Title

A comparison of marginal odds ratio estimators

Journal

Statistical Methods in Medical Research, 26(1)

ISSN

0962-2802

Authors

Loux, Travis M
Drake, Christiana
Smith-Gagen, Julie

Publication Date

2017-02-01

DOI

10.1177/0962280214541995

Peer reviewed

A comparison of marginal odds ratio estimators

Travis M Loux, Christiana Drake and Julie Smith-Gagen

Stat Methods Med Res published online 8 July 2014
DOI: 10.1177/0962280214541995

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–21

DOI: 10.1177/0962280214541995
smm.sagepub.com
SAGE

A comparison of marginal odds ratio estimators

Travis M Loux,[1] Christiana Drake[2] and Julie Smith-Gagen[3]

Abstract

Uses of the propensity score to obtain estimates of causal effect have been investigated thoroughly under assumptions of linearity and additivity of exposure effect. When the outcome variable is binary relationships such as collapsibility, valid for the linear model, do not always hold. This article examines uses of the propensity score when both exposure and outcome are binary variables and the parameter of interest is the marginal odds ratio. We review stratification and matching by the propensity score when calculating the Mantel–Haenszel estimator and show that it is consistent for neither the marginal nor conditional odds ratio. We also investigate a marginal odds ratio estimator based on doubly robust estimators and summarize its performance relative to other recently proposed estimators under various conditions, including low exposure prevalence and model misspecification. Finally, we apply all estimators to a case study estimating the effect of Medicare plan type on the quality of care received by African-American breast cancer patients.

Keywords

causal inference, confounding, counter-factual inference, doubly robust estimator, propensity score, stratification, inverse probability of treatment weighting

1 Introduction

The odds ratio is a common measure for the association between exposure to a specific factor and presence of disease. A complication arises, however, in that the population-average, or marginal, odds ratio is not equal to the unit-specific, or conditional, odds ratio within the population. Two important consequences of this fact are that (1) one needs to be clear about the choice of marginal or conditional odds ratio as the effect of interest, and (2) distinct estimators for each parameter need to be developed. Estimating a unit-specific odds ratio entails estimating an odds ratio at each

[1]Department of Biostatistics, College for Public Health and Social Justice, Saint Louis University, Saint Louis, USA
[2]Department of Statistics, University of California – Davis, Davis, USA
[3]School of Community Health Services, University of Nevada – Reno, Reno, USA

Corresponding author:
Travis M Loux, Department of Biostatistics, College for Public Health and Social Justice, Saint Louis University, Saint Louis, MO 63104, USA.
Email: loux@slu.edu

combination of levels of covariates, often incorporating modeling assumptions. When estimating the marginal odds ratio, one needs to estimate the averages of the potential outcomes under exposure and non-exposure across all units in the population, regardless of actual exposure status. Estimating either parameter requires accounting for variables associated with exposure and/or outcome. Conditional estimation usually involves stratification or regression modeling, while common marginal estimation methods often incorporate the propensity score.

Addressing the choice of the effect of interest, someone in the role of a personal physician or therapist customizing treatment for individual patients at a time would be interested in a conditional odds ratio given the patient's risk factors. On the other hand, someone making a policy decision to be applied uniformly to a population or community would more likely be interested in the marginal odds ratio within that population. In this paper, we compare various marginal odds ratio estimators via simulation. We also apply these estimators to data on the quality of care received by African-American breast cancer patients in Medicare Managed Care Organization (MMCO) and Fee-For-Service (MFFS) plans. Since the result of such an analysis would be to advocate for a health insurance policy directed at an entire population, the parameter of interest would necessarily be marginal.

1.1 Confounding and collapsibility

Meaningful conditional and marginal estimates must be adjusted for confounding variables. We employ the comparability-based definition of confounding (e.g., Miettinen and Cook,[1] Greenland and Robins,[2] and Wickramaratne and Holford[3]), which is distinct from collapsibility. Thus, we will say X is a confounding variable when (1) X is unbalanced in sub-populations defined by exposure Z, i.e., $P(X|Z=1) \neq P(X|Z=0)$, and (2) X is a risk factor of the outcome Y after accounting for Z, i.e., $P(Y=1|X,Z) \neq P(Y=1|Z)$.

An effect measure is said to be collapsible over a covariate if the marginal and conditional measures are equal to one another.[4] Gail et al.[5] show that in the generalized linear model where $g(E[Y|X,Z]) = \beta_0 + \beta_1 X + \alpha Z$, independence between X and Z ensures collapsibility only when g is the identity or log link function. In particular, the odds ratio is not collapsible over covariates even if those covariates are independent of treatment and therefore not confounding the effect measure. Greenland et al.[6] discuss confounding and collapsibility in depth, extending the concept of collapsibility to a non-constant conditional effect by necessitating that the collapsed effect measure equal the appropriately weighted average of the conditional measures.

1.2 The potential outcomes model

To be explicit about the distinction between marginal and conditional parameters and their relationship via collapsibility, it will be useful to introduce the potential outcome approach suggested in Neyman[7] and developed by Rubin.[8,9]

We let Y_{1i} be the outcome experienced by unit i if exposed ($Z_i=1$) and Y_{0i} be the outcome experienced by unit i if not exposed ($Z_i=0$). The unit-specific effect of exposure is a comparison of Y_{0i} and Y_{1i}, while the population-average effect is a comparison of the means of Y_0 and Y_1. For a binary outcome, these means are given by $P(Y_0=1)$ and $P(Y_1=1)$, respectively.

The obvious problem arises that one cannot observe both Y_{1i} and Y_{0i} for a given i. The value which is observed depends on Z and is called the observable outcome, $Y_i = Z_i Y_{1i} + (1-Z_i)Y_{0i}$. Since only one of Y_{1i} and Y_{0i} is observed, inference in the potential outcomes model can be thought of as a missing data problem.

When exposure is independent of the distribution of potential responses, i.e., $(Y_0, Y_1) \perp\!\!\!\perp Z$ (Dawid[10]), and $0 < P(Z_i = 1) < 1$ for all i, exposure is said to be *strongly ignorable*. Exposure can also be strongly ignorable given a set of covariates, X, i.e., $0 < P(Z_i = 1|X) < 1$ and $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp Z|X$. In order for causal inferences to be made, it must be assumed that exposure is strongly ignorable given X and X are measured before exposure or otherwise not affected by exposure.

Strong ignorability requires that each unit has a non-zero probability of being exposed and a non-zero probability of being unexposed, or that any unit in one subgroup could well have been in the other. In fact, this is necessary in order for the potential outcomes Y_1 and Y_0 to be meaningful – if an unexposed unit could not possibly have been exposed, it makes no sense to describe its outcome under that impossible condition. One consequence of this philosophical restriction is that non-manipulable characteristics such as race or gender generally cannot be considered exposures or causes of an outcome.[11] However, when such comparisons are of interest, the methods described below can still be used to estimate population average adjusted probabilities. For example, instead of estimating $P(Y_1 = 1)$, it may be of interest to estimate $E_X[P(Y = 1|X = x, Z = 1)]$, where the expectation is taken with respect to the distribution of X in the full population (sample) of exposed and unexposed units. Due to the lack of potential outcomes, these estimates can no longer be considered causal, but are still useful as they allow comparisons among entire sub-populations while controlling for imbalance in relevant covariates. For a more detailed explanation and worked example, the reader is referred to Li et al.[12]

1.3 Definitions of odds ratio parameters

We define the simple comparison of the odds of the disease among the exposed relative to the odds of disease among the unexposed as the crude odds ratio

$$\psi_{crude} = \frac{P(Y_1 = 1|Z = 1)}{P(Y_1 = 0|Z = 1)} \Big/ \frac{P(Y_0 = 1|Z = 0)}{P(Y_0 = 0|Z = 0)}$$

Notice that $P(Y_1 = 1|Z = 1)$ can be estimated consistently using a random sample of exposed units because Y_1 is observable for every unit with $Z = 1$, and similarly for $P(Y_0 = 1|Z = 0)$. As the crude odds ratio compares the odds of outcome among two different subgroups of individuals, those with $Z = 1$ vs. those with $Z = 0$, without accounting for the possible confounding influences of X, the crude odds ratio has a merely associative interpretation, rather than a causal one.

The marginal, or population average, odds ratio compares the odds of disease if every unit in the population were exposed to the odds if none were exposed

$$\psi_{marg} = \frac{P(Y_1 = 1)}{P(Y_1 = 0)} \Big/ \frac{P(Y_0 = 1)}{P(Y_0 = 0)}$$

Generally, $P(Y_1 = 1)$ and $P(Y_0 = 1)$ cannot be estimated directly, so additional assumptions are needed to estimate ψ_{marg}. If exposure is strongly ignorable, as is often the case in a randomized trial, then $(Y_1, Y_0) \perp\!\!\!\perp Z$ and $\psi_{crude} = \psi_{marg}$.

The conditional, or unit-specific, odds ratio is a function of the covariate X

$$\psi_{cond}(x) = \frac{P(Y_1 = 1|X = x)}{P(Y_1 = 0|X = x)} \Big/ \frac{P(Y_0 = 1|X = x)}{P(Y_0 = 0|X = x)}$$

When X contains all of the covariates which are predictive of disease, $\psi_{cond}(x)$ measures the effect of exposure for individuals with $X = x$. Though in general the conditional odds ratio is allowed to vary with x, the commonly used logistic model yields a constant conditional odds ratio when no exposure–covariate interactions are present.

A fourth odds ratio parameter often of substantive interest is the odds ratio among the exposed. Here, the (observable) odds of outcome among the exposed are compared to the (unobservable/counter-factual) odds of outcome among the same units had they not been exposed, i.e.,

$$\psi_{exp} = \frac{P(Y_1 = 1|Z = 1)}{P(Y_1 = 0|Z = 1)} \Big/ \frac{P(Y_0 = 1|Z = 1)}{P(Y_0 = 0|Z = 1)}$$

Similarly, we could be interested in the odds ratio among the unexposed, which would condition all probabilities on $Z = 0$. Throughout this paper, we will focus on the marginal odds ratio, noting these alternative causal parameters for completeness only.

2 The Mantel–Haenszel estimator

2.1 Matching and subclassification on covariates

The Mantel–Haenszel (MH) estimator[13] is a common method to obtain a combined odds ratio from subclassified or stratified data. The notation for the kth subclass is given in Table 1, where a_k, b_k, c_k, and d_k are cell frequencies and n_k is the total number of observations in table k. The within-table odds ratio is estimated as $\hat{\psi}_k = \frac{a_k d_k}{b_k c_k}$ and the MH estimate is given by

$$\hat{\psi}_{MH} = \frac{\sum_k \frac{a_k d_k}{n_k}}{\sum_k \frac{b_k c_k}{n_k}} = \sum_k \hat{w}_k \hat{\psi}_k$$

where $\hat{w}_k = (\frac{b_k c_k}{n_k})/(\sum_l \frac{b_l c_l}{n_l})$ so that $\sum_k \hat{w}_k = 1$.

If $\psi_{cond}(x) = \psi_{cond}$ is constant across values of x and observations are subclassified so that covariate values are constant within tables, the MH estimate is consistent for the conditional odds ratio.[14] Such subclassification can arise two ways: (1) observations can be subclassified on unique combinations of discrete covariates, leading to standard large-strata asymptotics, or (2) observations can be matched on covariates, leading to sparse asymptotics.[14]

When observations are subclassified into bins defined by continuous covariates, covariate values are not constant within the subclass. Since the odds ratio is not collapsible across covariates, within-table odds ratios do not estimate ψ_{cond} and the MH estimator is not consistent for ψ_{cond}.

Table 1. A Mantel–Haenszel subclass.

	$Z = 1$	$Z = 0$	
$Y = 1$	a_k	b_k	m_{1k}
$Y = 0$	c_k	d_k	m_{0k}
	n_{1k}	n_{0k}	n_k

2.2 Matching and subclassification on the propensity score

The propensity score was defined by Rosenbaum and Rubin[15] as the probability of exposure given a set of covariates, $e(x) = P(Z = 1|X = x)$. Rosenbaum and Rubin show that exposed and unexposed subpopulations with common propensity score have balanced distributions of X and, when exposure is strongly ignorable given X, it is strongly ignorable given $e(X)$ as well. These properties allow the researcher to summarize the confounding of X, which may be high-dimensional, by a univariate score.

Two of the most common uses of the propensity score are matching and subclassification.[16] However, in a Monte Carlo study, Austin[17] shows that matching and stratifying on the propensity score lead to biased estimates of the marginal odds ratio. In another simulation paper, Austin et al.[18] show that such methods also lead to biased estimates of the conditional odds ratio.

As we show below, the bias of the propensity-matched MH estimator is related to the amount of variability of the prognostic score (as defined by Hansen[19]) within propensity-defined strata. Given logit models for exposure and outcome with logit $P(Y = 1|X = x, Z = z) = \eta(x) + \alpha z$, the prognostic score is given by $\eta(x)$ and we have $\psi_{cond}(x) = e^{\alpha}$ for all x. If there exists a function f such that $\eta(x) = f(e(x))$, in other words, if $\eta(x)$ is constant in propensity-defined strata, then matching on $e(x)$ will lead to cancellation of $\eta(x)$ and will estimate the conditional odds ratio. On the other hand, if $\eta(x)$ is independent of $e(x)$, the distribution of $\eta(x)$ within strata will equal the distribution across the population, and matching on $e(x)$ will estimate the marginal odds ratio. Proofs of the convergence of $\hat{\psi}_{MH}$ in these extreme cases are given in Appendix 1. When $\eta(x)$ and logit $e(x)$ are moderately correlated, $\hat{\psi}_{MH}$ converges to a value between ψ_{cond} and ψ_{marg}, as shown in the simulations of Section 4.8.

3 Recently proposed marginal odds ratio estimators

3.1 Stratified probability estimator

In response to Austin,[17] a number of potential estimators for the marginal odds ratio were proposed. The first comes in Graf and Schumacher[20] and is similar to the MH estimate described in Section 2.2. Using the notation from Table 1, Graf and Schumacher subclassify observations based on the propensity score and suggest estimating the probabilities of outcomes within each subclass by

$$\hat{p}_{1k} = \frac{a_k}{n_{1k}} \quad \text{and } \hat{p}_{0k} = \frac{b_k}{n_{0k}}$$

then weighting within-subclass estimates by the relative sample size within each subclass to obtain the estimates

$$\hat{p}_{j,SP} = n^{-1} \sum_k n_k \hat{p}_{jk} \quad \text{for } j = 0, 1$$

The marginal odds ratio can then be estimated by

$$\hat{\psi}_{SP} = \frac{\hat{p}_{1,SP}(1 - \hat{p}_{0,SP})}{(1 - \hat{p}_{1,SP})\hat{p}_{0,SP}}$$

As with the MH estimator, the simulations of Section 4 show the convergence of $\hat{\psi}_{SP}$ will depend on the size and number of subclasses, and a substantial bias may remain if the subclasses are

particularly large. However, as seen in Appendix 2, matching will also lead to a biased estimator as within-strata exposure rates should estimate the propensity score. This is not generally the case in a matched analysis.

3.2 Regression-based estimators

Graf and Schumacher also suggested an estimator based on a logistic (or more generally any binomial) regression model. This estimator was also examined by Zhang.[21] If a logistic model is assumed to hold so that

$$\text{logit } P(Y = 1|X, Z) = \beta_0 + \beta^T X + \alpha Z$$

then maximum likelihood estimates $\hat{\beta}_0$, $\hat{\beta}$, and $\hat{\alpha}$ can be obtained and used to estimate

$$\hat{p}_{j,ML} = n^{-1} \sum_{i=1}^{n} \frac{\exp\left\{\hat{\beta}_0 + \hat{\beta} x_i + \hat{\alpha} j\right\}}{1 + \exp\left\{\hat{\beta}_0 + \hat{\beta} x_i + \hat{\alpha} j\right\}} \quad \text{for } j = 0, 1$$

with the marginal odds ratio estimated by

$$\hat{\psi}_{ML} = \frac{\hat{p}_{1,ML}(1 - \hat{p}_{0,ML})}{(1 - \hat{p}_{1,ML})\hat{p}_{0,ML}}$$

In fact, this is how Austin[17] calculated ψ_{marg}, running a regression on the entire simulated population. Zhang states that "this can be regarded as an imputation method that replaces each potential outcome, observed or not, by its predicted value based on X," a reference to the missing data aspect of the potential outcomes model.

Another logistic regression-based estimate was proposed by Stampf et al.[22] In this estimator, the estimated propensity score, $\hat{e}_i = \hat{e}(x_i)$ is used as a covariate in the model for the outcome

$$\text{logit } P(Y = 1|Z, \hat{e}) = \beta_0 + \beta_1 \hat{e} + \alpha Z$$

As with the previous regression estimate, maximum likelihood estimates are obtained and used to estimate the marginal probabilities

$$\hat{p}_{j,COV} = n^{-1} \sum_{i=1}^{n} \frac{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \hat{e}_i + \hat{\alpha} j\right\}}{1 + \exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \hat{e}_i + \hat{\alpha} j\right\}} \quad \text{for } j = 0, 1$$

with the marginal odds ratio estimated by

$$\hat{\psi}_{COV} = \frac{\hat{p}_{1,COV}(1 - \hat{p}_{0,COV})}{(1 - \hat{p}_{1,COV})\hat{p}_{0,COV}}$$

3.3 Inverse propensity weighted estimator

Another estimator spurred by Austin[17] was given by Forbes and Shortreed.[23] Here, the estimated propensity score is used as the basis for a weighting scheme in the vein of Lunceford and Davidian[24]

to estimate the marginal probabilities of outcome under exposure and no exposure. The estimates for $P(Y_1 = 1)$ and $P(Y_0 = 1)$ are, respectively

$$\hat{p}_{1,IPW} = \left(\sum_{i=1}^{n} \frac{Z_i}{\hat{e}_i}\right)^{-1} \sum_{i=1}^{n} \frac{Z_i Y_i}{\hat{e}_i} \quad \text{and } \hat{p}_{0,IPW} = \left(\sum_{i=1}^{n} \frac{1 - Z_i}{1 - \hat{e}_i}\right)^{-1} \sum_{i=1}^{n} \frac{(1 - Z_i) Y_i}{1 - \hat{e}_i}$$

Weighting with the inverse probability of being exposed ($Z = 1$) and not being exposed ($Z = 0$) can be interpreted similar to the Horvitz-Thompson estimator in the sample survey literature.[25] In this context, the Y_i of an exposed unit represents $\hat{e}_i^{-1}$ units in the population and similarly for the Y_i of an unexposed unit when weighted by $(1 - \hat{e}_i)^{-1}$. Previous authors (e.g. Hernan and Robins[26]) have shown $\hat{p}_{1,IPW}$ and $\hat{p}_{0,IPW}$ to be consistent for $E(Y_1) = P(Y_1 = 1)$ and $E(Y_0) = P(Y_0 = 1)$, respectively. The estimate

$$\hat{\psi}_{IPW} = \frac{\hat{p}_{1,IPW}(1 - \hat{p}_{0,IPW})}{(1 - \hat{p}_{1,IPW})\hat{p}_{0,IPW}}$$

is then consistent for ψ_{marg} by continuous mapping.

3.4 Doubly robust estimator

An odds ratio estimator which has not been explicitly discussed incorporates the doubly robust (DR) estimator of Robins et al.,[27] also called the augmented inverse propensity weighted (IPW) estimator. Let $m_j(X, \alpha_j)$ be the outcome model for the subgroup with $Z = j$, with covariate vector α_j estimated by $\hat{\alpha}_j$ via maximum likelihood. Then

$$\hat{p}_{1,DR} = n^{-1} \sum_{i=1}^{n} \frac{Z_i Y_i}{\hat{e}_i} + \left(1 - \frac{Z_i}{\hat{e}_i}\right) m_1(X_i, \hat{\alpha}_1) \quad \text{and } \hat{p}_{0,DR} = n^{-1} \sum_{i=1}^{n} \frac{(1 - Z_i) Y_i}{1 - \hat{e}_i} + \left(1 - \frac{1 - Z_i}{1 - \hat{e}_i}\right) m_0(X_i, \hat{\alpha}_0)$$

are consistent estimates of $P(Y_1 = 1)$ and $P(Y_0 = 1)$, respectively, and the marginal odds ratio can be estimated by

$$\hat{\psi}_{DR} = \frac{\hat{p}_{1,DR}(1 - \hat{p}_{0,DR})}{(1 - \hat{p}_{1,DR})\hat{p}_{0,DR}}$$

The estimators $\hat{p}_{1,DR}$ and $\hat{p}_{0,DR}$ are called DR because Robins et al show that consistency is ensured so long as at least one of the propensity score model or outcome model is correctly specified. Further, when both models are correctly specified, these estimates are known to be asymptotically efficient in the class of semi-parametric estimators. Though consistent, neither $\hat{p}_{1,DR}$ nor $\hat{p}_{0,DR}$ are restricted to be between 0 and 1, potentially yielding a negative $\hat{\psi}_{DR}$, particularly in small samples or samples where $Z_i/\hat{e}_i$ or $(1 - Z_i)/(1 - \hat{e}_i)$ may be extremely large.

3.5 Handling extreme weights

Both the IPW and DR estimators incorporate the inverse of the propensity scores as weights, which can lead to exceptionally variable estimates with even a few extreme propensity scores, i.e., $e(x_i) \approx 0$ for an exposed unit or $e(x_i) \approx 1$ for an unexposed unit. In the simulations below, we estimate the marginal odds

ratio without concern for this issue, as well as using stabilized weights (Robins et al.[28]) and truncated weights (Cole and Hernan[29]). Stabilized weighting, denoted below as STAB, replaces the weight $Z_i/\hat{e}_i$ with $Z_i p/\hat{e}_i$ for the exposed subjects and $(1-Z_i)/(1-\hat{e}_i)$ with $(1-Z_i)(1-p)/(1-\hat{e}_i)$ for the unexposed subjects, where p is the proportion of the sample which is exposed. For all three estimators, weights are successively truncated at 100, 50, 20, and 10, with odds ratio estimates given for each.

4 Simulations

Except where otherwise noted, all simulations began with a population of two million units, each consisting of three covariates, X_1, X_2, and X_3, simulated from mutually independent standard normal distributions. The binary exposure, Z, was simulated from logistic models logit $P(Z=1|X=x)=\gamma^T x$, and the binary potential responses Y_1 and Y_0 were also simulated through logistic models, with logit $P(Y_0=1|X=x)=\beta^T x$ and logit $P(Y_1=1|X=x)=\beta^T x+\alpha$. For each population, the true values of ψ_{cond}, ψ_{crude}, and ψ_{marg} were calculated. With Y_1 and Y_0 explicitly generated for all units, $P(Y_1=1)$ and $P(Y_0=1)$ can be computed by simple proportions, unlike the calculation of ψ_{marg} in Austin,[17] which involved Monte Carlo integration similar to that done in Section 3.2. We drew 1000 samples of size 2000 and observed X, Z, and Y, the observable response. For each sample, the marginal odds ratio was estimated using the methods described in Section 3.

The propensity score was estimated from the sample via logistic regression using X_1, X_2, and X_3 as predictors. Subclassification on the propensity score was done using quintiles and deciles. The regression-based estimators were calculated using logistic regression for the outcome model. In estimating $\hat{\psi}_{DR}$, we used logistic regression on the exposed and unexposed units, respectively, to model $m_1(X,\hat{\alpha}_1)$ and $m_0(X,\hat{\alpha}_0)$.

Funk et al.[30] show that the model-based standard errors for the DR propensity-based estimators lead to less than nominal confidence interval coverage and suggest bootstrapping for standard errors and confidence intervals. We follow these suggestions, resampling 1000 bootstrap simulations in each sample, returning the bootstrap standard error as well as 2.5% and 97.5% bootstrap quantiles for 95% confidence intervals.

The simulations in Sections 4.1 through 4.6 below were designed to investigate the estimators' performance under three model specification scenarios: correct specification, missing a quadratic covariate, and discretizing a continuous covariate, each under moderate (50%) and low (10%) exposure prevalence. Section 4.8 investigates the bias of the MH estimator when matched on the propensity score.

4.1 Correct specifications, moderate exposure rate

For each unit exposure was generated by setting $\gamma^T=(1,1/2,1/3)$. Potential responses were simulated with $\beta^T=(2,-1,1)$ and $\alpha=\log 3$. Calculating the population parameters yielded $\psi_{crude}=4.073$, $\psi_{cond}=3.000$, and $\psi_{marg}=1.799$. Table 2 shows summary data from the 1000 samples. The columns give the empirical bias, percent bias, empirical standard error (SE) from the 1000 samples, average bootstrap standard error, the ratio of the empirical error to the boostrap error, root mean squared error, true coverage rates of 95% bootstrap confidence intervals, and, for the weighted estimators, the average number of truncated weights per sample. The numbers in parenthesis indicate the number of strata for the stratified probability (SP) estimators and the truncation weight for the IPW, STAB, and DR estimators.

By subclassifying on the propensity score, SP yields noticeable bias with five subclasses. The other estimators, with the possible exception of IPW under extreme truncation, are all essentially unbiased

Table 2. Results for Simulation 4.1. $\psi_{marg} = 1.799$.

Method	Bias	% Bias	Emp SE	Boot SE	SE ratio	rMSE	Coverage	Avg. trunc
MH (5)	0.357	19.85	0.202	0.202	0.997	0.410	0.521	
SP (5)	0.149	8.26	0.163	0.168	0.970	0.221	0.854	
MH (10)	0.270	15.02	0.192	0.194	0.987	0.331	0.678	
SP (10)	0.056	3.09	0.157	0.162	0.971	0.167	0.948	
ML	0.006	0.34	0.139	0.139	0.997	0.139	0.952	
COV	0.013	0.70	0.144	0.145	0.994	0.144	0.948	
IPW	0.011	0.61	0.160	0.162	0.987	0.160	0.950	
STAB	0.011	0.61	0.160	0.162	0.987	0.160	0.950	
IPW (100)	0.011	0.61	0.160	0.162	0.987	0.160	0.950	0.000
STAB (100)	0.011	0.61	0.160	0.162	0.987	0.160	0.950	0.000
IPW (50)	0.012	0.66	0.160	0.162	0.987	0.160	0.950	0.039
STAB (50)	0.011	0.61	0.160	0.162	0.987	0.160	0.950	0.000
IPW (20)	0.024	1.36	0.157	0.160	0.985	0.159	0.949	0.966
STAB (20)	0.013	0.71	0.159	0.161	0.988	0.160	0.951	0.096
IPW (10)	0.078	4.32	0.158	0.160	0.986	0.176	0.920	8.307
STAB (10)	0.024	1.35	0.157	0.160	0.986	0.159	0.949	0.958
DR	0.009	0.49	0.144	0.147	0.978	0.144	0.953	
DR (100)	0.009	0.49	0.144	0.147	0.978	0.144	0.953	0.000
DR (50)	0.009	0.49	0.144	0.147	0.978	0.144	0.953	0.039
DR (20)	0.008	0.45	0.144	0.146	0.983	0.144	0.953	0.966
DR (10)	0.008	0.44	0.143	0.144	0.989	0.143	0.952	8.307

MH: Mantel–Haenszel; SP: stratified probability; ML: maximum likelihood; COV: propensity covariate adjusted; IPW: inverse propensity weighted; STAB: stabilized weighting; DR: doubly robust.

for ψ_{marg}. Of these, the ML estimator is the most efficient but the differences between ML and DR are negligible. This would be expected since maximum likelihood estimates are well known be to asymptotically efficient under the correct model. The bootstrap standard errors accurately estimate the true empirical standard error throughout.

4.2 Correct specifications, low exposure rate

In the above simulation there were few extreme weights so the implications of stabilization and truncation were not well-exposed. Now, we simulate data using models similar to those in Simulation 1, with the inclusion of an intercept in the exposure model: logit $P(Z = 1|X = x) = -2.7 + x_1 + (1/2)x_2 + (1/3)x_3$. The low exposure prevalence will lead to low propensity scores, even among the exposed units, which will induce large weights. The outcome models were exactly the same as in Simulation 1. The odds ratio parameters were $\psi_{crude} = 4.587$, $\psi_{cond} = 3.037$, and $\psi_{marg} = 1.798$. Table 3 gives the resulting summary data.

Broadly speaking, Table 3 shows more bias in this situation than in the more balanced exposure situation of Section 4.1. We will focus our attention on the IPW, STAB, and DR estimators, as those are the ones affected by weight truncation. With truncation set to 20, about 1% of observations are affected, while truncation to 10 affects about 3%. In all IPW and DR estimates the coverage level is, at best, slightly below the nominal 95%. We see a bias for IPW, likely due to the large weights of some observations. Truncating the weights, though, does little to improve the estimate. The bias increases, as expected, but the reduction in standard error is not enough to compensate for the

Table 3. Results for Simulation 4.2. $\psi_{marg} = 1.798$.

Method	Bias	% Bias	Emp SE	Boot SE	SE ratio	rMSE	Coverage	Avg. trunc
MH (5)	0.570	31.68	0.454	0.473	0.959	0.728	0.657	
SP (5)	0.211	11.74	0.588	0.605	0.971	0.625	0.960	
MH (10)	0.424	23.59	0.426	0.445	0.958	0.601	0.789	
SP (10)	0.132	7.34	0.539	0.522	1.032	0.554	0.954	
ML	0.037	2.05	0.257	0.263	0.979	0.260	0.943	
COV	0.180	9.99	0.339	0.351	0.966	0.384	0.910	
IPW	0.108	6.01	0.563	0.555	1.014	0.573	0.925	
STAB	0.108	6.01	0.563	0.555	1.014	0.573	0.925	
IPW (100)	0.154	8.59	0.522	0.534	0.978	0.544	0.928	0.701
STAB (100)	0.108	6.02	0.563	0.555	1.014	0.573	0.925	0.002
IPW (50)	0.280	15.56	0.496	0.520	0.954	0.570	0.905	3.789
STAB (50)	0.110	6.10	0.561	0.554	1.012	0.571	0.925	0.007
IPW (20)	0.707	39.32	0.516	0.543	0.950	0.875	0.587	21.179
STAB (20)	0.118	6.57	0.548	0.548	0.999	0.560	0.925	0.098
IPW (10)	1.251	69.54	0.588	0.615	0.956	1.382	0.140	57.360
STAB (10)	0.153	8.52	0.523	0.534	0.978	0.544	0.929	0.695
DR	0.037	2.05	0.391	0.678	0.576	0.393	0.940	
DR (100)	0.032	1.78	0.368	0.377	0.974	0.369	0.943	0.701
DR (50)	0.029	1.62	0.352	0.362	0.972	0.353	0.947	3.789
DR (20)	0.034	1.86	0.341	0.353	0.965	0.342	0.941	21.179
DR (10)	0.030	1.68	0.340	0.351	0.969	0.341	0.947	57.360

MH: Mantel–Haenszel; SP: stratified probability; ML: maximum likelihood; COV: propensity covariate adjusted; IPW: inverse propensity weighted; STAB: stabilized weighting; DR: doubly robust.

introduced bias; in fact, the standard error increases for truncations more strict than 50. Overall, the root mean squared error increases for truncations more strict than 100, and the coverage rates never increase, holding steady only with the initial truncation and shrinking dramatically from there.

For the STAB and DR estimates, the results are better. Weight stabilization reduces the influence of extreme weights, with an average number of truncations always less than 1. The truncations lead to a negligible increase in bias and an only slightly larger decrease in standard error. This leads to a slight reduction in the rMSE, though coverage rates remained constant. Truncation does not introduce a bias to the DR estimate due to the double robustness property (the outcome model is still correct), and the standard errors reduce slightly. The coverage rates of the bootstrap confidence intervals are consistently between 94 and 95 percent. While weight truncation has little effect on the empirical standard error and rMSE of DR, even a very liberal truncation can greatly improve the overly conservative DR bootstrap standard errors dramatically, bringing it more in line with the truth without adversely affecting coverage rates.

4.3 Quadratic misspecification, moderate exposure rate

In the current simulation we included a quadratic term in the data generating process. The linear part of the exposure generating model was $-1/2 + x_1 + (1/2)x_2 + (1/3)x_3 + (1/2)x_3^2$. The potential outcomes were generated using $-1/2 + 2x_1 - x_2 + x_3 + (1/2)x_3^2$, including an extra log(3) term for the exposed outcomes. The analysis of the samples ignored the quadratic term, using the covariates only in linear terms. The results of the simulations are given in Table 4.

Table 4. Results for Simulation 4.3. $\psi_{marg} = 1.794$.

Method	Bias	% Bias	Emp SE	Boot SE	SE ratio	rMSE	Coverage	Avg. trunc
MH (5)	0.781	43.51	0.245	0.243	1.009	0.818	0.031	
SP (5)	0.515	28.68	0.200	0.202	0.993	0.552	0.157	
MH (10)	0.694	38.65	0.236	0.235	1.005	0.733	0.067	
SP (10)	0.423	23.60	0.194	0.196	0.991	0.466	0.297	
ML	0.367	20.43	0.177	0.174	1.013	0.407	0.356	
COV	0.362	20.16	0.179	0.177	1.007	0.403	0.386	
IPW	0.369	20.54	0.200	0.196	1.020	0.419	0.447	
STAB	0.369	20.54	0.200	0.196	1.020	0.419	0.447	
IPW (100)	0.369	20.55	0.200	0.196	1.020	0.419	0.447	0.004
STAB (100)	0.369	20.54	0.200	0.196	1.020	0.419	0.447	0.000
IPW (50)	0.370	20.60	0.199	0.196	1.018	0.420	0.446	0.017
STAB (50)	0.369	20.55	0.200	0.196	1.019	0.419	0.447	0.004
IPW (20)	0.378	21.06	0.197	0.194	1.013	0.426	0.412	0.571
STAB (20)	0.370	20.63	0.199	0.196	1.018	0.420	0.445	0.039
IPW (10)	0.420	23.41	0.193	0.193	0.999	0.462	0.299	6.383
STAB (10)	0.378	21.05	0.197	0.194	1.013	0.426	0.416	0.571
DR	0.398	22.16	0.188	0.188	1.001	0.440	0.334	
DR (100)	0.398	22.16	0.188	0.188	1.001	0.440	0.334	0.004
DR (50)	0.398	22.16	0.188	0.188	1.001	0.440	0.334	0.017
DR (20)	0.397	22.14	0.187	0.187	1.001	0.439	0.332	0.571
DR (10)	0.393	21.88	0.184	0.184	1.001	0.433	0.325	6.383

MH: Mantel–Haenszel; SP: stratified probability; ML: maximum likelihood; COV: propensity covariate adjusted; IPW: inverse propensity weighted; STAB: stabilized weighting; DR: doubly robust.

Biases were relatively consistent across estimators, with a slight increase in bias for the IPW estimator with strict truncation. Coverage of all confidence intervals was very low, usually between 35% and 45%. The IPW and STAB estimators had larger standard errors than the DR estimators, but that translated into increased confidence interval coverage.

4.4 Quadratic misspecification, low exposure rate

To simulate low exposure, the same models were used as in 4.3, though an intercept term of –3.3 was included in the exposure model. Again, the quadratic term was not included in the data analysis. Results are given in Table 5.

The biases are substantially larger here than in the previous simulation. Also, truncation of the IPW estimators leads to very poor confidence interval coverage, much like in Section 4.2. Under this misspecification, though, STAB seems to perform better than DR. In fact, DR without truncated weights resulted in a highly variable estimate. Even mild truncation brought the DR estimate in line with the others, but did little to improve confidence interval coverage.

4.5 Categorical misspecification, moderate exposure rate

Here we use the same data generating model as in Section 4.1; however, we categorize X_3 into tertiles before running any analyses. The results are shown in Table 6.

Table 5. Results for Simulation 4.4. $\psi_{marg} = 1.797$.

Method	Bias	% Bias	Emp SE	Boot SE	SE ratio	rMSE	Coverage	Avg. trunc
MH (5)	1.362	75.79	0.617	0.681	0.906	1.495	0.124	
SP (5)	0.936	52.06	0.649	0.738	0.878	1.138	0.594	
MH (10)	1.168	64.97	0.580	0.641	0.906	1.304	0.224	
SP (10)	0.860	47.88	0.661	0.733	0.903	1.085	0.665	
ML	0.660	36.74	0.384	0.409	0.938	0.764	0.458	
COV	0.698	38.85	0.432	0.472	0.915	0.821	0.524	
IPW	0.552	30.74	0.793	0.844	0.939	0.966	0.841	
STAB	0.552	30.74	0.793	0.844	0.939	0.966	0.841	
IPW (100)	0.596	33.16	0.657	0.726	0.905	0.887	0.823	1.514
STAB (100)	0.551	30.68	0.788	0.834	0.945	0.962	0.841	0.006
IPW (50)	0.743	41.37	0.614	0.682	0.900	0.964	0.683	5.487
STAB (50)	0.546	30.38	0.769	0.819	0.939	0.943	0.843	0.047
IPW (20)	1.221	67.93	0.620	0.685	0.905	1.369	0.228	23.906
STAB (20)	0.551	30.66	0.715	0.776	0.922	0.903	0.843	0.360
IPW (10)	1.865	103.79	0.696	0.767	0.907	1.990	0.017	59.164
STAB (10)	0.599	33.32	0.655	0.725	0.903	0.887	0.820	1.612
DR	1.849	102.89	20.404	45.188	0.452	20.478	0.586	
DR (100)	1.139	63.36	0.795	0.959	0.829	1.389	0.466	1.514
DR (50)	0.968	53.85	0.626	0.664	0.943	1.152	0.470	5.487
DR (20)	0.835	46.45	0.548	0.582	0.941	0.998	0.519	23.906
DR (10)	0.815	45.37	0.540	0.572	0.943	0.977	0.534	59.164

MH: Mantel–Haenszel; SP: stratified probability; ML: maximum likelihood; COV: propensity covariate adjusted; IPW: inverse propensity weighted; STAB: stabilized weighting; DR: doubly robust.

For the non-stratified estimators, the bias incurred due to categorizing a continuous covariate is on the order of excluding a quadratic covariate. The bias of the SP estimators is reduced to the level of the others. Bootstrap standard errors accurately reflect the empirical standard errors, but the noticeable bias leads to poor confidence interval coverage for all estimators. No estimator sticks out as clearly better or worse than the others.

4.6 Categorical misspecification, low exposure rate

Here we use the same data generating model as in Section 4.2, categorizing X_3 into tertiles before running any analyses. Table 7 gives the results of the simulations.

Again, bias is noticeable for all estimators. The reduction in SE obtained by truncating the IPW weights does not outweigh the increase in bias as coverage levels drop dramatically for truncation cutoffs more strict than 50. The STAB and DR estimates are less affected by truncation, with the STAB having a slightly larger standard error and coverage level than DR.

4.7 Summary

Both forms of misclassification introduced substantial relative bias into the estimates. Even a bias of 0.35, roughly the bias seen in Sections 4.3, 4.5, and 4.6, denotes a relative bias in effect estimate of nearly 20% for an odds ratio of 1.8 ($0.35/1.8 \approx 0.194$). In both misspecification cases, severe truncation (truncation at 20 or 10) increases the bias of the IPW estimators without reciprocal

Table 6. Results for Simulation 4.5. $\psi_{marg} = 1.8$.

Method	Bias	% Bias	Emp SE	Boot SE	SE ratio	rMSE	Coverage	Avg. trunc
MH (5)	0.583	32.41	0.224	0.225	0.995	0.625	0.137	
SP (5)	0.350	19.47	0.186	0.189	0.981	0.397	0.465	
MH (10)	0.544	30.25	0.220	0.222	0.994	0.587	0.195	
SP (10)	0.320	17.76	0.186	0.190	0.979	0.370	0.539	
ML	0.304	16.87	0.173	0.175	0.990	0.349	0.533	
COV	0.270	15.02	0.170	0.171	0.994	0.319	0.598	
IPW	0.315	17.50	0.188	0.189	0.994	0.367	0.554	
STAB	0.315	17.50	0.188	0.189	0.994	0.367	0.554	
IPW (100)	0.315	17.50	0.188	0.189	0.994	0.367	0.554	0.000
STAB (100)	0.315	17.50	0.188	0.189	0.994	0.367	0.554	0.000
IPW (50)	0.315	17.50	0.188	0.189	0.994	0.367	0.554	0.000
STAB (50)	0.315	17.50	0.188	0.189	0.994	0.367	0.554	0.000
IPW (20)	0.315	17.51	0.188	0.189	0.994	0.367	0.554	0.052
STAB (20)	0.315	17.50	0.188	0.189	0.994	0.367	0.554	0.000
IPW (10)	0.320	17.79	0.186	0.188	0.992	0.370	0.539	2.872
STAB (10)	0.315	17.51	0.188	0.189	0.994	0.367	0.554	0.047
DR	0.316	17.56	0.182	0.183	0.990	0.364	0.529	
DR (100)	0.316	17.56	0.182	0.183	0.990	0.364	0.529	0.000
DR (50)	0.316	17.56	0.182	0.183	0.990	0.364	0.529	0.000
DR (20)	0.316	17.56	0.182	0.183	0.990	0.364	0.529	0.052
DR (10)	0.317	17.60	0.181	0.183	0.990	0.365	0.526	2.872

MH: Mantel–Haenszel; SP: stratified probability; ML: maximum likelihood; COV: propensity covariate adjusted; IPW: inverse propensity weighted; STAB: stabilized weighting; DR: doubly robust.

gain in efficiency. Exposure prevalence also influenced the results, with estimates becoming much more volatile and sensitive to misspecification when exposure rates were low. In most cases the ML, STAB, and DR (appropriately truncated) estimators tended to out perform the COV and IPW estimators, though recommendations for making decisions between ML, STAB, and DR are difficult to make, as it is unclear whether one is consistently better than the others. Even though poor model specification may lead to biased estimators of the coefficients β_j used in the ML estimator (which relies on a correct specification of the outcome model), Monte Carlo integration seems to "average away" these potentially substantial biases in $\hat{\beta}_j x_{ij}$ into a reasonably reliable estimate of the marginal odds ratio. Similarly, any gains from using DR over STAB seem to be relatively minor and may not necessitate a recommendation for the more complicated estimator.

We also found the bootstrap standard errors to accurately estimate the empirical standard errors throughout and so second the suggestion of Funk et al.[30] in applying bootstrapping for standard error and confidence interval estimation.

4.8 The relationship between exposure and outcome models

In this subsection, we return our focus to the MH estimator, demonstrating that when using data matched on the propensity score the MH estimate converges to a value between ψ_{cond} and ψ_{marg}. If we assume that $\text{logit}\, e(x) = \gamma^T x$ and $\text{logit}\, P(Y = 1|X = x, Z = z) = \beta^T x + \alpha z$ are the correct models, then the bias of $\hat{\psi}_{MH}$ in estimating ψ_{marg} is strongly related to the correlation between $\gamma^T X$ and $\beta^T X$.

Table 7. Results for Simulation 4.6. $\psi_{marg} = 1.799$.

Method	Bias	% Bias	Emp SE	Boot SE	SE ratio	rMSE	Coverage	Avg. trunc
MH (5)	0.823	45.76	0.520	0.529	0.982	0.973	0.455	
SP (5)	0.427	23.76	0.598	0.641	0.932	0.735	0.914	
MH (10)	0.784	43.57	0.515	0.523	0.986	0.938	0.492	
SP (10)	0.394	21.93	0.608	0.603	1.007	0.724	0.904	
ML	0.394	21.88	0.335	0.339	0.989	0.517	0.723	
COV	0.540	30.02	0.403	0.412	0.979	0.674	0.638	
IPW	0.370	20.56	0.572	0.599	0.956	0.681	0.886	
STAB	0.370	20.56	0.572	0.599	0.956	0.681	0.886	
IPW (100)	0.378	21.00	0.560	0.586	0.955	0.675	0.885	0.420
STAB (100)	0.370	20.56	0.572	0.599	0.956	0.681	0.886	0.000
IPW (50)	0.458	25.47	0.533	0.565	0.944	0.703	0.828	3.907
STAB (50)	0.370	20.56	0.572	0.599	0.956	0.681	0.886	0.000
IPW (20)	0.919	51.10	0.565	0.588	0.960	1.079	0.397	21.667
STAB (20)	0.370	20.55	0.571	0.597	0.956	0.680	0.886	0.010
IPW (10)	1.571	87.32	0.657	0.675	0.973	1.702	0.038	54.463
STAB (10)	0.377	20.97	0.560	0.587	0.955	0.675	0.886	0.393
DR	0.388	21.55	0.498	0.533	0.934	0.631	0.868	
DR (100)	0.388	21.59	0.493	0.502	0.981	0.627	0.867	0.420
DR (50)	0.382	21.21	0.474	0.479	0.991	0.609	0.867	3.907
DR (20)	0.360	20.03	0.454	0.457	0.995	0.580	0.861	21.667
DR (10)	0.353	19.64	0.451	0.453	0.995	0.573	0.857	54.463

MH: Mantel–Haenszel; SP: stratified probability; ML: maximum likelihood; COV: propensity covariate adjusted; IPW: inverse propensity weighted; STAB: stabilized weighting; DR: doubly robust.

As discussed in Section 2.2 and Appendix 1, if $cor(\gamma^T X, \beta^T X) = 0$, then $\hat{\psi}_{MH}$ is consistent for the marginal odds ratio; however, if $cor(\gamma^T X, \beta^T X) = 1$, ψ_{MH} is consistent for the conditional odds ratio.

To detail the relationship between $cor(\gamma^T X, \beta^T X)$ and $E(\hat{\psi}_{MH})$, we ran simulations using 13 distinct $\gamma \times \beta$ combinations with $\psi_{cond} = 4$. After generating two million units, we computed the conditional and marginal odds ratios for the population. We then took 10,000 samples of size 2000 and computed the MH estimator matched on the propensity score. Figure 1 gives the bias for the empirical mean of $\hat{\psi}_{MH}$ for various correlations between $\gamma^T X$ and $\beta^T X$. Simulations using the same $\gamma \times \beta$ combinations with $\psi_{cond} = 2$ yielded similar results (not shown). From these simulations, it is clear that the stronger the correlation between $\gamma^T X$ and $\beta^T X$, the closer $\hat{\psi}_{MH}$ converges to the conditional odds ratio.

5 Comparing breast cancer quality of care

The data used in this section come from the linked Surveillance, Epidemiology and End Results Medicare database. Here, interest is in comparing the quality of breast cancer care for African-American patients in MMCO insurance plans against those in MFFS plans. For our purposes, adequate quality of care is defined as having received radiation therapy after breast-conserving surgery[31] (1 for adequate, 0 for inadequate).

We began with 2856 (30.5% MMCO) African-American women diagnosed with breast cancer between 1992 and 2004 prior to receiving breast-conserving surgery. After a variable selection procedure which excluded those variables which may have been affected by exposure (Medicare

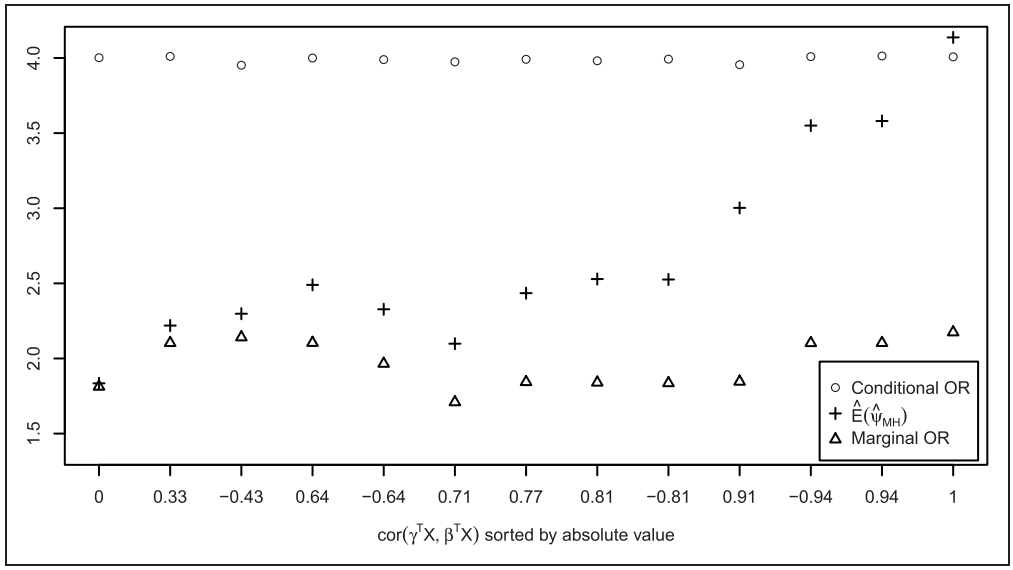

Figure 1. Relationship between $E\left(\hat{\psi}_{MH}\right)$ and $cor\left(\gamma^T X,\ \beta^T X\right)$.
Note: *x*-axis is not linear.

service) and those with no significant association to either exposure or quality of care, the relevant covariates included indicators for urban v. non-urban residence and marriage status, age at diagnosis, year of diagnosis (three categories: 1992–1996, 1997–2000, 2001–2004), and indicators for state and health service area (HSA). For simplicity, subjects with missing values, mostly with respect to radiation therapy, were dropped from the analysis. This left a total of 2720 (30.3% MMCO) women in the sample.

Table 8 gives the distribution of the discrete covariates and outcome in the remaining sample. The average ages at diagnosis were 72.8 years for the MMCO group (SD = 5.82) and 75.2 years for the MFFS group (SD = 7.17). From the last row of Table 8, the crude (unadjusted) odds ratio is estimated as 1.74.

Propensity scores were estimated via logistic regression on the covariates mentioned above, using state and HSA as random effects.[12] A review of the fitted propensity scores showed encouraging signs of comparability among the MMCO and MFFS groups – the propensity scores in the MMCO and MFFS groups had the same support (see Figure 2(a)) and there were very few extreme weights. The propensity scores ranged from 0.02 to 0.65 with median 0.26. The weights in the MMCO group ($Z_i/\hat{e}_i$) ranged from 1.55 to 22.13 with median 2.39, while the weights in the MFFS group ($(1 - Z_i)/(1 - \hat{e}_i)$) ranged from 1.021 to 2.818 with median 1.268. The total of the MMCO weights was 2682 while the total of the MFFS weights was 2715, both very close to the total sample size of 2720. Figure 2(a) shows the smoothed densities of the propensity scores – though both groups have support over the same range, the distributions are markedly different. The weighted densities given in Figure 2(b) are much more similar, and also very similar to the distribution of the propensity scores in the combined group of all patients. Table 9 gives the distributions of the discrete covariates weighted by the (unstabilized) propensity scores. For all covariates, the propensity-weighted proportions brings the distributions in the MMCO and MFFS groups closer together – in many cases the difference is almost indistinguishable.

Table 8. Proportions for discrete variables.

	MMCO	MFFS
Urban	0.925	0.830
Married	0.383	0.310
Year 1992–1996	0.306	0.293
Year 1997–2000	0.432	0.417
Year 2001–2004	0.262	0.290
Adequate care	0.708	0.581

MMCO: Medicare Managed Care Organization; MFFS: Medicare Fee-For-Service.

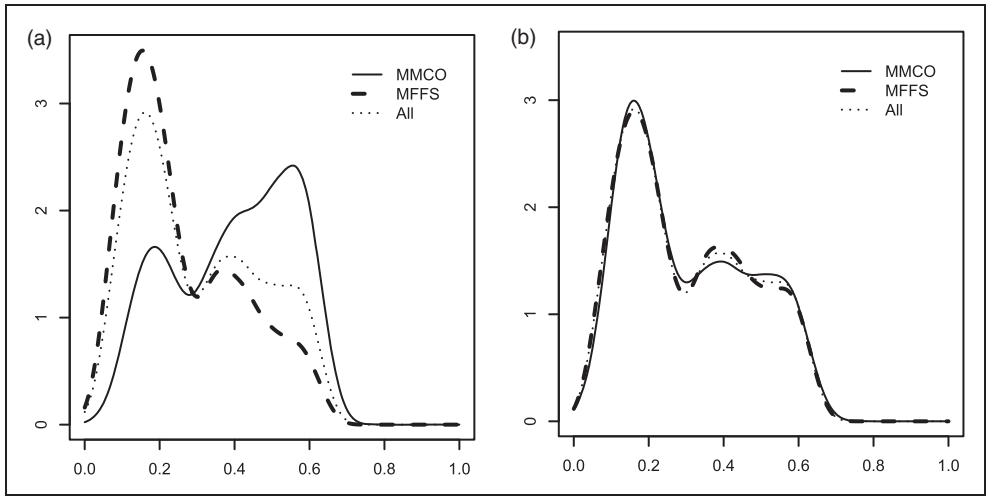

Figure 2. Density plots of fitted propensity scores.

The weighted average of the ages in the MMCO group was 74.1 (weighted SD = 6.20), compared to the MFFS group weighted average age of 74.5 (weighted SD = 6.97). Again, both the weighted means and standard deviations are closer than the unadjusted values. These results suggest weighting can effectively remove the bias introduced by the recorded variables, lending credibility to the IPW, STAB, and DR approaches of estimating the marginal odds ratio.

Using the marginal odds ratio estimators discussed in Section 3, estimates of ψ_{marg} are shown in Table 10. The standard errors come from the standard deviation of 1000 bootstrap replications of the data; similarly, the 95% confidence intervals come from the 2.5th and 97.5 percentiles of the bootstrap distributions. For the IPW, STAB, and DR estimates, the final column gives the number of weights which were truncated at the values in the parentheses. The estimates are all fairly consistent, showing the MMCO patients are more likely to receive radiation therapy after breast-conserving surgery, with odds ratios ranging from about 1.39 to 1.57 compared to MFFS patients. Similarly, the confidence intervals all span roughly 1.2 to 1.8. As mentioned previously, the weights are reasonably well-behaved. With the most restrictive weight of 7.5, only 47 (about 1.7%) of the

Table 9. Weighted proportions for discrete variables.

	MMCO	MFFS
Urban	0.852	0.858
Married	0.340	0.329
Year 1992–1996	0.301	0.296
Year 1997–2000	0.423	0.421
Year 2001–2004	0.276	0.283
Adequate care	0.696	0.600

MMCO: Medicare Managed Care Organization; MFFS: Medicare Fee-For-Service.

Table 10. Estimates of ψ_{marg}.

Estimator	$\hat{\psi}$	St. err.	95% CI	Truncated
SP (5)	1.57	0.174	(1.369, 1.786)	
SP (10)	1.51	0.231	(1.200, 1.768)	
ML	1.39	0.140	(1.270, 1.588)	
COV	1.39	0.154	(1.264, 1.601)	
IPW	1.53	0.191	(1.361, 1.796)	
STAB	1.53	0.191	(1.361, 1.796)	
IPW (20)	1.52	0.191	(1.361, 1.796)	1
STAB (20)	1.53	0.191	(1.361, 1.796)	0
IPW (15)	1.52	0.189	(1.376, 1.803)	2
STAB (15)	1.53	0.191	(1.361, 1.796)	0
IPW (10)	1.52	0.192	(1.397, 1.823)	10
STAB (10)	1.53	0.191	(1.361, 1.796)	0
IPW (7.5)	1.53	0.198	(1.397, 1.831)	47
STAB (7.5)	1.53	0.191	(1.361, 1.796)	0
DR	1.47	0.169	(1.284, 1.685)	
DR (20)	1.47	0.169	(1.284, 1.685)	1
DR (15)	1.47	0.168	(1.290, 1.689)	2
DR (10)	1.46	0.171	(1.294, 1.697)	10
DR (7.5)	1.45	0.178	(1.283, 1.695)	47

SP: stratified probability; DR: doubly robust; IPW: inverse propensity weighted.

weights were truncated. In this application, weight truncation yielded little change in the estimate or standard error.

6 Discussion

We have refined the results of Austin,[17] showing that stratifying on the propensity score leads to the MH estimator converging to a value between ψ_{marg} and ψ_{cond}, with its location relative to each dependent upon the relationship between the prognostic function of the covariates in the outcome model and the linear function of the covariates in the propensity model. If the prognostic function in

the response model is independent of the propensity score, then the estimator converges to ψ_{marg}, but if the prognostic function and propensity score are highly correlated, convergence is towards ψ_{cond}.

In Section 4.1, which could be considered an ideal situation for causal inference, all of the estimators were relatively similar. When all models were correctly specified, ML was the most efficient, though the efficiency trade-off for guarding against misspecification by using DR is minimal. When exposure was rare in the population (Section 4.2), there was little change in the relative efficiency of the estimators, and the double robustness property of DR was displayed in the ability to remain unbiased for the marginal estimate while truncating extreme weights, essentially purposely misspecifying the true propensity score model. When compared to DR, the simpler IPW estimator performed nearly as well in cases where neither exposure group was small, but was affected more by the extreme weights found in Section 4.2 when exposure had low probability across the population. The stabilization of STAB mitigates the effect of the truncation, yielding less biased estimates with better confidence interval coverage rates.

The model misspecifications in this analysis were chosen to represent near correct models, under the assumption that in practice researchers will have some notion of how covariates relate to exposure and outcome without necessarily perfect knowledge. In these cases, the ML, STAB, and DR estimators outperformed the COV and IPW estimators, though there is little evidence to suggest the use of one over the others. Previous work[32] has shown that misspecifications in the propensity score lead to smaller biases than misspecifications in outcome model. With this in mind, we may prefer STAB and DR over ML to mitigate particular misspecifications not investigated here.

The data analysis of Section 5 was qualitatively similar to the simulation in Section 4.1 in that neither exposure group was particularly small, there was good overlap in the support of the propensity scores, and the weights were reasonably well-behaved. Here again, the estimation methods yielded similar point and interval estimates for the effect of insurance model on rate of radiation therapy after breast-conserving surgery.

Odds ratios can be delicate parameters to estimate due to their non-linear interpretation as well as their non-collapsibility. All of the methods in Section 3 proved adequate at estimating the marginal odds ratio and yielded similar results under ideal circumstances. In less than ideal circumstances, some clear distinctions arise, though it is still difficult to make a general statement about the best estimator in terms of both bias and efficiency, regardless of what theory dictates. As a general rule, the results of these simulations agree with advice that often the design of an observational study is more influential to the results than the details of the analysis (e.g., Rubin[33]).

Conflict of interest

None declared.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Miettinen OS and Cook EF. Confounding: essence and detection. *Am J Epidemiol* 1981; **114**: 593–603.
2. Greenland S and Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol* 1986; **15**: 413–419.
3. Wickramaratne PJ and Holford TR. Confounding in epidemiologic studies: the adequacy of the control group as a measure of confounding. *Biometrics* 1987; **43**: 751–765.
4. Whittemore AS. Collapsibility of multidimensional contingency tables. *J R Stat Soc B* 1978; **40**: 328–340.

5. Gail MH, Wieand S and Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984; **71**: 431–444.
6. Greenland S, Robins JM and Pearl J. Confounding and collapsibility in causal inference. *Stat Sci* 1999; **14**: 29–46.
7. Neyman J. On the application of probability theory to agricultural experiments, essay on principals, Section 9. *Rocznik Nauk Roiniczych* 1923. In: Dabrowska DM and Speed TP (eds) On the application of probability theory to agricultural experiments, essay on principals, Section 9. *Rocznik Nauk Roiniczych* 1923. *Stat Sci* 1990; **5**: 465–472.
8. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974; **66**: 688–701.
9. Rubin D. Bayesian inference for causal effects: the role of randomization. *Ann Stat* 1978; **6**: 34–58.
10. Dawid AP. Conditional independence in statistical theory. *J R Stat Soc B* 1979; **41**: 1–31. [with comments].
11. Holland PW. Statistics in causal inference. *J Am Stat Assoc* 1986; **81**: 945–960. [With comments].
12. Li F, Zaslavsky AM and Landrum MB. Propensity score weighting with multilevel data. *Stat Med* 2013; **32**: 3373–3387.
13. Mantel N and Haenszel W. Statistical aspects of the analysis of data from the retrospective analysis of disease. *J Natl Cancer I* 1959; **22**: 719–748.
14. Breslow N. Odds ratio estimators when the data are sparse. *Biometrika* 1981; **68**: 73–84.
15. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
16. Shah BR, et al. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol* 2005; **58**: 550–559.
17. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med* 2007; **26**: 3078–3094.
18. Austin PC, et al. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Stat Med* 2007; **26**: 754–768.
19. Hansen BB. The prognostic analogue of the propensity score. *Biometrika* 2008; **95**: 481–488.
20. Graf E and Schumacher M. Comments on 'The performance of different propensity score methods for estimating marginal odds ratios'. *Stat Med* 2008; **27**: 3915–3917.
21. Zhang Z. Estimating a marginal causal odds ratio subject to confounding. *Commun Stat-Theor M* 2008; **38**: 309–321.
22. Stampf S, et al. Estimators and confidence intervals for the marginal odds ratio using logistic regression and propensity score stratification. *Stat Med* 2010; **29**: 760–769.
23. Forbes A and Shortreed S. Inverse probability weighted estimation of the marginal odds ratio: correspondence regarding 'The performance of different propensity score methods for estimating marginal odds ratios'. *Stat Med* 2008; **27**: 5556–5559.
24. Lunceford JK and Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med* 2004; **23**: 2937–2960.
25. Horvitz DG and Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc* 1952; **47**: 663–685.
26. Hernan MA and Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Commun H* 2006; **60**: 578–586.
27. Robins JM, Rotnitzky A and Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc* 1995; **90**: 106–121.
28. Robins JM, Hernán MA and Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**: 550–560.
29. Cole SR and Hernan MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008; **168**: 656–644.
30. Funk MJ, et al. Doubly robust estimation of causal effects. *Am J Epidemiol* 2011; **173**: 761–767.
31. Haggstrom DA, Quale C and Smtih-Bindman R. Differences in quality of breast cancer care among vulnerable populations. *Cancer* 2005; **104**: 2347–2358.
32. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 1993; **49**: 1231–1236.
33. Rubin D. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv Out Res Methodol* 2001; **2**: 169–188.
34. Breslow NE and Day NE. In: Davis W (ed.) *Statistical methods in cancer research Vol 1 – the analysis of case-control studies*. No. 32 in IARC Scientific Publications. Lyon: International Agency for Research on Cancer, 1980.
35. Robins J, Breslow N and Greenland S. Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* 1986; **42**: 311–323.
36. Armitage P. The use of the cross-ratio in aetiological surveys. In: Gani J (ed.) *Perspectives in probability and statistics: papers in honour of M.S. Bartlett*. London: Applied Probability Trust, 1975, pp.349–355.

Appendix 1. Matching on the propensity score leads to an inconsistent estimate of ψ_{cond} and ψ_{marg}

Following Breslow and Day,[34] let

$$\begin{aligned} p_1(x) &= P(Y = 1|Z = 1, X = x) \\ &= P(Y_1 = 1|X = x) \\ p_0(x) &= P(Y = 1|Z = 0, X = x) \\ &= P(Y_0 = 1|X = x) \end{aligned}$$

with $q_j(x) = 1 - p_j(x)$ for $j = 0, 1$. Assuming the conditional odds ratio is constant for simplicity, we can write $\psi_{cond} = \frac{p_1(x)q_0(x)}{q_1(x)p_0(x)}$ and $\psi_{marg} = \frac{E(p_1(X))E(q_0(X))}{E(q_1(X))E(p_0(X))}$. Assume units are matched on the propensity score, $e(X)$, allowing the covariates to vary within strata defined by the propensity score. Define X_{1k} and X_{0k} as the covariates for the exposed and unexposed units in the kth discordant pair, respectively, $k = 1, \ldots, n$, and similarly for the outcomes Y_{1k} and Y_{0k}. Let $R_k = Y_{1k}(1 - Y_{0k})$ and $S_k = (1 - Y_{1k})Y_{0k}$.

Case 1:

$$p_0(x) = f(e(x)).$$

To show $\hat{\psi}_{MH}$ is unbiased for ψ_{cond}, it is enough to show that $E[\sum_k (R_k - \psi_{cond}S_k)|\{e(X_{1k}) = e(X_{0k})\}_{k=1}^n] = 0$ (per Robins et al.[35])

$$\begin{aligned}
&E\left[\sum_k (R_k - \psi_{cond}S_k)|\{e(X_{1k}) = e(X_{0k})\}_{k=1}^n\right] \\
&= \sum_k E[R_k|e(X_{1k}) = e(X_{0k})] - \psi_{cond}E[S_k|e(X_{1k}) = e(X_{0k})] \\
&= \sum_k p_1(X_{1k})q_0(X_{0k}) - \psi_{cond}q_1(X_{1k})\,p_0(X_{0k}) \\
&= \sum_k p_1(X_{1k})q_0(X_{1k}) - \psi_{cond}q_1(X_{1k})\,p_0(X_{1k}) \\
&= 0
\end{aligned}$$

Thus, when $p_0(x) = f(e(x))$, $\hat{\psi}_{MH}$ is consistent for the conditional odds ratio.

Remark:

When ψ_{cond} is constant, $p_1(x) = \frac{p_0(x)\psi_{cond}}{1-p_0(x)+p_0(x)\psi_{cond}}$, so $p_1(x)$ is also a function of $e(x)$.

Case 2:

$$p_1(X_{1k}) \underline{\parallel} p_0(X_{0k})|e(X_{1k}) = e(X_{0k}).$$

Similar to above, we need to show $E\left[E\left[\sum_k (R_k - \psi_{marg}S_k)|\{e(X_{1k}) = e(X_{0k})\}_{k=1}^n\right]\right] = 0$

$$\begin{aligned}
&E\left[E\left[\sum_k (R_k - \psi_{marg}S_k)|\{e(X_{1k}) = e(X_{0k})\}_{k=1}^n\right]\right] \\
&= \sum_k E[E[R_k|e(X_{1k}) = e(X_{0k})]] - \psi_{marg}E[E[S_k|e(X_{1k}) = e(X_{0k})]] \\
&= \sum_k E[p_1(X_{1k})q_0(X_{0k})] - \psi_{marg}E[q_1(X_{1k})\,p_0(X_{0k})] \\
&= \sum_k E[p_1(X_{1k})]E[q_0(X_{0k})] - \psi_{marg}E[q_1(X_{1k})]E[p_0(X_{0k})] \\
&= 0
\end{aligned}$$

So, $\hat{\psi}_{MH}$ is consistent for the marginal odds ratio.

Remark:

Since $p_1(X)$ is a function of $p_0(X)$ for a constant conditional odds ratio (see above), this case occurs when $(p_1(X), p_0(X)) \perp\!\!\!\perp e(X)$

Since the convergence of $\hat{\psi}_{MH}$ depends on the relationship between the propensity score and outcome models, in general $\hat{\psi}_{MH}$ is consistent for neither the conditional nor marginal odds ratio.

As the marginal odds ratio is closer to unity than the conditional odds ratio (Armitage[36]), matching on the propensity score will tend to underestimate the conditional effect and overestimate the marginal effect.

Appendix 2. SP estimator using matched data

In Section 3.1, the SP estimator for $P(Y_1 = 1)$ is written as

$$\hat{p}_{1,SP} = \sum_{k=1}^{K} \frac{n_k}{n} \frac{a_k}{n_{1k}}$$

notice we can write $a_k = \sum_{i \in I_k} Z_i Y_i$, where I_k indexes the observations in strata k. Thus, $\hat{p}_{1,SP}$ becomes

$$\begin{aligned} \hat{p}_{1,SP} &= \sum_{k=1}^{K} \sum_{i \in I_k} \frac{n_k}{n} \frac{Z_i Y_i}{n_{1k}} \\ &= n^{-1} \sum_{k=1}^{K} \sum_{i \in I_k} \frac{Z_i Y_i}{n_{1k}/n_k} \end{aligned}$$

Compare this to

$$n^{-1} \sum_{i=1}^{n} \frac{Z_i Y_i}{\hat{e}_i}$$

which is known to be unbiased for $P(Y_1 = 1)$. If observations are stratified based solely on $\hat{e}(x)$, then n_{1k}/n_k gives an estimate of $P(Z = 1)$ within the stratum of $\hat{e}(x)$. In the matching case, however, n_{1k}/n_k is artificially fixed (and often constant) in all strata. For example, in $1:1$ matching $n_{1k}/n_k = 1/2$ for all k as we get

$$E(\hat{p}_{1,SP}) = 2n^{-1} \sum_{k=1}^{K} \sum_{i \in I_k} E(Z_i Y_i)$$

where $E(Z_1 Y_i) = E[E(Z_i Y_i | Y_{1i}, x_i)] = \hat{e}_i E(Y_1) = \hat{e}_i P(Y_1 = 1)$, so

$$E(\hat{p}_{1,SP}) = P(Y_1 = 1) \cdot 2n^{-1} \sum_{k=1}^{K} \sum_{i \in I_k} \hat{e}_i$$

Thus, in this case n_{1k}/n_k does not give an appropriate propensity score-based weight, leading to a biased estimate of $P(Y_1 = 1)$. A similar argument shows $\hat{p}_{0,SP}$ is biased for $P(Y_0 = 1)$.