The International Journal of Biostatistics

Volume 6, *Issue* 2 2010 *Article* 14

CAUSAL INFERENCE

Comparing Approaches to Causal Inference for Longitudinal Data: Inverse Probability Weighting versus Propensity Scores

Ashkan Ertefaie, *McGill University*
David A. Stephens, *McGill University*

Recommended Citation:
Ertefaie, Ashkan and Stephens, David A. (2010) "Comparing Approaches to Causal Inference for Longitudinal Data: Inverse Probability Weighting versus Propensity Scores," *The International Journal of Biostatistics*: Vol. 6: Iss. 2, Article 14.
DOI: 10.2202/1557-4679.1198

Comparing Approaches to Causal Inference for Longitudinal Data: Inverse Probability Weighting versus Propensity Scores

Ashkan Ertefaie and David A. Stephens

Abstract

In observational studies for causal effects, treatments are assigned to experimental units without the benefits of randomization. As a result, there is the potential for bias in the estimation of the treatment effect. Two methods for estimating the causal effect consistently are Inverse Probability of Treatment Weighting (IPTW) and the Propensity Score (PS). We demonstrate that in many simple cases, the PS method routinely produces estimators with lower Mean-Square Error (MSE). In the longitudinal setting, estimation of the causal effect of a time-dependent exposure in the presence of time-dependent covariates that are themselves affected by previous treatment also requires adjustment approaches. We describe an alternative approach to the classical binary treatment propensity score termed the Generalized Propensity Score (GPS). Previously, the GPS has mainly been applied in a single interval setting; we use an extension of the GPS approach to the longitudinal setting. We compare the strengths and weaknesses of IPTW and GPS for causal inference in three simulation studies and two real data sets. Again, in simulation, the GPS appears to produce estimators with lower MSE.

KEYWORDS: inverse probability weighting, propensity scores, longitudinal data

Author Notes: The authors thank the reviewer and editor for their constructive comments that have improved the paper considerably. The first author is grateful for the support of a Schulich Graduate Fellowship in the Department of Mathematics and Statistics at McGill. The second author acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC).

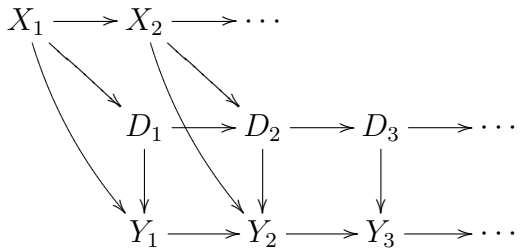
1 Introduction

Establishing the causal effect of a time-varying treatment in longitudinal studies is complicated because of the potential presence of time-varying confounding. Two methods of causal adjustment, *inverse probability of treatment weighting* (IPTW) and *propensity score* (PS) methods are commonly used. The two methods are constructed in a similar fashion; a model for treatment received is proposed and fitted, and then a regression model for the conditional expectation of the response variable is fitted either using weighting (IPTW) or matching/conditioning (PS). However, the precise implementation details differ, and our study will demonstrate that the use of the treatment received model has an influence on the quality of the estimation.

The theoretical properties of these two types of adjustment procedure have been studied, but rarely directly compared. Tan (2007) shows that, under correct model specification, the IPTW estimator is no more efficient than the outcome regression estimator which assumes a parametric model for $\mathrm{E}[Y|D = j, X]$ in estimating the expectations $\mathrm{E}[Y(j)]$, $j = 0, 1$, using the Rao-Blackwell theorem. Hirano et al. (2003) show that an estimator based on weighting by the reciprocal of the estimated propensity score is asymptotically equivalent to an efficient estimator that directly controls for all pretreatment variables, such as the estimator of Hahn (1998). On the other hand, Robins et al. (1992) shows that the least-squares estimator based on regressing on the correctly specified propensity score can have variance no less than the semiparametric efficiency bound, but possibly larger.

In this paper, we examine these results numerically in a one interval case, and then for longitudinal data. We look at the performance - specifically, the bias, variance and MSE - of the two methods for establishing the magnitude of a *direct* effect of treatment, that is, the *unconfounded* and *unmediated* effect on expected response. In simulation, we find that propensity score methods seem to give estimators with smaller variance and lower mean square error. Note that our focus is on direct effects, as this is the only setting in which IPTW and PS can be readily compared, although IPTW adjustments also play a role in the estimation of other causal effects.

We consider a longitudinal study with treatment doses D_{ij}, responses Y_{ij}, and covariates X_{ij} for subjects $i = 1, 2, \ldots, n$ on repeated observation $j = 1, \ldots, K_i$. All variables including dose can be binary, categorical or continuous. A directed acyclic graphic (DAG) representation of the data generating processes we consider is depicted below:



Note that X_i confounds the effect of D_i on Y_i. We also admit the possibility that the confounder, treatment and response sequences exhibit autocorrelation. Under an assumption of time-homogeneity, the direct effect of D on Y can be assessed.

1.1 Causal Adjustment Methods

We begin with the following simple setting, identical to Robins, Hernan and Brumback (2000). For a given subject, let $Y(1)$ denote outcome if treated, and $Y(0)$ outcome if untreated. The causal effect of treatment on this subject is $Y(1) - Y(0)$, but in most cases, only one of the outcomes is observed for each subject.

In the binary outcome case we can use the following (structural) representation in the form of a logistic regression: $Y(d) \sim Bernoulli(p(d))$, and

$$\text{logit}\{p(d)\} = \text{logit}\{p(Y = 1|D = d)\} = \beta_0 + \beta_1 d \tag{1}$$

so that

$$\text{E}[Y(1)-Y(0)] = \frac{\exp\{\beta_0 + \beta_1\}}{1 + \exp\{\beta_0 + \beta_1\}} - \frac{\exp\{\beta_0\}}{1 + \exp\{\beta_0 + \beta_1\}} = \frac{\exp\{\beta_0\}(\exp\{\beta_1\} - 1)}{1 + \exp\{\beta_0 + \beta_1\}}$$

when treatment is unconfounded the estimated parameters are unbiased for the corresponding causal parameters; however, these parameters individually or jointly are not interpretable as an *average treatment effect* (ATE) on the observable scale. Replacing equation (1) by the linear relation

$$p(d) = \beta_0 + \beta_1 d$$

with the implicit constraints that $\beta_0 > 0$ and $0 < \beta_0 + \beta_1 < 1$ does render β_1 the ATE. This is equivalent to modelling the average causal risk difference between treated and untreated potential outcomes.

DOI: 10.2202/1557-4679.1198

Inverse Probability of Treatment Weighting (IPTW): If treatment is confounded, given data on measured confounders X and under the assumption of no unmeasured confounders, unbiased estimates of the causal parameter can be obtained using a weighted analysis. IPTW proceeds by using a weighted regression. For example, for a binary treatment (where subjects are simply untreated or treated), for each subject i the weight $w_i = 1 + e^{-\eta_i}$ is assigned, where

$$\eta_i = \text{logit}\{p(D = d_i|X = x_i)\} = \alpha_0 + \alpha_1 x_i \tag{2}$$

where (d_i, x_i) are the observed dose and confounder data for subject i, and w_i is the estimated inverse probability of treatment weight. This weight can then be used in a weighted regression of Y on observed D, and possibly components of X. Robins et al. (2000) generalized this idea to the cases with multilevel or continuous treatment. For a continuous treatment one can still obtain the unbiased estimates of the causal parameter via IPTW by fitting the regression model for D given $X = x$ to obtain stabilized weights

$$w_i^s \propto \frac{f(d_i)}{f(d_i|x_i)}$$

provided that the model in equation (2) is correctly specified. For example, for doses on $\mathbb{R}$, $f(d_i|x_i)$ can be modelled using a linear regression model to yield

$$\widehat{f}(d|x) = (2\pi\widehat{\sigma}^2)^{-1/2} \exp(-(d - (\widehat{\alpha}_0 + \widehat{\alpha}_1 x))^2/(2\widehat{\sigma}^2)).$$

To estimate the numerator $f(d)$, one might specify normal density with the average of observed D and empirical variance as a mean and variance of the density.

Propensity Score (PS): Another method which gives an unbiased estimator for causal effect is based on the propensity score. Rosenbaum and Rubin (1983) define the propensity score for binary treatment as

$$\pi(x) = p(D = 1|x)$$

and demonstrate that it is the coarsest function of covariates that has the *balancing property*, that is, where treatment assignment is independent of covariates given the propensity score, $D \perp X|\pi(X)$. The ATE can be computed using iterated expectation

$$\mu = \text{E}[Y(1) - Y(0)] = \text{E}_{\pi(X)}[\text{E}[Y(1)|\pi(X)] - \text{E}[Y(0)|\pi(X)]] \tag{3}$$

where $\mathrm{E}_{\pi(X)}$ denotes expectation with respect to the distribution of $\pi(X)$ in the entire population. That is, under the *strong ignorability* assumption which states that $(Y(0), Y(1)) \perp D|X$, subjects with the same value of propensity score but different treatments can be considered as a controls for each other, in the sense that the (conditional) expected difference in their responses equals the average treatment effect for that value of π.

Equation (3) suggests an estimator based on some form of conditional expectation modelling of Y given π, averaged over the empirical distribution of π. Typically this is achieved using a stratification (matching) estimator, or a regression-based estimator. Within each stratum, or in the regression model, other covariates or confounders may be included.

When treatment dose is a continuous variable, a relevant quantity of interest is the *Average Potential Outcome* (APO) at dose level d, $\mu(d)$. The APO can be estimated using the propensity score by noting that

$$\mu(d) = \mathrm{E}[Y(d)] = \mathrm{E}_X[\mathrm{E}[Y(d)|\pi(X)]].$$

A general method for producing an estimate of μ or $\mu(d)$ is described in detail in Section 2; first Model I for π given X is constructed and estimated, and then Model II for Y given D and π (and possibly X) is considered, with both models being estimated using the observed data $(y_i, x_i, d_i), i = 1, \ldots, n$. The estimate for μ or $\mu(d)$ is then obtained by predicting Y at the counterfactual dose d in Model II for each observed (x_i, π_i) pair, and averaging the predictions over all pairs in the sample. In the case of binary doses

$$\mu = \mu(1) - \mu(0).$$

Uncertainty estimates can be obtained by making parametric assumptions, or using the sandwich variance, or by bootstrap.

For propensity score-type methods, the ATE and the APO can be estimated consistently in the presence of confounding provided that the balancing property, that is, $D \perp X|T(X)$, holds given for some score, $T(X)$ say. Thus, for the balancing score approach to estimate the causal parameter without bias, correct specification of $\pi(X)$ is a sufficient but not necessary condition. The adequacy of **any** proposed score model rests on whether or not balance is achieved; this can be checked by examining (in sample) the distribution of covariates/confounders X for different values (strata) of D for each of a collection of values (strata) of π. Note that although any score $T(X)$ that achieves balance will provide unbiased estimates of $\mu(d)$, whose variance will depend on the specific definition of $T(X)$.

DOI: 10.2202/1557-4679.1198

Assumptions: As with all models for observational data, causal models require certain modelling assumptions to be appropriately specified (Robins, 1997; Robins et al., 2000). Specifically, we make the *stable unit treatment value assumption* (Rubin, 1978), which states that a subject's outcome is not influenced by other subjects' treatment allocation. We further assume *weak unconfoundedness*: for all $d \in D$, the potential outcome $Y_i(d)$ and the dose received D_i are presumed conditionally independent given the covariates X_i, that is $Y_i(d) \perp D_i | X_i$ which implies the no unmeasured confounders assumption. Note that weak unconfoundedness is weaker than the strong ignorability assumption as it does not require joint independence of all potential outcomes.

1.2 Simulated Examples

A simple single interval example illustrates the potential differences between results obtained for the causal parameter of interest using IPTW and PS methods. Suppose that the causal (structural) relationship of interest is encapsulated by the equation

$$Y_i = \beta_0 + \beta_1 D_i + \xi_i$$

where $\text{Cov}[D_i, \xi_i] \neq 0$; in fact, suppose that

$$\begin{aligned} D_i &\sim Bernoulli(\text{expit}\{\alpha_0 + \alpha_1 X_i\}) \\ \xi_i &= \gamma_0 + \gamma_1 X_i + \sigma\epsilon_i \end{aligned}$$

where $X_i \sim \mathcal{N}(0, \tau)$. Regressing Y on D without adjustment will lead to bias; we consider fitting correctly specified models for D in terms of X to obtain the probability weights/propensity score, and examine in simulation the bias, variance and mean square error (MSE) of the estimators. We take two different sample sizes ($n = 300, 5000$) and use 10000 replicates. The four causal adjustment methods are

- IPTW: Inverse probability weighting, with weight w_i given by

$$w_i^{-1} = \widehat{\pi}_i = \frac{\exp\{D_i(\widehat{\alpha}_0 + \widehat{\alpha}_1 X_i)\}}{1 + \exp\{(\widehat{\alpha}_0 + \widehat{\alpha}_1 X_i)\}}$$

- IPTW.t: Inverse probability weighting, with data point retained only if fitted probability of treatment received satisfies $0.05 < \widehat{\pi}_i < 0.95$

- PS (match): propensity score quintile matching, where estimator is

$$\frac{1}{5n}\sum_{s=1}^{5}\left\{\frac{\sum_{i=1}^{n}\mathbb{I}\{D_i=1\ \&\ \widehat{\pi}_i\in Q_s\}Y_i}{\sum_{i=1}^{n}\mathbb{I}\{D_i=1\ \&\ \widehat{\pi}_i\in Q_s\}}-\frac{\sum_{i=1}^{n}\mathbb{I}\{D_i=0\ \&\ \widehat{\pi}_i\in Q_s\}Y_i}{\sum_{i=1}^{n}\mathbb{I}\{D_i=0\ \&\ \widehat{\pi}_i\in Q_s\}}\right\}$$

  where $Q_1,\ldots,Q_5$ form the quintile partition of the sample space, and

$$\widehat{\pi}_i=\frac{\exp\{(\widehat{\alpha}_0+\widehat{\alpha}_1X_i)\}}{1+\exp\{(\widehat{\alpha}_0+\widehat{\alpha}_1X_i)\}}$$

- PS (regress): propensity score regression, where estimator is OLS estimator of $\beta_1^{(R)}$ in the model

$$Y_i=\beta_0^{(R)}+\beta_1^{(R)}D_i+\beta_2^{(R)}\widehat{\pi}_i+\epsilon_i$$

  In fact, $\beta_1^{(R)}$ is identical to the causal parameter β_1 in this model, as, using the earlier argument

$$\begin{aligned}\mu(d)=\mathrm{E}[Y(d)] &= \mathrm{E}_\pi[\mathrm{E}[Y(d)|\pi]]=\mathrm{E}_\pi[\beta_0^{(R)}+\beta_1^{(R)}d+\beta_2^{(R)}\pi]\\ &= \beta_0^{(R)}+\beta_1^{(R)}d+\beta_2^{(R)}\mathrm{E}_\pi[\pi]\end{aligned}$$

  and hence

$$\beta_1^{(R)}=\mu(1)-\mu(0)=\mu=\beta_1.$$

The parameter settings used were arbitrarily chosen to introduce relatively high dependence between D and X. This simulation is somewhat unrealistic, as X would be typically conditioned upon in the regression equation as well as the treatment model, but serves to illustrate the strengths of the various estimators.

In a second simulation, the same causal relationship is used, but where

$$\begin{aligned}D_i &\sim Bernoulli(\mathrm{expit}\{\alpha_0+\alpha_1X_i+\alpha_3Z_i\})\\ \xi_i &= \gamma_0+\gamma_1X_i+\gamma_2Z_i+\sigma\epsilon_i\end{aligned}$$

where $X_i, Z_i\sim\mathcal{N}(0,\tau)$. Table 2 contains results for the different adjustment methods with the confounder Z omitted from the estimation models.

In the first simulation, all adjustment methods are effective in bias removal, but there is a marked difference in terms of variance and MSE with PS re-

Table 1: Bias, Variance and MSE for $\beta_1 = 10$ using a correctly specified model. Settings of the other parameters are $\gamma_0 + \beta_0 = 10, \gamma_1 = 0.8, \alpha_0 = 1.0, \alpha_1 = 0.6, \sigma = \tau = 0.5$.

	$n = 300$			$n = 5000$		
	Bias	s.d.	MSE	Bias	s.d.	MSE
Unadjusted	5.574	0.688	31.545	5.572	0.168	31.071
IPTW	0.873	1.985	4.702	0.178	1.135	1.321
IPTW.t	0.022	1.060	1.123	0.001	0.262	0.069
PS (match)	0.640	1.226	1.911	0.708	0.402	0.663
PS (regress)	0.008	0.894	0.798	0.003	0.218	0.048

Table 2: Bias, Variance and MSE for $\beta_1 = 10$ in incorrectly specified model. Settings of the other parameters are $\gamma_0 + \beta_0 = 10, \gamma_1 = 0.8, \gamma_2 = 0.1, \alpha_0 = 1.0, \alpha_1 = 0.6, \alpha_3 = -0.4, \sigma = \tau = 0.5$.

	$n = 300$			$n = 5000$		
	Bias	s.d.	MSE	Bias	s.d.	MSE
Unadjusted	5.039	0.739	25.940	5.050	0.183	25.531
IPTW	-0.084	1.993	3.980	-0.663	1.002	1.444
IPTW.t	-0.955	1.129	2.186	-0.974	0.275	1.024
PS (match)	-0.313	1.294	1.773	-0.234	0.412	0.224
PS (regress)	-0.961	0.935	1.799	-0.951	0.228	0.956

gression providing the estimator with the lowest MSE, outperforming IPTW with truncation, IPTW.t, as its nearest competitor, with the discrepancy diminishing as sample size increases. In the second simulation, significant finite sample bias remains even for very large sample sizes due to the unmeasured confounding; however, again, PS methods seem to produce estimators with lower mean square error. In this example, matching outperforms regression, but both methods seem superior to IPTW. About 35 % of the IPTW samples are excluded by the truncation mechanism in these simulations, but note the MSE reduction achieved by truncation; note also that truncation increases the magnitude of bias, but reduces variance. Note also that the experimental treatment assignment assumption - that each subject has a non-negligible probability of being treated and of not being treated - is required for IPTW, and may explain some of the bias observed in the IPTW results. This assumption is also required for PS matching, and may explain the larger bias in the PS (match) results. Note, however, that PS matching still appears to produce

a smaller mean-square error than IPTW when the experimental treatment assignment assumption is violated, and that PS regression is unaffected.

The issue addressed in this paper is the variance/MSE improvement of PS regression methods over IPTW. The simulation in Tables 1 and 2 implies that PS regression methods produce estimators with lower variance for even moderately large sample sizes, at least when treatment models are correctly specified. We investigate this supposition in longitudinal studies.

We address direct effects of treatment, but this may not reflect the inferential objective of the study in all cases. In longitudinal studies, treatment regimes followed over time may have different effects on *overall* outcome, that is, some response measured only at the end of the study. Marginal structural models (MSMs) are a class of causal models for the estimation from observational data, of the causal effect of a time-dependent exposure in the presence of time-dependent covariates. Typically, MSMs are utilized to estimate the *total* (causal) effect of treatment on an end of study outcome. The parameters of MSMs can be consistently estimated using IPTW, but not by PS methods, as conditioning on the propensity score explicitly blocks (in DAG terms) the path between treatment and subsequent response.

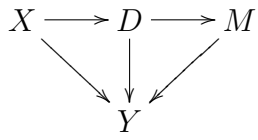
1.3 Binary Treatment with a Mediating Variable

Another case where PS performs better than IPTW in terms of MSE is in the presence of mediating variable, and it also appears that PS method is more successful in removing bias. In this section we report results of a small simulation study with a mediating variable. We have one time independent covariate X_i, one posttreatment intermediate variable M_i that may serve as a mediator for the treatment outcome, a treatment indicator D_i and response Y_i. We use the following densities:

$$\begin{aligned} X_i &\sim \mathcal{N}(2, 10) \\ M_i &\sim \mathcal{N}(d_i, 5) \\ D_i &\sim Bernoulli(p(x_i)) \qquad p(x_i) = \frac{1}{1 + \text{expit}\{-2 - 0.2x_i\}} \\ Y_i &\sim \mathcal{N}(-d_i + x_i + m_i, 5) \end{aligned}$$

DOI: 10.2202/1557-4679.1198

for $i = 1, ..., n$. Note that M can be written as $d_i + \mathcal{N}(0, 5)$ so the true ATE is zero. A DAG representation of the data generating processes is as follows:



We generated 1000 data sets of size 300 and 5000. We have used the following models for response variable using PS and IPTW and propensity score:

$$\text{logit}\{\pi_i\} = \text{logit}\{p(D_i = 1|X)\} = \alpha_0 + \alpha_1 x_i$$
$$y_i = \beta_0 + \beta_1 d_i + \beta_2 \pi_i + \epsilon_i \qquad \text{(PS)}$$

$$w_i^{-1} = \frac{\exp\{D_i(\alpha_0 + \alpha_1 X_i)\}}{1 + \exp\{(\alpha_0 + \alpha_1 X_i)\}}$$
$$y_i = \beta_0' + \beta_1' d_i + \beta_2' m_i + \epsilon_i \qquad \text{(IPTW)}$$

We also utilize the truncated version of IPTW, IPTW.t, in which we have retained only those observations with either $0.05 \leq \pi_i \leq 0.95$, where $\pi_i = w_i^{-1}$. Using the idea in VanderWeele (2009) we fitted another model considering M as a response variable to deal with counterfactuals in M,

$$\widehat{m}_i = \lambda_0 + \lambda_1 d_i$$

so that the total causal effect using IPTW is $\beta_1' + \beta_2' \lambda_1$. Table 3 shows the estimated ATE based on IPTW, IPTW.t and PS.

Table 3: ATE estimates based on IPTW, IPTW.t and PS for causal parameter $\beta^* = 1$, yielding ATE $= 0$.

	$n = 300$			$n = 5000$		
	ATE	s.d.	MSE	ATE	s.d.	MSE
IPTW	1.729	4.083	19.657	0.207	2.017	4.113
IPTW.t	0.174	1.907	3.667	0.015	0.435	0.190
PS	0.088	1.402	1.973	0.001	0.332	0.110

Under the assumption of correct model specification for the probability of treatment, IPTW has larger bias and standard deviation compared to PS for $n = 300$ and $n = 5000$. In the presence of the mediator, the PS method is

more successful in removing the bias and also has smaller variance than the IPTW methods.

The remainder of this paper is structured as follows: Section 2 introduces generalized propensity score (GPS) methodology. Section 3 develops the GPS for repeated measures data. Section 4 compares the repeated measures GPS with traditional regression and IPTW methods via simulations, two real data sets. Our simulation studies will help us to see the performance of these two estimators under correctly specified models and real data data examples show if one of them can outperform the other method in the presence of possible model misspecification.

2 The Generalized Propensity Score

In this section we define the Generalized Propensity Score which is the generalization of the classical binary treatment propensity score. We first examine the single interval case. When treatment is a continuous random variable, it is possible to construct a balancing score using an approach based on the *Generalized Propensity Score* (GPS). Following Imbens (2000) and Hirano and Imbens (2004), we define the (observed) GPS, $\pi(d, x)$ for dose d and covariate x by

$$\pi(d, x) = f_{D|X}(d|x) \tag{4}$$

that is, the conditional mass/density function for D given $X = x$ evaluated at $D = d$. Additionally $\pi(d, X)$ and $\pi(D, X)$ are corresponding random quantities. It has been shown by Hirano and Imbens (2004) that GPS random quantity $\pi(d, X)$ acts as a balancing score, in that D and X are conditionally independent given $\pi(d, X)$. Secondly, for any d, the allocation of the treatment dose is conditionally independent of the potential response, given the propensity score, $Y(d) \perp D \,|\, \pi(d, X)$, that is, we have unconfoundedness of $Y(d)$ and D given $\pi(d, X)$. Therefore $\pi(d, X)$ breaks the dependence between D and X, and hence the causal effect of D on X can be estimated by conditioning on $\pi(d, X)$ for each d in turn, and then averaging over the distribution of $\pi(d, X)$. The role of the GPS in estimating the APO is made clear by identity given in Imbens (2000)

$$\mu(d) = \mathrm{E}[Y(d)] = \mathrm{E}_X[\mathrm{E}[Y(d)|\pi(d, X)]] = \mathrm{E}_{\pi(d,X)}[\mathrm{E}[Y(d)|\pi(d, X)]]$$

We used the same algorithm to estimate the APO as Moodie and Stephens (2010) here:

I Form the GPS Model: Using the regression approach, construct the propensity model for D given X, $\pi(d,x) = f_{D|X}(d|x,\alpha)$. Estimate parameters α using data $\{(d_i, x_i), i = 1, \ldots, n\}$.

II Compute the Fitted GPS Model: Compute the estimated GPS,

$$\widehat{\pi}_i = f_{D|X}(d_i|x_i, \widehat{\alpha}).$$

III Form the Observable Model: Using the regression approach, construct a predictive model for the conditional expectation of density

$$f_{Y|D,\pi(d,X)}(y|d, \pi(d,x), \beta).$$

Estimate parameters β using data $\{(y_i, d_i, \widehat{\pi}_i), i = 1, \ldots, n\}$.

IV Estimate the APO: Estimate the APO for each d by

$$\widehat{\mu}(d) = \widehat{\mathrm{E}}[Y(d)] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{Y|D,\pi(d,X)}[Y_i(d)|d, \widehat{\pi}_i, \widehat{\beta}]$$

then $\widehat{\mu}(d)$ is the GPS-adjusted estimated dose-response function.

An alternative approach proposed by Hirano and Imbens (2004) suggests that the APO may be approximated by estimating the dose-response effect within strata defined by the linear predictor of the treatment density function, and then combining these estimates to form a single, weighted average. This approach is straightforward to implement and often provides an estimate of the dose-response relationship that has little or no residual bias, although it may be less efficient than the regression approach described above.

3 Causal Adjustment for Repeated Measures Data

3.1 The Multivariate GPS

In the case of dose response estimation from repeated measures or multi-interval data because of correlation structure in the data the potential patterns of time-varying confounding are more complex that can be dealt with using a univariate GPS approach. The GPS approach introduced in this section is suitable for the analysis of repeated measures response data with interval-dependent dosing. We denote Y_{ij} as a response of ith unit, $i = 1, ..., n$ in

interval j, $j = 1, ..., n_i$; dose and covariate variables are similarly subscripted. Furthermore, sequential weak unconfoundedness can be defined as

$$Y_{ij}(d) \perp D_{ij} | X_{i1}, ..., X_{ij}.$$

That is, at each interval, assignment to dose D_{ij} is weakly unconfounded with the response during interval j given covariates, previous response, and dose values measured up to the start of the jth interval. Moodie and Stephens (2010) show that if we define $\bar{X}_{ij} = (X_{1j}, ..., X_{ij})$ as a history of covariates, response and previous doses and let $\pi_{ij}(d, \bar{X}_{ij})$ be the multivariate GPS then, for every dose d,

$$Y_{ij}(d) \perp D_{ij} | \pi_{ij}(d, \bar{X}_{ij})$$

that is, for $d \in D$, current potential response $Y_{ij}(d)$ is conditionally independent of the distribution of dose received D_{ij} given the MGPS π_{ij} , for all i and j. In the same paper, it has also been shown that the APO obtained by averaging $\mathrm{E}[Y_{ij}(d)|\pi(d, \bar{X}_{ij})]$ over the distribution of the covariates $\bar{X}_{ij}$, is an unbiased estimator of the dose response function $\mu(d) = \mathrm{E}[Y_{ij}(d)]$. Note that a univariate GPS analysis that does not construct π by conditioning on $\bar{X}_{ij} = \bar{x}_{ij}$ for each j does not necessarily achieve bias removal.

We have carried out extensive testing of the MGPS approach and performed comparisons with non-causal and standard GPS methods. Our examples demonstrate the importance of the use of the multivariate extension of the GPS.

3.2 The IPTW Estimator for Repeated Measure Data

To implement IPTW in the repeated measures setting, the following model is fitted for response variable to estimate the total treatment effect,

$$\mathrm{E}[Y_{ij}|D_{ij} = d_{ij}] = \beta_0 + \beta_1 d_{ij}$$

with the stabilized weights

$$w_{ij}^s = \frac{p(D_{ij} = d_{ij}|D_{i(j-1)} = d_{i(j-1)})}{p(D_j = d_{ij}|D_{i(j-1)} = d_{i(j-1)}, \overline{X}_k = x_{ij})}$$

where $\overline{D}_{-1} = 0$. Thus the outcome at interval j is weighted with the inverse probability of treatment at that interval, modelled as a function of previous covariates, responses and doses. This is the natural extension of IPTW to the time-homogeneous repeated measures setting.

DOI: 10.2202/1557-4679.1198

Note the difference between the approach here and the typical MSM approach. Here we do not have a single response at the end of follow up, but responses and weights corresponding to each interval. Our weighted model produces the pseudo-populations based on observed treatment doses at each time point, rather than the pseudo-population through received treatment doses path up to end of follow-up.

4 Simulation Studies and Examples

4.1 Binary Treatment

In this section we report results of a small longitudinal simulation study carried out to evaluate the performance of the IPTW and PS explained in this paper. We have one time independent covariate X_{ij}, treatment indicator D_{ij} and response Y_{ij} with the following densities:

$$
\begin{aligned}
&X_{ij} \sim \mathcal{N}(1,2)\\
&D_{ij} \sim Bernoulli(\text{expit}\{\mathbb{I}\{j=1\}(2-x_{ij}) + \mathbb{I}\{j>1\}(2-0.2Y_{i(j-1)} - x_{ij})\})\\
&Y_{ij} \sim \mathcal{N}(D_{ij} + 2X_{ij}, 5)
\end{aligned}
$$

for $i = 1, ..., n$ and $j = 1, ..., 5$, where $\text{expit}\{x\} = \exp(x)/(1+\exp(x))$ and $\mathbb{I}\{.\}$ is an indicator function. We generated 1000 data sets of size 300 and 5000. Table 4 shows the estimated ATE based on IPTW and PS.

Table 4: ATE estimates based on IPTW, IPTW.t and PS for causal parameter $\beta^* = 1$, yielding ATE = 1.

	$n = 300$			$n = 5000$		
	ATE	s.d.	MSE	ATE	s.d.	MSE
IPTW	0.536	1.511	2.499	0.949	0.589	0.349
IPTW.t	0.946	0.952	0.910	1.004	0.235	0.055
PS	0.948	0.910	0.627	0.997	0.193	0.037

Under the assumption of correct model specification for weights and propensity score, IPTW method has larger bias and standard deviation compare to PS for $n = 300$ and $n = 5000$. Although weight truncation helps the IPTW method to reduce the MSE, it still has a slightly larger MSE than PS method. As we expected both methods are successful in removing the bias in the large sample size, $n = 5000$.

4.2 Simulation: Nonlinear, Nonadditive Treatment Effect

Here, we use the same simulation study as in Moodie and Stephens (2010) which is a version of the study presented in Hirano and Imbens (2004) extended to a two interval setting.

Data Generation: Suppose that at first and second interval, have

$$Y_1(d)|X_{11}, X_{12} \sim \mathcal{N}(d + (X_{11} + X_{12}) \exp(-d(X_{11} + X_{12})), 1)$$
$$Y_2(d)|X_{21}, X_{12} \sim \mathcal{N}(d + (X_{21} + X_{12}) \exp(-d(X_{21} + X_{12})), 1)$$

The marginal distribution of each of X_{11}, X_{12}, and X_{21} are all unit exponential and the marginal mean of the response in both intervals is identical. Let $D_1 \sim Exp(X_{11} + X_{12})$, $D_2 \sim Exp(X_{21} + X_{12})$. The APO at dose d, $\mu(d)$, can be obtained by integrating out the covariates analytically, yielding

$$\mu(d) = d + \frac{2}{(1+d)^3}.$$

In this section we want to compare the performance of estimators of APO based on IPTW, GPS and MGPS. As suggested by Robins et al. (2000), stabilized weights are estimated using a normal density for the IPTW analysis, and weighted splines has been used to fit the model for responses on dose. In GPS analysis, a multivariate GPS analysis, involves the GPS vector $\pi^M = (\pi_1, \pi_2)$:

$$\pi_1 = (X_{11} + X_{12}) \exp(-d(X_{11} + X_{12}))$$
$$\pi_2 = (X_{21} + X_{12}) \exp(-d(X_{21} + X_{12}))$$

where consists of correctly specified models. A univariate GPS analysis might fail to include information from the previous interval and hence the GPS used would be $\pi^U = (\pi_1, \pi_2^*)$ where π_1 is as before, but $\pi_2^* = X_{21} \exp(-dX_{21})$.

We generated 1000 data sets of size 250. The estimated APO using MGPS are exactly correct, while the UGPS and IPTW analysis are clearly biased. The general shape of the UGPS and IPTW APO are correct, however these estimators do not catch the curve (see Figure 1).

Table 5 shows the bias, variance and MSE's of the estimated APO using IPTW and MGPS. The bias and MSE obtained by MGPS are significantly smaller. As pointed out by Hirano et al. (2003), the efficiency of the GPS estimator can be improved by using the estimated GPS. In this simulation, the GPS can be estimated using a Gamma generalized linear model, for example.

DOI: 10.2202/1557-4679.1198

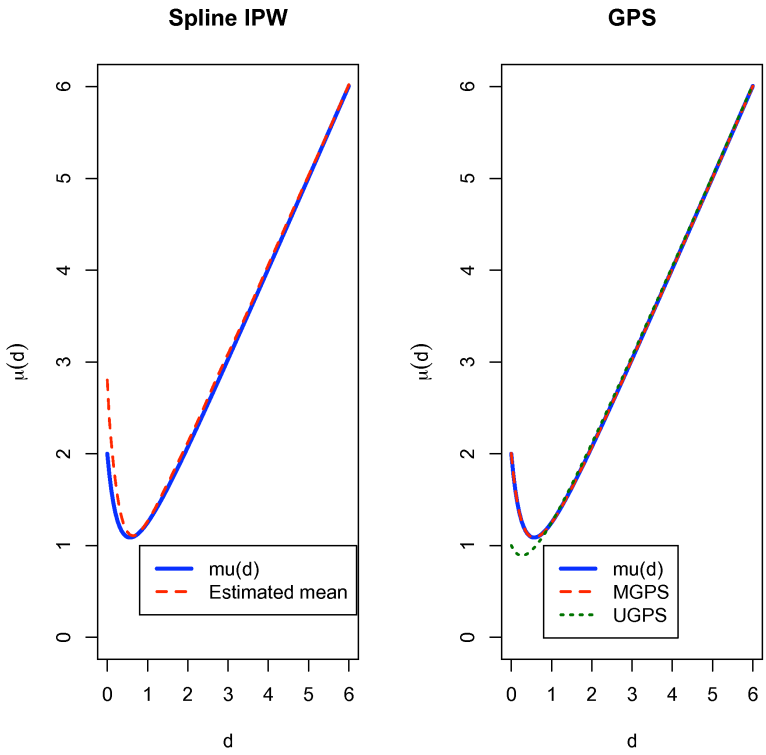


Figure 1: Simulated Example 2: The dose-response APO curves for the IPTW and GPS analyses.

4.3 Example: The MSCM Study

Alexander and Markowitz (1986) studied the relationship between maternal employment and paediatric health care utilization. The investigation was motivated by the major social and demographic changes that have occurred in the US since 1950. The Mothers' Stress and Children's Morbidity Study (MSCM) enrolled 167 preschool children between the ages of 18 months and 5 years that attended an inner-city paediatric clinic. Each individual provided information regarding their family and work outside the home. Daily measures of maternal stress and child illness were recorded during 4 weeks of follow-up. We use these data to determine casual effect of stress on child illness. We used logistic regression to fit the model for weights and propensity score over each interval with employment (e), married (m), previous stress (s) and previous illness (i)

Table 5: Pointwise bias estimates for causal curve based on IPTW using splines and GPS.

	IPTW			MGPS		
	$\mu(d) - \widehat{\mu}(d)$	Var	MSE	$\mu(d) - \widehat{\mu}(d)$	Var	MSE
$d = 0.05$	-0.595	0.063	0.417	0.001	0.043	0.043
$d = 0.10$	-0.519	0.030	0.300	0.001	0.030	0.030
$d = 0.20$	-0.350	0.019	0.141	0.000	0.016	0.016
$d = 0.55$	-0.037	0.018	0.020	0.000	0.005	0.005
$d = 0.65$	-0.016	0.023	0.024	0.000	0.005	0.005
$d = 1.00$	-0.012	0.030	0.030	0.000	0.006	0.006
$d = 2.50$	-0.053	0.102	0.105	0.001	0.006	0.006
$d = 5.50$	-0.023	0.317	0.317	0.000	0.009	0.009

as covariates, as follows:

$$\begin{aligned}\text{logit}\{p(s_{i1} = 1)\} &= \alpha_0 + \alpha_1 e + \alpha_2 m \\ \text{logit}\{p(s_{it} = 1)\} &= \beta_0 + \beta_1 s_{i(t-1)} + \beta_2 i_{i(t-1)} + \beta_3 e + \beta_4 m\end{aligned}$$

for $t = 1$ and $t > 1$ respectively. Since our response, illness, is a dichotomous random variable we fitted the following logistic models for IPTW and PS methods:

$$\text{logit}\{p(i_{it} = 1)\} = \gamma_0 + \gamma_1 s_{it} \qquad \text{(IPTW)}$$
$$\text{logit}\{p(i_{it} = 1)\} = \theta_0 + \theta_1 s_{it} + \theta_2 \pi_i(x) \qquad \text{(PS)}$$

where $\pi(x)$ is the propensity score. In order to see the effect of sample size in our estimators, we estimate the ATE for different sample sizes by randomly deleting individuals. Results are presented in Table 6.

Table 6: Parameter estimates based on IPTW and MGPS for MSCM study.

	$\widehat{\gamma}_1$	s.e.$(\widehat{\gamma}_1)$	$\widehat{\theta}_1$	s.e.$(\widehat{\theta}_1)$
$n = 50$	0.708	0.144	0.622	0.205
$n = 70$	0.644	0.129	0.617	0.189
$n = 100$	0.520	0.109	0.576	0.150
$n = 167$	0.547	0.083	0.537	0.115

As the sample size increases estimators become more similar, and for each sample size the IPTW standard errors are slightly smaller. Since there is a

large overlap between estimated parameter confidence intervals using IPTW and PS, neither is preferable to the other one in this example. We have also checked the truncated weights, IPTW.t, estimators, but the results are omitted because they were fairly similar.

In the next example we have a longitudinal data set with continuous response and treatment dose and we will compare the performance of univariate GPS, multivariate GPS and IPTW approaches.

4.4 Example: MOTAS Amblyopia Study

Amblyopia is a common childhood vision disorder characterized by reduced visual function in one eye which is often treated by occlusion therapy (patching) of the properly functioning eye. The Monitored Occlusion Treatment of Amblyopia Study (MOTAS) (Stewart et al. 2004) was the first clinical study aimed at quantifying the dose response relationship of occlusion, facilitated by the use of an electronic occlusion dose monitor. The MOTAS design and a full description of the study base have been published previously (Stewart et al. 2002, 2004); see also Moodie and Stephens (2010). The response variable, visual acuity was measured for each child at the ends of approximately two week intervals, and improvement corresponded to a *decrease* in the value of the response variable. We analyze the data of the 68 children who took part in the occlusion phase of MOTAS, who were prescribed six hours of occlusion daily, but received varying doses because of incomplete concordance. For child i, the response, Y_{ij}, is the change in visual acuity during interval j, and D_{ij} is the random occlusion dose (in hours) received in interval j.

In the study, 60 out of 404 (about 15%) of intervals in the occlusion phase had a zero dose, so we assume

$$D_{ij} \stackrel{\mathcal{L}}{=} \psi(\bar{x}_{ij}, \gamma)\mathbb{I}\{d = 0\} + (1 - \psi(\bar{x}_{ij}, \gamma))\mathbb{I}\{d \neq 0\}D^{+}_{ij}$$

where D^{+}_{ij} is strictly positive random variable and $0 < \psi(\bar{x}_{ij}, \gamma) < 1$ is a mixing weight which can be estimated using logistic model on binary ($D_{ij} = 0/D_{ij} > 0$) dose data.

Following the fitted model by Moodie and Stephens (2010), we included the visual acuity at start of interval, age, sex, interval number, length of interval (in days), and amblyopic type (anisometropic, strabismic, mixed) as a covariate in the GPS or IPTW model and if we add the previous dose to these covariates MGPS can be fitted. These covariates were used to predict both the probability of having any occlusion at all ($D/D > 0$) in a logistic

Table 7: Estimated parameters in APO models based on UGPS and MGPS, estimated variances are in brackets.

	UGPS	MGPS
β_1	-0.107(0.031)	-0.135(0.046)
β_2	9.00e-6(1.84e-4)	1.740e-4(2.28e-4)
β_3	2.917(1.580)	5.668(2.264)
β_4	0.080(0.047)	0.069(0.066)

model and the probability of receiving a particular dose (greater than zero) of occlusion in a Gamma model. The UGPS used is

$$\widehat{\pi}(d, x_{ij}) = \widehat{\psi}(x_{ij}, \widehat{\gamma})\mathbb{I}\{d = 0\} + (1 - \widehat{\psi}(x_{ij}, \widehat{\gamma}))\mathbb{I}\{d \neq 0\} f(d|x_{ij}, \widehat{\phi}, \widehat{\alpha})$$

where $f(d|x_{ij}, \widehat{\phi}, \widehat{\alpha})$ is a Gamma density with shape ϕ and scale determined by α. We used the same model to assign the weights for each individual in IPTW method. The fitted model model for MGPS is identical with x_{ij} replaced by $\bar{x}_{ij}$ which includes the previous dose.

$$\begin{aligned}\widehat{\pi}(d, x_{ij}, d_{i(j-1)}) &= \widehat{\psi}(x_{ij}, d_{i(j-1)}, \widehat{\gamma})\mathbb{I}\{d = 0\} \\ &\quad + (1 - \widehat{\psi}(x_{ij}, d_{i(j-1)}, \widehat{\gamma}))\mathbb{I}\{d \neq 0\} f(d|x_{ij}, d_{i(j-1)}, \widehat{\phi}, \widehat{\alpha})\end{aligned}$$

As response in the MOTAS is the vector of changes in visual acuity, there is little observed serial correlation in the data. The observable model for change in visual acuity, Y, in the GPS method is modelled via the expectation

$$\mathrm{E}_{Y|D,\pi}[Y|D = d, \pi, \beta] = \beta_0 + \mathbb{I}\{\pi < 0.05\}(\beta_1 + \beta_2 d + \beta_3 \pi + \beta_4 d.\pi)$$

and in order to decrease the bias in IPTW estimator, we have used the semi-parametric regression using weighted splines to fit the model for Y on D. A plot of the dose-response curve is presented in Figure 2. The MGPS, univariate GPS and IPTW APOs are plotted for comparison with 95% confidence interval based on MGPS.

As Figure 2 shows, there is no significant difference between the estimated APO using either IPTW or GPS method. Numerical values of the estimated parameters using least square estimates, $\beta_1, ..., \beta_4$, are presented in Table 7 for UGPS and MGPS.

The plot indicates that the direct effect of dose on visual acuity, when confounding between dose and the responses is adjusted for using the GPS ap-

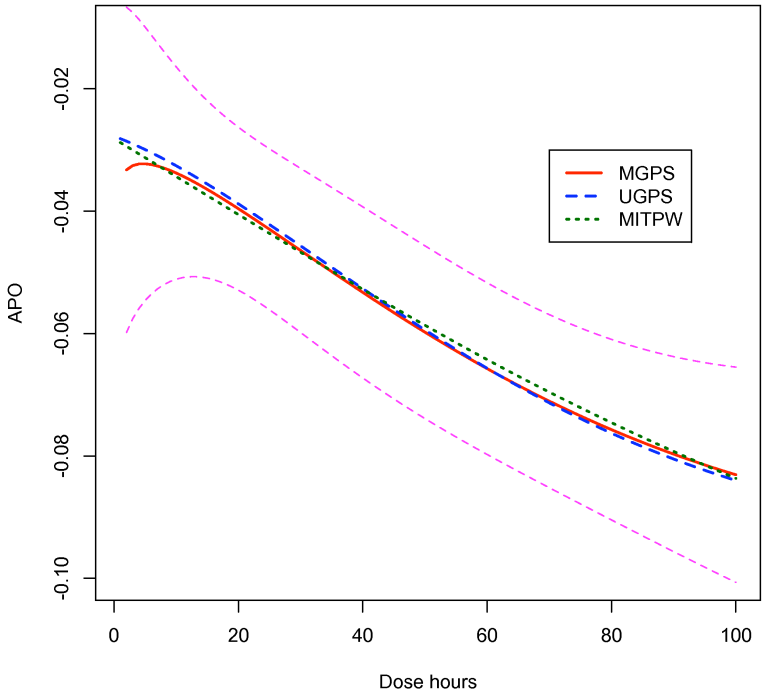


Figure 2: MOTAS data: The estimated average potential change in visual acuity (APO) vs dose for multi-interval IPTW (MIPW), UGPS and MGPS. Pointwise 95% confidence interval (light dashed) computed for MGPS.

proach, is appreciable; the average potential effect on change in visual acuity measurement Y_{ij} is significantly negative (corresponding to vision improvement) over the entire range of positive doses considered.

5 Summary

Our studies clearly demonstrate that in a range of simulation studies in single and multiple interval settings, PS methods outperform IPTW in terms of MSE. Therefore, in the context of moderate to high-dimensional covariate/confounder vectors, the scalar propensity score provides a straightforward causal adjustment approach which seems to have superior finite dimensional performance.

We outlined the Generalized Propensity Score, a generalization of the classical binary treatment propensity score, and showed that since the confounding

pattern is more complex in longitudinal data, the GPS needs to take into account the correlation between observations. We explained how the GPS can be modified to keep the balancing property in the context of repeated measures data. We compared the performance of the IPTW and GPS approach to estimate the average potential outcome through simulation studies, MSCM and MOTAS data. Our studies reveal that the ATE estimator using propensity score regression adjustment has a smaller variance and is more successful in removing bias than corresponding methods that use weighting, under correct model specification.

Our studies here are entirely empirical, but theoretical results demonstrating the superiority of PS-based estimators are also available. It can be shown using results from the theory of semiparametric estimation that, under the assumption of a correct model for treatment assignment, the PS matching or regression estimator has asymptotic variance which equals the semiparametric efficiency bound of Chamberlain (1987), whereas the variance bound for IPTW estimators - see Hahn (1998) and Hirano et al. (2003) - exceeds the Chamberlain bound. See Ertefaie and Stephens (2010) for further details. One limitation of PS methods at this stage is that they have only been developed for use in the estimation of direct effects, and cannot be used for the estimation of total effects, whereas the marginal structural models approach that utilizes IPTW does allow the estimation of total effects.

References

Alexander, G. S. and Markowitz, R. (1986) "Maternal employment and use of pediatric clinic services", *Medical Care*, 24(2), 134-147.

Chamberlain, G. (1987) "Asymptotic efficiency in estimation with conditional moment restrictions", *Journal of Econometrics*, 34, 305334.

Ertefaie, A. and Stephens, D. A. (2010) "On the Efficiency of Propensity Score Matching", Technical Report, Department of Mathematics and Statistics, McGill University.

Hahn, J. (1998) "On the role of the propensity score in efficient semiparametric estimation of average treatment effects". *Econometrica* 66 (2), 315-331.

Hirano, K. and Imbens, G. W. (2004) "The Propensity Score With Continuous Treatments". In Gelman, A. and Meng, X.-L. (eds.), *Applied Bayesian*

Modeling and Causal Inference from Incomplete-Data Perspectives", 73-84. Oxford, UK: Wiley.

Hirano, K. Imbens, G. W. and Ridder, G.(2003) "Efficient Estimation of Average treatment Effects Using the Estimated Propensity Score", *Econometrica*, 71, 4, 1161-1189.

Imbens, G. W. (2000) "The Role of the Propensity Score in Estimating Dose-Response Functions", *Biometrika*, 87:706-710.

Moodie, E. E. M. and Stephens, D. A. (2010) "Estimation of Dose-Response Functions for Longitudinal Data", *Statistical Methods in Medical Research*. To Appear.

Rosenbaum, P. R. and Rubin, D. B. (1983)"The central role of the propensity score in observational studies for causal effects", *Biometrika*, 70:41-55.

Robins, J. M. "Analytic methods for HIV treatment and cofactor effects", *AIDS Epidemiology Methodological Issues*. Eds. Ostrow DG; Kessler R. Plenum Publishing, New York. pp. 213-290.

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

Robins, J. M., Hernan, M. A. and Brumback, B. (2000). "Marginal structural models and causal inference in epidemiology", *Epidemiology*, 11, 550-560.

Robins, J. M., Mark, S. D. and Newey, W. K. (1992) "Estimating exposure effects by modeling the expectation of expisure conditional on confounders", *Biometrics* 48, 2, 479-495.

Rubin, D. B. (1978). "Bayesian inference for causal effects: The role of randomization", *Annals of Statistics* 6, 34-58.

Stewart, C. E., A. R. Fielder, D. A. Stephens, and M. J. Moseley (2002). "Design of the monitored occlusion treatment of amblyopia study (MOTAS)", *British Journal of Ophthalmology*, 86, 915-919.

Stewart, C. E., M. J. Moseley, D. A. Stephens, and A. R. Fielder (2004). "Treatment dose response in amblyopia therapy: the monitored occlusion treatment of amblyopia study (MOTAS)", Investigations in Ophthalmology and Visual Science 45, 3048-3054.

Tan, Z. (2007). "Comment: Understanding OR, PS and DR", *Statistical Science*, 22, 4, 560-568.

VanderWeele, T. J. (2009). "Marginal structural models for the estimation of direct and indirect effects", *Epidemiology* 20, 1, 18-26.