to logistic models, although researchers have applied it in such settings. Table 2 presents the results of logistic models fit on the observed data. We see in Model 1 that wearing eyeglasses is protective of the development of cortical cataracts and is statistically significant after controlling for the possible confounders. In Model 2, the effect of glasses changes slightly and is no longer significant. In Model 3, glasses use is a significant predictor of UVB exposure after controlling for the possible confounders in a linear regression, which is a necessary condition for mediation to occur under Baron and Kenny's framework. This last finding was not surprising since glasses use was used to derive the observed UVB exposure in equation (5.1).

To practitioners of Baron and Kenny's method, the analysis on the observed data provides limited if any evidence of a mediational effect. After controlling for the hypothesized mediator, glasses use was no longer a statistically significant predictor of cataracts, but the magnitude of the estimated coefficient barely changes. However, the coefficient of UVB exposure is not statistically significant in the full model, which is not consistent with a hypothesis that UVB exposure is in the causal pathway. Since job status was related to UVB exposure, we repeated the analysis but excluded the job indicators from the logistic models; the results were not substantially changed. As discussed in Section 3, however, such an analysis is causally flawed.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 2: Logistic models of cataract development (with coefficients reported as odds ratios) and linear model of observed UVB exposure.

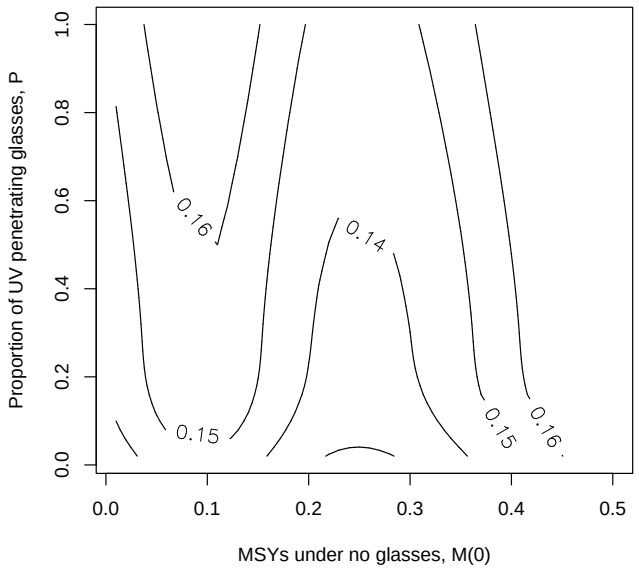
Variable	Model 1	95% CI	Model 2	95% CI
Cataract Models				
Age	1.17	(1.07, 1.28)	1.17	(1.07, 1.28)
Age spline term	0.89	(0.78, 1.03)	0.89	(0.78, 1.03)
Diabetic	1.43	(1.02, 2.00)	1.43	(1.02, 2.00)
Male	0.64	(0.45, 0.92)	0.63	(0.44, 0.91)
Black	4.23	(3.13, 5.72)	4.22	(3.12, 5.71)
<12 years education	1.10	(0.81, 1.48)	1.09	(0.81, 1.48)
Worked over water	Reference		Reference	
Worked outside	0.50	(0.20, 1.27)	0.52	(0.20, 1.32)
Worked inside	0.64	(0.25, 1.66)	0.70	(0.26, 1.90)
Worked as homemaker	0.54	(0.19, 1.51)	0.57	(0.20, 1.61)
Glasses	0.74	(0.56, 0.99)	0.78	(0.57, 1.09)
UVB			1.80	(0.30, 10.76)

Variable	Model 3	95% CI
UVB Model		
Intercept	0.31	(0.16, 0.45)
Age	0.00	(0.00, 0.00)
Age spline term	0.00	(0.00, 0.00)
Diabetic	0.00	(-0.01, 0.01)
Male	0.03	(0.02, 0.04)
Black	0.00	(-0.01, 0.01)
<12 years education	0.01	(0.00, 0.01)
Worked over water	Reference	
Worked outside	-0.05	(-0.08, -0.02)
Worked inside	-0.14	(-0.17, -0.11)
Worked as homemaker	-0.08	(-0.11, -0.05)
Glasses	-0.10	(-0.10, -0.09)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper113

Figure 2: Estimates of $P[Y(0) = 1 \mid P = p, M(0) = m]$: Probabilities of developing cataracts under no glasses within strata.

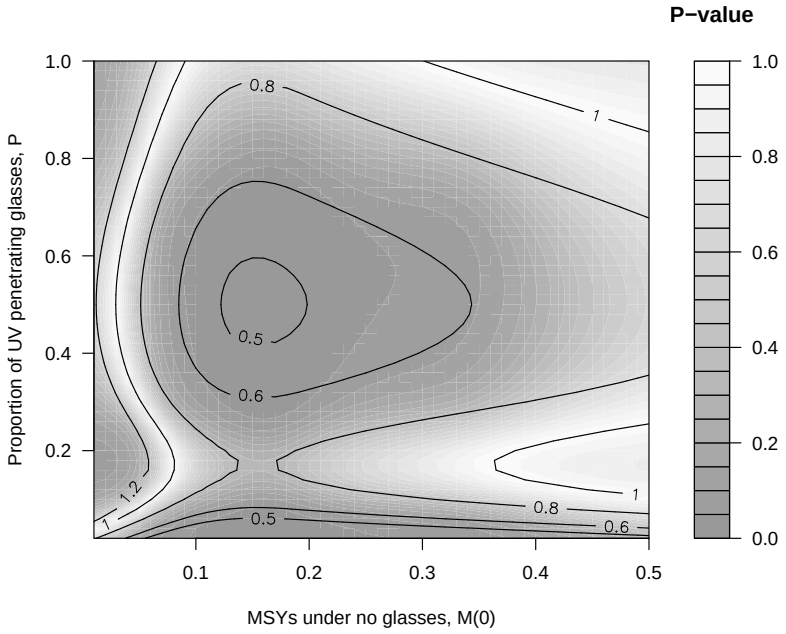


In the sample, 16.1% of those who did not wear glasses and 11.6% of those who did developed cortical cataracts. Figure 2 presents a contour plot of estimates of $P[Y(0) = 1|P, M(0)]$, the probabilities of developing cataracts under no glasses within levels of P and $M(0)$. There does not seem to be much impact of either UVB exposure or UVB reduction on the probability of developing cataracts under no glasses. The lack of a strong relationship between outcomes under no glasses and exposure under no glasses was not expected. We would expect that the probability of developing cataracts would increase as UVB exposure increases, as was found by West *et al.* (1998) This finding could be related to the data structure. We have missing cortical cataract data in those who had surgery; it is possible those in the dataset with high UVB exposure developed cataracts early and hence had surgery by the time of the study. Hence, we might be underestimating the true probability of developing cataracts for those with high levels of exposure.

Figure 3 presents estimates of $\nu(p, m)$, the relative risk of developing cataracts related to glasses use given the proportion of UVB rays not filtered by glasses and MSY exposure under no glasses. As expected, glasses do not have a protective effect on cataract development when baseline exposure under no glasses is 0 or when glasses do not filter any UVB. The RR is generally near or greater than 1 when $P = 1$ or $M(0) = 0$;

Hosted by The Berkeley Electronic Press

Figure 3: Estimates of $\nu(P = p, M(0) = m)$: RR of developing cataracts under glasses versus under no glasses given strata.



however, there are relatively few observations with P near 1 or $M(0)$ near 0 with which to estimate such effects precisely. The contour plot achieves a minimum when $P = 0.50$ and $M(0) = 0.15$ MSYs (RR=0.48, 95% CI 0.28, 0.83). Darker areas of the plot indicate regions in which the effect was statistically significant; that is, the upper bound of the 95% confidence interval did not cross 1. We see that glasses are protective at moderate levels of MSYs under no glasses and moderate reductions in the proportion of UVB rays reaching the eye.

In the $M(0)$ direction, the point estimates of the relative risk range from 0.82 to 1.97 when there is no baseline exposure ($M(0) = 0$). No relative risk estimates were statistically significant near $M(0) = 0$. The RR decreases as $M(0)$ increases near 0; this suggests that glasses are more protective as baseline UVB exposure increases. This intuitively makes sense since glasses cannot have much of an impact on cataracts via reduction in UVB exposure when people do not have much baseline exposure. The RR starts to increase slightly at a baseline exposure of approximately 0.15 MSYs, although the effect is still generally protective as $M(0)$ increases. This might indicate that glasses are not as effective for an individual with high baseline exposure if the individual is still receiving substantial UVB exposure under glasses. The effect of UVB on cataract development might be nonlinear and there could be a threshold effect in which exposure

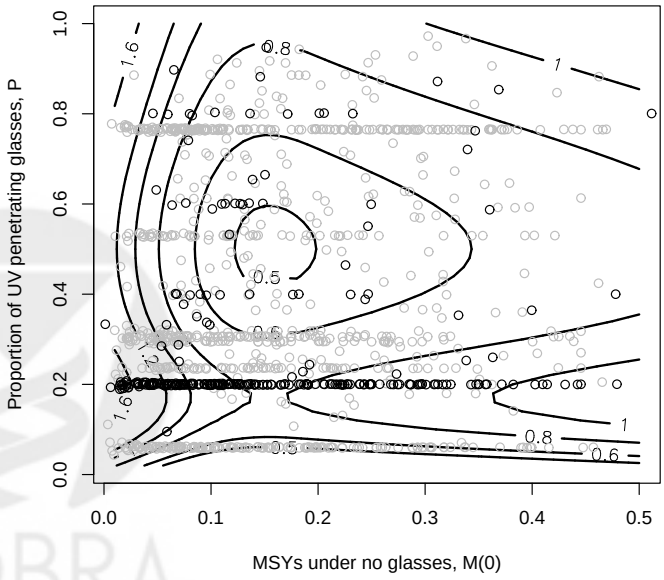
COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper113

above a certain level does not significantly increase cataract risk.

Another reason that the hyperplane might increase at higher levels of $M(0)$ might be due to nonrandom missing data among those who had surgeries. Indeed, 221 (10.14%) of those eligible for inclusion in the study had surgery. However, it is unlikely that this significantly affected the results. While the rate of cortical cataracts was likely higher in those who had surgery, the increased rate was probably not high enough to change the inferences. Among subjects who had unilateral surgery in the SEE data, a group probably comparable to those who had bilateral surgery, 17.9% of participants had cortical cataracts noted in the good eye. This is slightly lower than the 21.5% pre-surgery rate of cortical cataracts for those who underwent surgery reported by Lewis et al. who also used SEE data (Lewis *et al.* 2004) This compares with a rate of 13.0% in the SEE population without a history of unilateral or bilateral surgery. Hence, the vast majority of those who had surgery probably did not have surgery due to cortical cataracts and hence the addition of the unseen outcomes would be unlikely to change the overall results.

Figure 4: Estimates of $\nu(P = p, M(0) = m)$ and P vs. $M(0)$ among those who wore glasses; jitter of points used for clarity. Those who wore corrective glasses but not sunglasses are noted in black.



COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

In the P direction, the RR ranges from 0.84 to 1.36 when P=1, except for the effect at very small values of $M(0)$ when the RR reaches values up to 1.97. the relative

Hosted by The Berkeley Electronic Press

risk decreases as P decreases near 1. Again, this intuitively makes sense in that glasses are likely to be more effective at reducing the risk of cataracts as they block a larger proportion of UVB rays. At approximately $P = .50$, the hyperplane begins to increase in the P direction before decreasing again, although the increase is generally slight. The shape of the hyperplane in the P direction may result from differences in protective effects of corrective glasses versus sunglasses. Figure 4 highlights the values of P vs. $M(0)$ among those who wore corrective glasses but not sunglasses. We see many of those who wore corrective glasses only have values of P around 0.2, near the peak of the uptick in the hyperplane. It is possible those who wore corrective glasses only had different strategic behavior in their wearing of eyeglasses such that eyeglasses were not as protective of cataracts in this group. There did not seem to be any other striking data patterns that could explain the non-monotonicity of the hyperplane.

7 Discussion

Our causal analysis provides evidence that the protective effect of eyeglasses on cataracts is mediated by their effect on reducing exposure to sunlight UVB. The shape of the hyperplane is as we would expect it would be for complete mediation to occur. There is no statistically significant effect of glasses on cataracts and the point estimates of the relative risk are generally around 1 when a person receives little baseline sunlight exposure or glasses have little shielding effect. However, eyeglasses are protective at moderate levels of baseline exposure or UVB shielding ability. The results are consistent with a hypothesis that there is no causal direct effect of eyeglasses on cataract development; the mechanism by which eyeglasses affect cataracts is through their impact on UVB exposure.

Our method not only provides important epidemiological evidence about the utility of wearing glasses outdoors, but also demonstrates the richness of analyses that are possible by using the potential outcomes framework to investigate mediation. The naive logistic analysis gave only weak support for a mediated effect; the magnitude of the coefficients and statistical significance between the reduced and full models did not change very much. In contrast, the causal analysis gave a rich depiction of the extent to which the effect of glasses on cataracts is likely mediated by reducing UVB exposure. We were able to observe not just a mediational effect, but also able to see how much

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper113

data was available to support our findings.

The SEE data and nature of this problem afforded us some unique opportunities. Because the ability of eyeglasses to reduce UVB exposure is scientifically well-defined, the counterfactual exposure level was available for those in the dataset who wore glasses. This gave us valuable information that helped us identify the joint distribution of the potential outcomes. The fact that the counterfactuals could be considered observable provides a case study of a situation when a potential outcomes analysis is not inherently epistemological.

There were limitations to this study. One was that we did not have complete outcome information on those who had a prior history of surgery. However, as discussed above, it is unlikely that the inclusion of that information would dramatically change the overall conclusions of this study. Another limitation comes from the retrospective cohort nature of the SEE data. We did not follow participants prospectively from their thirties into later life. Hence, our study is subject to recall biases and left truncation due to death. However, since the development of late life cortical cataracts is not likely to be associated with death, we do not think that losses due to death severely bias the results.

This study provides evidence that wearing eyeglasses in midlife can protect against the development of cataracts in later life. We also demonstrate how scientists can use the potential outcomes framework in order to define causal effects that are more meaningful than the direct and indirect effects of path analytic techniques. The potential outcomes framework allows us to define meaningful causal estimands, and also to investigate mediation when outcomes are not continuous in ways not possible under traditional path analysis.

8 References

Ayalon, L., Arean, P.A., Alvidrez, J. (2005), "Adherence to antidepressant medication in blacks and latino elderly patients," *American Association for Geriatric Psychiatry*, 13:572-780.

Baron, R.M., and Kenny, D.A. (1986), "The moderator-mediator variable distinction in social psychological research conceptual, strategic, and statistical considerations," *Journal of Personality and Social Psychology*, 51(6): 1173-1182.

CORRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Bridge, J.A., Day, N.L., Richardson, G.A., Birmaher, B., Brent, D.A. (2003), "Major depressive disorder in adolescents exposed to a friend's suicide," *Journal of the American Academy of Child and Adolescent Psychiatry*, 42(11):1294-1300.

Chang, M.A., Congdon, N.G., Bykhovskaya, I., Munoz, B., West, S.K. (2005), "The association between myopia and various subtypes of lens opacity," *Ophthalmology*, 112: 1395-1401.

Cruickshanks, K.J., Klein, B.E.K., Klein, R. (1992), "Ultraviolet light exposure and lens opacities: The Beaver Dam Eye Study," *American Journal of Public Health*, 82(12): 1658-1662.

Delcourt, C., Carriere, I., Ponton-Sanchez, A., Lacroux, A., Covacho, M-J., Papoz, L. (2000), "Light exposure and the risk of cortical, nuclear, and posterior subcapsular cataracts," *Archives of Ophthalmology*, 118: 385-392.

Dominici, F., Zeger, S.L., Parmigiani, G., Katz, J., Christian, P. "Estimating Percentile-Specific Causal Effects: A Case Study of Micronutrient Supplementation, Birth Weight, and Infant Mortality" (December 16, 2004). Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 68.

Duncan, D.D., Munoz, B., Bandeen-Roche, K., West, S.K. (1997), "Assessment of ocular exposure to Ultraviolet-B for population studies," *Photochemistry and Photobiology*, 66(5):701-709.

Ferrari, S., Cribari-Neto, F. (2004), "'Beta regression for modelling rates and proportions," *Journal of Applied Statistics*, 31(7): 799-815.

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

Fried, L.P., Ferrucci, L., Darer, J., Williamson, J.D., Anderson, G. (2004), "Untangling concepts of disability, frailty, and comorbidity: Implications for improved targeting and care," *Journal of the Gerontology: Medical Sciences*, 59(3):255-263.

Holland, P.W. (1988), "Causal inference, path analysis, and recursive structural equations models," *Sociological Methods*. 18: 449-484.

Huber, P.J. (1964), "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, 35(1):73-101.

Lewis, A., Congdon, N., Munoz, B., Bowie, H., Lai, H., Chen, P., West, S.K. (2004), "Cataract surgery and subtype in a defined, older population: The SEECAT Project," *British Journal of Ophthalmology*, 88(12):1512-7.

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper113

O'Leary, A., Hoff, C.C., Purcell, D.W., Gomez, C.A., Parsons, J.T., Hardnett, F., Lyles, C.M. (2005), "What happened in the SUMIT trial? Mediation and behavior change," *AIDS*, 19(Suppl 1):S111-S121.

Pearl, J. *Causality: Models, Reasoning, and Inference,* Cambridge: Cambridge University Press.

Pearl, J. (2001) Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (San Francisco, CA: Morgan Kaufmann). pp. 411-420.

Robins, J.M., Greenland, S. (1992), "Identifiability and exchangeabiltiy for direct and indirect effects," *Epidemiology*, 3(2):143-155.

Robins, J.M. (2003), "Semantics of causal DAG models and the identification of direct and indirect effects," in *Highly Structured Stochastic Systems* eds. P.J. Green, N.L. Hjort, and S. Richardson, Oxford: Oxford University Press.

Rubin, D.B. (1980), Comment on "Randomization Analysis of Experimental Data: The Fisher Randomization Test," by D. Basu, *Journal of the American Statistical Association*, 75:591-593.

Rubin, D.B. (2004), "Direct and indirect causal effects via potential outcomes," *Scandinavian Journal of Statistics* 31:161-170.

Tan, Z. (2006), "Efficient and robust causal inference: A distributional approach," *Journal of the American Statistical Association*, in press.

Taylor, J.M.G., Wang, Y., Thiebaut, R. (2005), "Counterfactual links to the proportion of treatment effect explained by a surrogate marker," *Biometrics* , 61(4):1102-1111.

van der Laan, M.J, Petersen, M.L. (2004), "Estimation of direct and indirect causal effects in longitudinal studies," (August 23, 2004). U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 155.

West, S.K., Munoz, B., Rubin, G.S., Schein, O.D., Bandeen-Roche, K., Zeger, S., German, S., Fried, L.P. (1997), "Function and visual impairment in a population-based study of older adults. The SEE project. Salisbury Eye Evaluation," *Investigative Ophthalmology and Visual Science*, 38(1):72-82.

West, S.K., Duncan, D.D., Munoz, B., Rubin, G.S., Fried, L.P., Bandeen-Roche, K., Schein, O.D. (1998), "Sunlight exposure and risk of lens opacities in a population-based study," *JAMA*, 280(8): 714-718.

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press