Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
21(3) 273–293

DOI: 10.1177/0962280210394483
smm.sagepub.com
SAGE

Propensity scores: From naïve enthusiasm to intuitive understanding

Elizabeth Williamson,[1,2,3] Ruth Morley,[1,2] Alan Lucas[4] and James Carpenter[5]

Abstract

Estimation of the effect of a binary exposure on an outcome in the presence of confounding is often carried out *via* outcome regression modelling. An alternative approach is to use propensity score methodology. The propensity score is the conditional probability of receiving the exposure given the observed covariates and can be used, under the assumption of no unmeasured confounders, to estimate the causal effect of the exposure. In this article, we provide a non-technical and intuitive discussion of propensity score methodology, motivating the use of the propensity score approach by analogy with randomised studies, and describe the four main ways in which this methodology can be implemented. We carefully describe the population parameters being estimated — an issue that is frequently overlooked in the medical literature. We illustrate these four methods using data from a study investigating the association between maternal choice to provide breast milk and the infant's subsequent neurodevelopment. We outline useful extensions of propensity score methodology and discuss directions for future research. Propensity score methods remain controversial and there is no consensus as to when, if ever, they should be used in place of traditional outcome regression models. We therefore end with a discussion of the relative advantages and disadvantages of each.

Keywords

confounding, inverse probability weighting, matching, observational study, propensity score, stratification

1 Introduction

Non-experimental studies are often conducted to investigate the effect of a binary exposure on an outcome. In the absence of random allocation of exposure, however, systematic differences between

[1]Murdoch Childrens Research Institute, Melbourne, Australia
[2]MEGA Epidemiology, School of Population Health, University of Melbourne, Melbourne, Australia
[3]Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Australia
[4]Childhood Nutrition Research Centre, Institute of Child Health, London, UK
[5]Medical Statistics Unit, London School of Hygiene & Tropical Medicine, London, UK

Corresponding author:
Elizabeth Williamson, Department of Epidemiology and Preventive Medicine, The Alfred Centre, 99 Commercial Road, Melbourne, VIC 3004, Australia
Email: elizabeth.williamson@monash.edu

exposed and unexposed subjects create the problem of confounding. In 1983, Rosenbaum and Rubin[1] published a seminal article introducing the propensity score in which they demonstrated that, under certain assumptions, this score can be used to unbiasedly estimate the (causal) effect of an exposure in the presence of confounding. The gradual acceptance and use of the propensity score approach has culminated in what has recently been described as 'overwhelming and sometimes naïve enthusiasm'.[2]

Although several articles detail the methodology of the propensity score approach (e.g.[3,4]) or provide example analyses,[5,6] few provide the reader with an intuitive understanding of the general approach and the links between the various ways in which it can be implemented. In this article, we provide a non-technical and intuitive discussion of these issues, paying particular attention to the definition of the estimand – the population parameter being estimated. This latter issue has received great attention in the econometrics literature in regard to propensity score methods, but has been comparatively neglected in the medical literature.[7] We comment on the current usage of the propensity score, outline useful extensions of propensity score methodology and discuss directions for future research.

In Section 2, we discuss three definitions of the causal effect of an exposure and in Section 3, we describe how the propensity score can be used to estimate these causal effects. In Section 4, we discuss estimation of the propensity score itself and in Section 5, we summarise important extensions of this approach. In Section 6, we demonstrate the propensity score methods using data from a study investigating the association between maternal choice to provide breast milk and the infant's subsequent neurodevelopment. Despite widespread use, propensity score methods remain controversial, due to frequent misuse and over-confidence in the ability of these methods to remove bias.[2,8] We therefore conclude, in Section 7, with a discussion of some of the advantages and disadvantages of the propensity score approach in comparison with traditional outcome regression models.

2 Causal effects

We begin by considering a simple, but common, scenario in which a cohort study has been undertaken in order to investigate the effect of a (non-randomised) binary exposure Z (1 = exposed, 0 = unexposed) on a continuous outcome Y, with information additionally collected on potential confounding variables $X = (X_1, \ldots, X_p)$.

2.1 A motivating example

Several observational studies have suggested that provision of maternal breast milk improves the infant's subsequent neurodevelopment.[9] Randomisation of breastfeeding is not feasible but confounding by parental education and socio-economic status (SES) is highly likely in the absence of randomisation. In Section 6, we investigate this question using a real dataset. For the purposes of illustration, we first introduce a simplified version of the dataset.

Figure 1 shows a dataset containing 20 children. The binary exposure is breastfeeding and the outcome is IQ at 7.5 years of age. The scale on which IQ is assumed to be measured is based on the abbreviated, Weschler Intelligence Scale for Children (revised Anglicised version: WISC-R UK).[10] We assume, at present, that the only confounder of the association between breastfeeding and IQ is SES status (of the mother) which we have dichotomised into low and high.

To allow calculation of the population (true) values of the causal estimands, suppose that breastfeeding causally increases IQ at age 7.5 by 5 points in children from low SES families, but

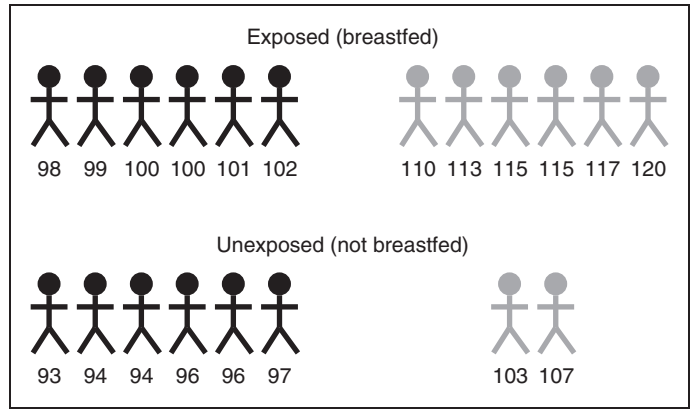

Figure 1. A hypothetical observational dataset ($n = 20$), where each figure represents a child from a low SES (black) or high SES (grey) household, with outcomes (IQ at 7.5 years) written below the figures.

by 10 points in children from high SES families. Suppose further that 40% of women in the population are classified as high SES. And finally, suppose that 50% of low SES women choose to breastfeed their children, compared with 75% of high SES women, resulting in a total of 60% of women in the population breastfeeding their children.

2.2 The causal effect of the exposure

In order to consider issues of cause and effect, we turn to the potential outcomes (or counterfactual) framework.[11] Each individual has two *potential* outcomes: the outcome that would occur if that individual was exposed, Y_1, and the outcome that would occur if that individual was not exposed, Y_0. We observe $Y = ZY_1 + (1 - Z)Y_0$. The intuitively defined causal effect for an individual, $Y_1 - Y_0$, can never be observed since only one of Y_1 and Y_0 occurs (the other is *counterfactual*). When interest lies in a group of individuals, we might define the causal effect as the mean of the individual causal effects. This mean can be estimated by using outcomes of other individuals to estimate the unobserved counterfactual outcomes. Since the individual causal effects may not all be equal, defining the group of individuals for whom we wish to estimate the mean causal effect of the exposure becomes important. Setting this issue within the standard framework of an infinite population from which our sample was randomly drawn, there are two popular choices for the group of interest: the whole population and the sub-population of those who are actually exposed. We call the resulting causal effects, respectively, the Average Causal Effect of the exposure (ACE_{All}) and the Average Causal Effect of the exposure on the Exposed (ACE_{Exp}). These are sometimes called the Average Treatment Effect (ATE) and Average Treatment effect on the Treated (ATT).[12] A third estimand that might be of interest, but is rarely discussed, concerns the sub-population of those who are actually unexposed, (ACE_{Un}). Other causal effects could be constructed but we will restrict our attention to the three mentioned.

The ACE_{All} is the mean of the individual causal effects in the whole population, $E[Y_1 - Y_0]$, or, equivalently, the difference between two hypothetical mean outcomes in the population: the mean outcome if everyone were exposed $E[Y_1]$ and the mean outcome if no-one were exposed $E[Y_0]$. The ACE_{Exp} restricts attention to the sub-group of the population who are in fact exposed, and is the mean of the individual causal effects in this sub-population, $E[Y_1 - Y_0|Z = 1]$ or, equivalently,

the difference between the mean outcome of all exposed people in the population $E[Y_1|Z=1]$ and the mean outcome of these *same* people had they not been exposed $E[Y_0|Z=1]$. Analogously, the ACE_{Un} restricts attention to the sub-group of the population who are in fact not exposed, $E[Y_1 - Y_0|Z=0]$.

For the simplified breastfeeding example (Section 2.1), the causal effect of breastfeeding on IQ at age 7.5 is 5 units for each low SES child and 10 for each high SES child. Since the population comprises 60% low SES children, the mean of the individual causal effects in this population is

$$ACE_{All} = \frac{60}{100} \times 5 + \frac{40}{100} \times 10 = 7$$

In the sub-population of the children who were breastfed, 50% are low SES, so the mean causal effect in this sub-population is

$$ACE_{Exp} = \frac{50}{100} \times 5 + \frac{50}{100} \times 10 = 7.5$$

In the sub-population of the children who were not breastfed, 75% are low SES, so the mean causal effect in this sub-population is

$$ACE_{Un} = \frac{75}{100} \times 5 + \frac{25}{100} \times 10 = 6.25$$

The effect of breastfeeding is larger – more beneficial – amongst those who are more likely to be breastfed (high SES children). Therefore, the ACE_{Exp} is larger than the ACE_{All} which in turn is larger than the ACE_{Un}.

Whether the ACE_{All}, the ACE_{Exp} or the ACE_{Un} should be estimated depends on the study question. For example, many primary (elementary) schools provide children with access to school computers to develop their computer literacy skills. This practice is more likely to occur in schools located in wealthier areas but children living in such areas are more likely to have home computers. Thus, the effect of school computers may be smallest for children currently receiving them. The three causal effects will therefore differ. A policy-maker assessing the cost-effectiveness of a proposed population-wide provision of free computers to schools may be interested in the ACE_{All}. A policy-maker specifically charged with increasing computer literacy amongst deprived children would be more interested in the ACE_{Un}. To address the question of whether the computers are currently having any effect, however, the ACE_{Exp} would be the causal effect of interest.

2.3 Assumptions

We make the following assumptions: The first, most vital, assumption is that the exposure Z must temporally precede its putative effect Y. The second assumption – sometimes called positivity[13] – is that, since the causal effect of the exposure for someone who could never (or would always) be exposed is undefined, each subject must have the potential to be exposed, and to be unexposed. For the breastfeeding example, this might be violated by babies born to HIV +ve mothers, since the latter are often advised not to breastfeed. The third assumption is often called the Stable Unit Treatment Value Assumption (SUTVA).[14] It states that the two potential outcomes for an individual are unaffected by any other individual's exposure status. This assumption is required for most individually randomised trials and observational studies. An example violation is given by

Little and Rubin: Suppose you and I are in the same room and we both have a headache. If you do not take aspirin my headache will remain whether or not I take aspirin since your whining will counteract the alleviating effect of my aspirin![15] We further assume that the data $\{Y_0, Y_1, Z, X\}$ were sampled independently from the population, in which case SUTVA follows automatically. The final assumption is often called the SITA assumption – Strongly Ignorable Treatment Assignment (given the observed covariates).[1] It states that the exposure and the potential outcomes are conditionally independent, given the observed covariates. In other words, there are no unobserved confounders. When exposure is not randomly allocated, it is impossible to be certain that this assumption has been satisfied. In the simplified breastfeeding example, the SITA assumption may be violated because the unmeasured gestational age of the child is likely to confound the association between breastfeeding and IQ. The assumed structure of the data and the SITA assumption are shown schematically in Figure 2.

3 Using the propensity score to estimate causal effects

3.1 Covariate balance in a randomised trial

For each participant of a simple two-arm randomised controlled trial (RCT), the probability of being exposed is 50%. As this probability is the same for each participant, the group who actually receives the exposure can be viewed as a random sample from the whole set of trial participants, leading us to expect that the exposed and unexposed arms will be comparable in terms of all characteristics other than exposure. Any pair of participants, one exposed and the other unexposed, will have different characteristics (one may be male and the other female, for example) but all characteristics will be balanced between the two trial arms *on average*, those both observed and unobserved. Provided that the SUTVA assumption holds, this balance ensures that the mean counterfactual outcome of the participants in one arm of the trial can be estimated by the mean outcome in the other arm, i.e. the difference between the mean outcomes in the exposed and unexposed arms is an unbiased estimate of $E[Y_1] - E[Y_0]$. Note that the randomisation artificially forces the ACE_{Exp} and ACE_{Un} to coincide with the ACE_{All}.

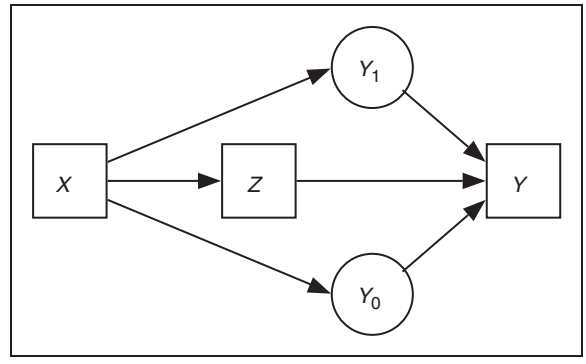

Figure 2. Structure of the data. Observed variables are depicted as squares and unobserved (potential outcomes) as circles. The SITA assumption is satisfied by the absence of arrows directly connecting exposure Z to the potential outcomes Y_0 and Y_1.

3.2 Balancing covariates using the propensity score

The propensity score is defined as the conditional probability of being exposed given the observed covariates, $e(X) = \mathbb{P}(Z = 1 \mid X)$. Each subject is assumed to have the possibility of being either exposed or unexposed (Section 2.3) so $0 < e(X) < 1$. The propensity score is typically unknown but can be estimated from the data. We defer discussion of its estimation until Section 4.

In an observational study, the covariates are unlikely to be balanced between the exposed and unexposed groups. Rosenbaum and Rubin,[1] however, demonstrated that observed covariates are balanced at each value of the propensity score. This is because, at any particular value of the propensity score, each subject with that propensity score has the same probability of being exposed (conditional on the observed covariates). Therefore, if the SITA assumption holds, we can essentially view those who are in fact exposed as a random sample of all subjects with that propensity score. Analogously to the RCT, any pair of subjects with the same propensity score, one exposed and the other unexposed, will have different characteristics. On average, however, all (observed) characteristics will be balanced between the two exposure groups. Provided that the SUTVA assumption holds, in addition to SITA, this balance ensures that *in the sub-sample of subjects with a particular value of the propensity score*, p, the mean counterfactual outcome of the subjects in one exposure group can be estimated by the mean observed outcome in the other exposure group, i.e. the difference between the mean outcomes in the exposed and unexposed groups is an unbiased estimate of $E[Y_1 - Y_0 \mid e(X) = p]$.

Thus, under the strong assumption that all confounders have been observed, the propensity score can be used to re-create a pseudo-randomised situation (at each value of the propensity score), allowing unbiased estimation of the exposure effect within strata of the propensity score.

3.3 Four propensity score methods

We describe four ways in which this general approach can be applied, once the propensity score has been estimated (Section 4), using the simplified breastfeeding example. Since there is a single binary confounder, the propensity score can be estimated by the proportion of children who were breastfed in the low and high SES groups, i.e. $e = \frac{1}{2}$ for low SES children, and $e = \frac{3}{4}$ for high SES children.

3.3.1 Stratification on the propensity score

Following the argument above, we might: (i) create strata within which all subjects have the same value of the estimated propensity score, (ii) within each stratum estimate the exposure effect by the difference in mean outcome between exposed and unexposed groups and (iii) calculate a weighted average of the within-strata estimates of the effect of the exposure, with the weight for a stratum equal to the fraction of the sample within that stratum. This process estimates the ACE_{All}. By changing the weights in step (iii), other causal effects can be estimated. The ACE_{Exp} is estimated by weighting strata by the fraction of the exposed subjects who are in each stratum, and the ACE_{Un} is estimated by weighting strata by the fraction of the unexposed subjects who are in each stratum.

The strata for the breastfeeding example are shown in Figure 3(a). The within-strata mean differences in outcome are 5 for stratum 1, and 10 for stratum 2. Estimates of the ACE_{All}, ACE_{Exp} and ACE_{Un} of breastfeeding on IQ at age 7.5, respectively, are

$$\widehat{ACE}_{All} = \frac{12}{20} \times 5 + \frac{8}{20} \times 10 = 7$$
$$\widehat{ACE}_{Exp} = \frac{6}{12} \times 5 + \frac{6}{12} \times 10 = 7.5$$
$$\widehat{ACE}_{Un} = \frac{6}{8} \times 5 + \frac{2}{8} \times 10 = 6.25$$

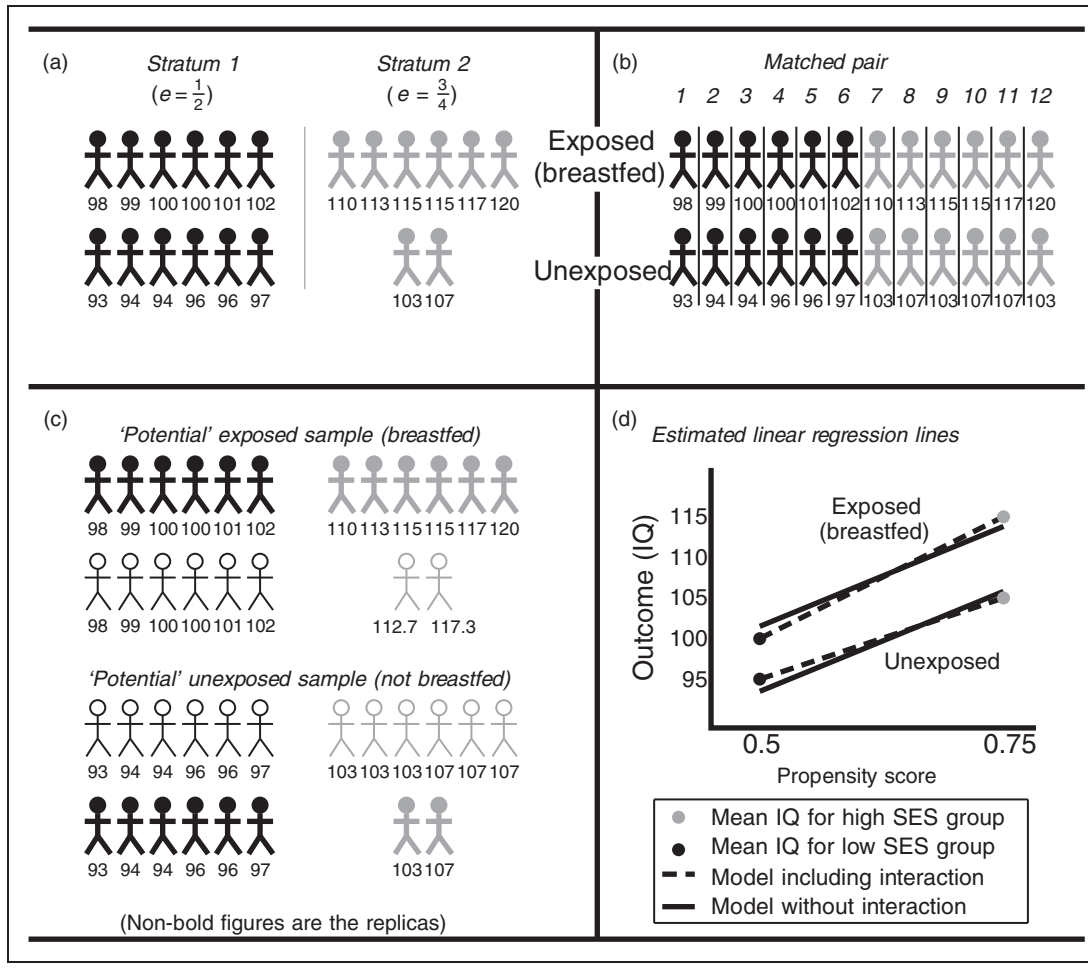

Figure 3. Pictorial representation of how four propensity score methods could be used to analyse the data shown in Figure 1. Black figures represent children from a low SES household and grey figures from a high SES household. The outcomes (IQ at 7.5 years) are written below the figures.

Ideally, each stratum should contain a single value of the propensity score but this is often not possible. Various ways of choosing strata boundaries can be employed,[16] the most common being to divide the sample at percentiles of the estimated propensity score creating equal-sized groups. Allowing strata to contain several values of the propensity score will cause bias (residual confounding[17]), but over 90% of bias due to imbalance of observed covariates will often be removed by the stratification.[18]

3.3.2 Matching on the propensity score

A popular alternative is to: (i) for each exposed subject, select a single unexposed subject who has the same value of the estimated propensity score, (ii) estimate the within-pair effect of the exposure by taking the difference in the two outcomes and (iii) calculate the average within-pair estimate of the effect of the exposure. This estimates the ACE_{Exp}. To estimate the ACE_{Un}, each unexposed subject must be allocated an exposed match. For the ACE_{All}, a match must be found for each subject in the sample. This means that, inevitably, some subjects will appear more than once in the matched sample.

Matched pairs for the breastfeeding example, selected to estimate the ACE_{Exp}, are shown in Figure 3(b). Since, unusually, more subjects are exposed than unexposed, matching must be with replacement to ensure a match for each exposed subject. For this particular set of matches selected, the effect of breastfeeding on IQ at age 7.5 is estimated as

$$\widehat{ACE}_{Exp} = \frac{1}{12}\,(5+5+6+4+5+5+7+6+12+8+10+17) = 7.5$$

By selecting different matches, estimates ranging from 6.8 to 8.2 can be obtained. The range of estimates resulting from selecting different matches will decrease as the sample size increases. Selecting a single exposed match for each unexposed child produces a mean estimate of $\widehat{ACE}_{Un} = 6.25$ ranging from 5.4 to 7.1, and selecting a match for each child in the sample produces a mean estimate of $\widehat{ACE}_{All} = 7$ ranging from 6.3 to 7.8.

Inexact matching leads to bias in the estimate of the causal effect. Various ways of selecting optimal matches have been proposed to minimise this bias.[3,8,19] A second problem is caused by discarding an often substantial number of subjects, leading to a loss in statistical power. This can also be avoided by more complex matching strategies (e.g. allowing varying numbers of subjects within a matched group). A third, more subtle, problem is that the population quantity being estimated may change when subjects are discarded due to the lack of a suitable match.[20] Finally, there is some debate about how (and whether) the within-pair correlation caused by matching should be accounted for in the analysis.[8,20]

3.3.3 Weighting by the inverse of the propensity score

The idea of this approach, when used to estimate the ACE_{All}, is to create two 'potential' samples (sometimes called a pseudo-population[13]) that are intended to represent the samples we would have observed if: (i) everyone had been exposed and (ii) no-one had been exposed. We demonstrate the construction of these for the simplified breastfeeding example. Since each low SES child has a propensity score of $\frac{1}{2}$, i.e. a 1-in-2 chance of being breastfed, we expect that for each low SES child who is breastfed there is another low SES child who is not. Although we cannot know the (counterfactual) outcome that would have occurred if this unexposed child had been breastfed, we can estimate it by the outcome of the child who *was* breastfed. To do this, we transform each breastfed low SES child into two children, assigning the 'replica' the outcome of the original child.

High SES children have a propensity score of $\frac{3}{4}$, i.e. a 3-in-4 change of being breastfed. Therefore, for each group of three breastfed high SES children, we expect there to be a single high SES child who was not breastfed. Thus, for each three breastfed high SES children, we create a single replica, assigning the replica the mean observed outcome of the three original children. Similarly, for each low SES child who was not breastfed, we create a single replica, and for each high SES child who was not breastfed we create three replicas (a 1-in-4 chance of not being breastfed means that for each child not breastfed, we expect there to be three who were). This process leads to the situation shown in Figure 3(c). The estimated ACE_{All} is simply the difference in mean outcome between the two potential samples. This estimates the effect of breastfeeding on IQ at age 7.5 as

$$\widehat{ACE}_{All} = \frac{(12 \times 100) + (8 \times 115)}{20} - \frac{(12 \times 95) + (8 \times 105)}{20} = 7$$

Equivalently, a weighted linear regression model of outcome on exposure can be fitted, where the weight for each exposed subject is $1/e(X)$ and the weight for each unexposed subject is $1/(1-e(X))$.

From Figure 3(c), it is clear that the two potential samples created by the weighting each mirror exactly the covariate distribution of the original dataset (in this case, the distribution of SES). To estimate the ACE_{Exp}, we need weights that will create two potential samples each of which exactly mirror the covariate distribution of the exposed subjects in the original dataset (i.e. 40% high SES). Weighting exposed subjects by 1 means that the potential exposed sample is identical to the exposed subset of the sample, and weighting the unexposed subjects by $e(X)/(1-e(X))$ produces a potential unexposed sample that takes the same covariate distribution as the exposed subset of the original dataset.[6] Similarly, to estimate the ACE_{Un}, the unexposed subjects are weighted by 1 and the exposed by $(1-e(X))/e(X)$.

If the sample contains an exposed (or unexposed) individual with a very low (or high) propensity score, that individual would be replicated many times in the process above so that random error in that individual's outcome would greatly influence the estimate of exposure effect. This leads to estimates of exposure effect with very high variance and thus wide confidence intervals.

3.3.4 *Including the propensity score as a covariate (covariate-adjustment)*

The simplest way of applying this popular method would be to fit a regression model of the outcome on both the exposure and the propensity score

$$E[\,Y \mid Z, X\,] = \alpha + \beta\, e(X) + \gamma\, Z.$$

The estimate of regression coefficient γ is the estimate of the exposure effect. This procedure (Figure 3(d), solid lines) estimates the effect of breastfeeding on IQ at age 7.5 as 6.67 which is equal to none of the causal effects discussed. Allowing the effect of the exposure to vary with the propensity score[1] fits the correct model in this example (Figure 3(d), dashed lines). Then, the ACE_{All}, ACE_{Exp} and ACE_{Un} can be correctly estimated by evaluating the exposure effect at three propensity score values: the sample mean, the sample mean in the exposed and the sample mean in the unexposed, respectively. Rather than assuming a linear relationship between propensity score and outcome, other functions of the propensity score such as the logit can be used as covariates in the model, or more flexible strategies including splines can be employed.

3.3.5 *Choosing a method*

In their most popular forms, inverse-weighting and stratification estimate the ACE_{All}, matching estimates the ACE_{Exp} and covariate-adjustment may estimate neither. All four methods can, as

described, be adapted to estimate all three causal effects considered. Each method has different advantages and disadvantages. Inverse-weighting produces unbiased estimates of the ACE_{All} but these can have very large variances if some subjects have large weights (i.e. are replicated many times). Stratified estimates of exposure effect can be written as inversely weighted estimates, using a 'crude' propensity score model.[21] If a small number of strata are used residual confounding within strata will cause bias. This bias can, however, be greatly reduced by creating more strata. The more strata used, the closer stratification comes to the matching procedure. Matching estimates can have large variance when many data are discarded in the matching process. Compared with inverse-weighting and covariate-adjustment, stratification and matching may be more robust to wrongly modelling the propensity score (Section 4). There is no clear optimal method. Since including the propensity score as a covariate relies on stronger assumptions than the other three methods, however, some advise using the other methods in preference.[22]

4 Estimating the propensity score

In medical applications, the propensity score is typically estimated using a logistic regression model of exposure on a group of observed covariates,[23] with the latter selected as either potential confounders or predictors of exposure using previous studies and expert knowledge.[8] Further selection is often performed using an automated procedure such as stepwise regression.

4.1 Functional form of propensity score model

Propensity score matching and stratification have been found to be robust to different choices of link functions (e.g. probit, logit and identity) under the SITA assumption (no unobserved confounding).[24,25] This robustness may not carry over to the covariate-adjustment and inverse-weighting methods since, whilst correct strata and matches would be selected using any function that correctly preserved the ordering of the sample,[25] these methods rely on the actual value of the propensity score. Non- and semi-parametric alternatives to the logistic model have been proposed but not frequently applied.[26,27]

4.2 Variable selection

Until fairly recently, received wisdom has been that very rich propensity score models including many interactions and non-linear terms were advisable. However, concerns have been raised about overfitting these models.[2,28] The propensity score model must include all confounders (Section 3). Failure to do this will result in a biased estimate of exposure effect.[24] Additionally including variables related only to either exposure or outcome, however, would not create bias. Including predictors of outcome but not exposure typically decreases the variance of the estimate of exposure effect because their inclusion will correct for any chance imbalances of these covariates between exposure groups.[4] This correction for chance as well as systematic imbalance is the reason why the estimated propensity score has often been found to perform better than the true (population) propensity score.[1,18,29] Conversely, including predictors of exposure but not outcome will generally increase the variance of the estimate of exposure effect.[22] Omitting confounders that are strongly related to exposure but only weakly to outcome may be desirable in smaller samples since the increase in bias may be dominated by the decrease in variance.[30]

4.3 Propensity score diagnostics

Just as the success of randomisation is assessed by the degree of balance achieved between trial arms, the success of a propensity score model is assessed by the balance achieved between exposure groups.[22] However, criteria for adequate balance have been described as 'ill-defined'.[20,23] Although hypothesis testing is frequently used to assess balance, this practice is not universally accepted due to the influence of sample size on the significance tests.[8] Graphical methods for assessing balance within the matched sample or strata include box-plots and quantile–quantile plots of the estimated propensity score comparing exposure groups. The degree of imbalance can be quantified by the percentage standardised difference – the difference in means of a covariate between exposure groups divided by a pooled standard deviation – within the matched sample, strata or between potential samples.[31,32] Assessing imbalance of standard deviations may also be advisable.[20,33] In practice, imbalance is often found after estimation of the propensity score necessitating further refinement of the propensity score model. This leads to a iterative process of estimation and balance-checking.[18]

4.4 Detecting and adjusting for omitted confounders

Large observational datasets often contain imperfectly measured variables and may not include all confounders. It is sometimes possible, however, to collect a richer set of data for a sub-sample including more accurate measurements and additional information. In this case, the relationship between the 'true' propensity score and the mis-measured propensity score can be assessed within the sub-sample and used to correct the estimates of exposure effect from the propensity score analysis. This is known as propensity score calibration.[34] Goodness-of-fit tests have not been found useful for identifying omitted confounders.[35] Sensitivity analyses for assessing the potential impact of unmeasured confounders have been proposed for propensity score matching and stratification methods[36] and extended to inverse-weighting.[26]

4.5 The common support condition

If an exposed subject has a higher estimated propensity score than any unexposed subject, this suggests that there is no unexposed subject who is comparable to the exposed subject, violating the positivity assumption (in this sample). It has therefore become common practice to impose the 'common support' condition – to exclude data from exposed subjects with a propensity score higher than any unexposed subject.[37] Data are similarly excluded from unexposed subjects with low propensity scores. As when unmatched subjects are discarded, imposing this condition may change the population parameter being estimated.

5 Extensions and further research

5.1 Non-linear link functions

Propensity score methods can be directly applied to binary outcomes to estimate the risk difference, and can be easily adapted to estimate the risk ratio or odds ratio, if desired.[38,39] In the odds ratio setting, however, the definition of the causal estimand becomes much more complex due to a property called the 'non-collapsibility' of the odds ratio.[40] For example, if p_1 denotes the proportion of the whole population who would have the outcome of interest if the whole population were exposed and p_0 the analogous unexposed proportion, the marginal odds ratio is $\{p_1/(1-p_1)\} \ / \ \{p_0/(1-p_0)\}$. By splitting the population into low and high SES, as in the

breastfeeding example, calculating a sub-population odds ratio within each SES group and combining the two in a weighted average, we arrive at a different population quantity – a conditional odds ratio – even if the two sub-population odds ratios are equal. This will typically be further from 1 than the marginal odds ratio. Since there are many characteristics that we could stratify (condition) on, there are many different population conditional odds ratios. This non-collapsibility property leads to the little appreciated fact that adding a covariate to a logistic regression model changes the population quantity being estimated – to a different conditional odds ratio.

A multivariable logistic regression model estimates an odds ratio conditional on all covariates in the model. Covariate-adjustment and stratification on the propensity score each estimate a conditional odds ratio, conditional only on the propensity score. By conditioning only on a single variable, the conditional odds ratio from these propensity score methods may be closer to the marginal odds ratio than that from a fully adjusted logistic regression model.[41] Matching on the propensity score (in its most popular form, estimating the ACE_{Exp}) restricts attention to the exposed sub-population and thus does not estimate the marginal odds ratio in the whole population. Conversely, inverse-weighting by the propensity score does estimate the marginal odds ratio. These ideas explain results from simulation studies showing that the conditional odds ratio estimated by a fully adjusted logistic regression model is poorly estimated by propensity score methods,[38] and that whilst the inverse-weighting method estimates the marginal odds ratio well[42] other propensity score methods perform less well.[43]

Since the marginal odds ratio is a single well-defined measure it could be argued that this is the preferred odds ratio.[42] However, it is a well-defined measure only given a well-defined population – different populations may well have different marginal odds ratios! And it could be argued that in a particular setting the conditional odds ratio given sex, for example, captures the mechanistic differences and thus is the most appropriate odds ratio. Due to this controversy, some question the use of the odds ratio as a measure of effect and advocate the use of collapsible measures such as the risk ratio.[20,41,44]

5.2 More than two exposure groups

When the exposure has K categories, each individual has K potential outcomes. One way of defining the causal effect would be to compare each level of exposure with a chosen baseline level, creating $K-1$ population mean differences similar to the ACE_{All}. A generalised propensity score (GPS) has been proposed, essentially a set of $K-1$ propensity scores of the form: $e_k(X) = \mathbb{P}(Z = k| X)$.[45] These can be estimated, for example, by a multinomial or ordinal regression model.[4,45] The GPS has been shown to possess the balancing property of the propensity score under an extended version of the SITA assumption, and can be incorporated into an inverse-weighting approach.[45]

5.3 Variance estimation

Variance estimation and confidence interval (CI) calculation for propensity score analyses is an area that has been relatively neglected. Although estimation of the propensity score is widely held to reduce the variance of the estimate of exposure effect this is rarely taken into account in the formation of 95% CIs. This omission may lead to conservative CIs and hypothesis tests. Theoretical variance formulæ, correcting for the estimation of the propensity score, have been derived for the inverse-weighting method[17] and stratification[46] but these corrections are not

routinely implemented in statistical software. Re-sampling approaches such as bootstrapping may be of use here and have shown promising results in simulation studies.[47,48]

6 Application

6.1 Introduction

In this section, we use data from two parallel trials of infant feeding to investigate the question introduced earlier: does provision of maternal breast milk improve subsequent neurodevelopment of the infant? A subset of these data has been previously used to investigate the same question.[49]

6.2 Methods

926 babies born at <37 weeks gestation, who were under 1850 g at birth and admitted to one of five centres in the United Kingdom were enrolled into two separate randomised trials. In each trial, the mother chose whether or not to provide expressed breast milk for the infant. The randomised trial diet (Trial 1: term formula vs pre-term formula; Trial 2: banked donor milk vs pre-term formula) was given until either discharge from the neonatal unit or achieving a weight of 2000 g. For mothers who chose to provide breast milk, the trial diet was supplementary and given according to the mother's success in providing her milk. More details of the trials can be found elsewhere.[50,51] Breast milk was delivered via tubes, removing effects due solely to the process of breastfeeding.

Children were assessed at approximately 7.5 years of age. IQ was measured as described previously (Section 2.1). Social class (SES) was coded using the Registrar General's classification based on occupation of the income-providing parent, or the father's where both parents were in paid employment. Mother's education was coded as: 1 = no educational attainments, 2 = up to four passes for the certificate of secondary education (CSE), 3 = any 'O' levels or more than four CSEs, 4 = any 'A' levels, and 5 = degree or higher professional qualifications. Information was collected about family structure and marital status (parents living together versus single mother). Details of the pregnancy, labour, delivery and the neonatal period were recorded.

6.2.1 Statistical analysis

The propensity score – the conditional probability of choosing to provide breast milk – was estimated using a logistic regression model, where the outcome was choice to provide breast milk (yes/no). It included the following covariates: marital status, social class, maternal educational level and age infant gender, birthweight, gestational age, and birth order in the family. All covariates except gender and marital status were entered into the model as (centred) continuous variables, after initial models indicated approximately linear trends. Non-linearities of all continuous variables were investigated, as were interactions of all covariates with the two binary covariates. Histograms, kernel density plots and box-plots were used to graph the estimated propensity score. To assess balance, the percentage standardised difference for each covariate – the mean difference of the covariate between exposure groups, divided by a pooled estimate of the standard deviation (SD), and multiplied by 100 – was calculated for the whole dataset, within the matched sample, and for the potential samples. For the stratified estimator, the within-strata differences were weighted according to the fraction of the sample in each strata before dividing by the pooled SD.

The three causal effects of the choice to provide breast milk on IQ – the ACE_{All}, the ACE_{Exp} and the ACE_{Un} – were estimated using the four methods described. For the stratification, 18 strata were used, obtained by collapsing the top three categories of 20 equal-sized strata to ensure adequate cell frequencies. Three matched samples, one for each causal effect, were each created by matching pairs

(one exposed and one unexposed) such that each pair of estimated propensity scores were no further than 0.05 apart. In each case, matching was performed with replacement. Three sets of potential samples were created as previously described for the inverse weighting method. Finally, a linear regression model for IQ on the estimated propensity score and exposure, with and without an exposure by propensity score interaction, was fitted. From the model including the interaction, the three causal effects were estimated by evaluating the exposure effect at three propensity score values: the sample mean, the mean amongst the exposed subset of the sample, and the mean amongst the unexposed subset of the sample. These will correctly estimate the ACE_{All}, the ACE_{Exp} and the ACE_{Un}, respectively, if the relationship between IQ and the propensity score is approximately linear within each exposure group. For the purposes of comparison, a traditional linear regression model for IQ was fitted including variables on the basis of whether or not their inclusion changed the estimated association between choice to breastfeed and IQ by more than 10%. All measured covariates were considered as possible confounders. Non-linearities and interactions were assessed. For all estimates, bootstrap bias-corrected and accelerated (BCa) CIs,[52] calculated from 5000 replications, are presented. The propensity score was re-estimated for each bootstrap sample, to allow for the estimation of the propensity score within the BCa CIs.

6.3 Results

In total, demographic and follow-up IQ data were obtained for 487 children. Of these, the mothers of 340 (69.8%) had chosen to provide breast milk and 147 (30.2%) had not. Mean follow-up age (range) was 7.6 (7.3, 8.2) years amongst children in the maternal breast milk group, and 7.6 (7.4, 8.0) years amongst those in the no maternal breast milk group. Demographic characteristics of the children included in the analysis are shown in Table 1, where SES, maternal educational level and birth order are shown as binary variables for simplicity. There are some suggestions of imbalance across the two exposure groups, most noticeably of SES, maternal educational level and birth order. The estimated propensity score model indicated that factors influencing the choice to provide breast milk were: maternal education, maternal age, SES, marital status and birth order. Non-linearity of the relationship between mother's education and the choice to provide breast milk was apparent. Imposing the common support condition dropped 10 children all of whom had received maternal breast milk and had mothers from the highest class of maternal

Table 1. Demographic characteristics by exposure group

Characteristic	No maternal breast milk ($n = 147$)		Maternal breast milk ($n = 340$)	
Social classes I and II	20	(13.6%)	117	(34.4%)
Parents living together	109	(74.1%)	297	(87.4%)
Maternal educational level 3–5	74	(50.3%)	243	(71.5%)
Mean maternal age – years (SD)	26.1	(5.9)	27.5	(5.2)
Mean gestational age – weeks (SD)	30.9	(2.7)	31.2	(2.6)
Mean birth weight – g (SD)	1371.2	(296.0)	1425.1	(292.3)
First child in family	77	(49.0%)	216	(63.5%)
Male child	63	(42.9%)	165	(48.5%)

education who were living with a partner. The estimated propensity score is shown in the upper panels of Figure 4, with a smoothed estimate of the density overlaid, by exposure group.

No subjects were dropped from any matched analysis due to the lack of a suitable match. However, for the matched sample estimating the ACE_{Exp}, 51 (34.7%) of the unexposed children were excluded from the analysis (they were not allocated as a match to any exposed child). For the matched sample estimating the ACE_{Un}, 244 (71.8%) of the exposed children were excluded from the analysis.

Since the estimated propensity score was a continuous function, the potential samples contained non-integer numbers of subjects. The numbers of subjects in the potential exposed and unexposed samples created to estimate the ACE_{All}, the ACE_{Exp}, and the ACE_{Un} were: (489.2, 515.1), (340, 368.1) and (149.2, 147), respectively.

The lower panels of Figure 4 show the distributions of the estimated propensity score by exposure group in one of the matched samples (estimating the ACE_{Exp}) and in one set of the potential samples (estimating the ACE_{All}). Both these methods appear to have balanced the propensity score distribution between exposure groups, as expected. Table 2 shows percentage standardised differences of the eight covariates. The large imbalances present in the original dataset have been greatly reduced by matching, stratification and inverse-weighting. Based on the change-in-estimate criterion, the outcome linear regression model included only the following covariates: choice to provide breast milk, maternal education (categorical), SES (categorical) and birthweight

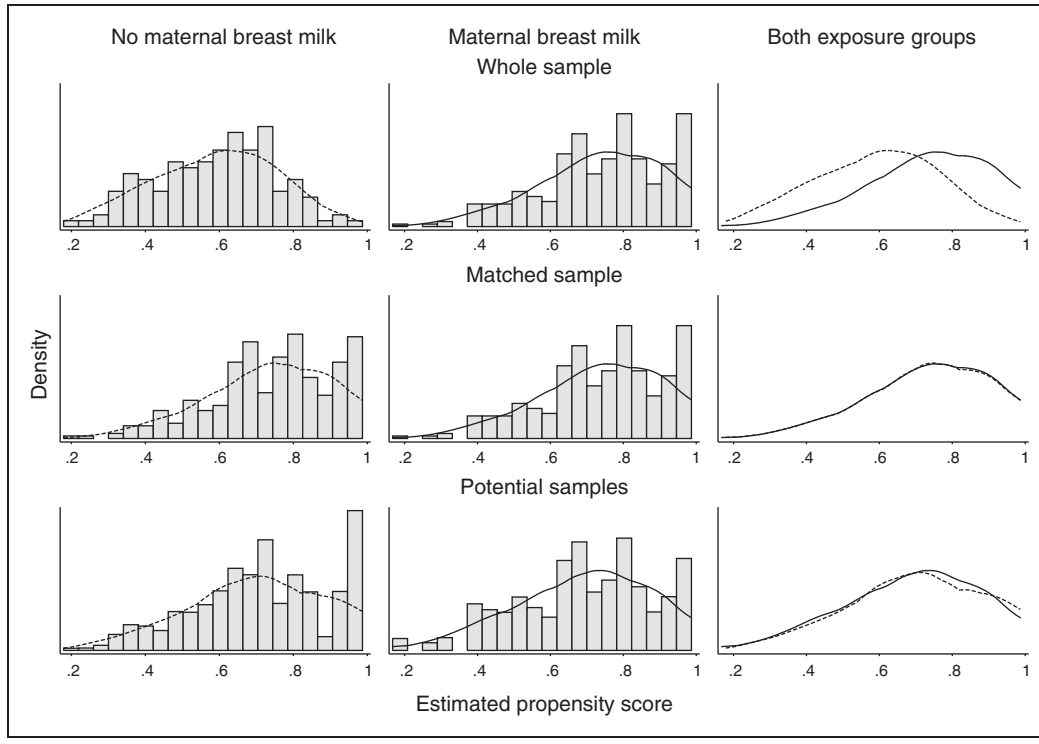

Figure 4. Propensity score distribution in the whole sample, the matched sample (estimating the ACE_{Exp}) and the potential samples created by inverse weighting (estimating the ACE_{All}).

Table 2. Balance of demographic characteristics

Characteristic	Percentage standardised difference			
	Whole sample	Matched sample	Potential samples	Strata
Social class category score	55.3	15.3	9.6	8.0
Parents living together	34.0	7.4	1.6	0.2
Maternal education category score	71.2	1.1	12.6	1.1
Mean maternal age (years)	24.5	7.9	16.2	3.0
Mean gestational age (weeks)	13.2	3.7	2.4	5.7
Mean birth weight (g)	18.3	13.7	16.3	4.7
First child in family	28.8	12.7	4.5	6.9
Male child	11.4	0	8.7	9.7

(continuous). This model was fitted with and without an interaction between choice to provide breast milk and maternal education.

Mean IQ at follow-up (SD) was 105.2 (12.5) amongst children in the maternal breast milk group, and 98.1 (12.4) amongst those in the no maternal breast milk group. In this example, the three causal effects are likely to be distinct from one another since the inclusion of an interaction term in the regression model of IQ on exposure and estimated propensity score indicated some evidence that the exposure effect increases with the propensity score (likelihood ratio test for interaction $p = 0.06$).

Table 3 shows the estimates of the causal effects of choice to provide maternal breast milk on IQ at 7.5 years obtained from the four different propensity score methods and traditional regression models, with all but the unadjusted estimate shown in Figure 5. The crude estimate of exposure effect, an increase of 7.1 IQ units (95% CI: 4.6–9.5), is reduced by all four propensity score methods.

The four propensity score methods each produced similar estimates of the three causal effects. In all cases, the ACE_{Exp} was higher than the ACE_{All} which in turn was higher than the ACE_{Un}. The inverse weighting method produced slightly higher estimates of the ACE_{All} and the ACE_{Exp} in comparison with the other propensity score methods. In this dataset, one unexposed subject had a weight of 75, five times larger than any other subject, when calculating the ACE_{All} and the ACE_{Exp} (but a weight of 1 when calculating the ACE_{Un}). After excluding this subject from the analysis, the estimates become $\widehat{ACE}_{All} = 4.08\,(1.38, 6.92)$ and $\widehat{ACE}_{Exp} = 4.53\,(1.55, 7.93)$. There was some evidence of an interaction between maternal education and choice to provide breast milk in the traditional outcome model (likelihood ratio test $p = 0.06$). Including this interaction term and evaluating the exposure effect at maternal education category 3 – slightly below the mean education level in the sample – produced a very high estimate of effect with an extremely wide CI.

6.4 Summary of findings about breastfeeding and IQ

All three causal effects might be of interest in this example, but the ACE_{Un} may be the most relevant for informing interventions promoting breastfeeding. In this example, there was some suggestion of a non-linear relationship between propensity score and outcome so the covariate-adjusted method without including an exposure-propensity score interaction is likely to produce biased conclusions. Similarly, the outcome regression model ignoring the suggestion of non-linearity in the exposure effect may produce misleading estimates. The inverse weighting method weighted one subject

Table 3. Estimates of the difference in mean IQ points associated with the choice to provide maternal breast milk

Method	Estimand	Estimate	(95% CI)
Outcome regression			
Unadjusted	–	7.09	(4.64, 9.49)
Adjusted	–	3.26	(0.77, 5.66)
Adjusted with interaction[a]	–	6.52	(0.36, 12.21)
Propensity score methods			
Stratification	ACE_{All}	4.13	(0.61, 6.75)
	ACE_{Exp}	4.84	(0.28, 7.87)
	ACE_{Un}	2.49	(−0.69, 5.41)
Matching	ACE_{All}	4.61	(1.79, 8.16)
	ACE_{Exp}	5.13	(1.18, 9.29)
	ACE_{Un}	3.39	(0.03, 8.37)
Weighting	ACE_{All}	5.10	(2.25, 8.20)
	ACE_{Exp}	6.06	(2.75, 9.88)
	ACE_{Un}	2.98	(0.29, 6.44)
Covariate-adjustment	ACE_{All}	3.78	(0.96, 6.44)
(with interaction)[b]	ACE_{Exp}	4.39	(1.01, 7.30)
	ACE_{Un}	2.39	(−0.09, 5.10)
(without interaction)[b]	–	2.97	(0.45, 5.41)

Notes: [a]of exposure and mother's education, the effect given is for education category 3.
[b]of exposure and propensity score.

disproportionately to the rest of the sample and estimates were sensitive to this subject's inclusion. The estimates produced by the covariate-adjusted method rely on the relationship between outcome and propensity score being linear within each exposure group. The similarity between these estimates and those from the other methods suggests that this assumption holds at least approximately.

In this example, all methods lead to fairly similar conclusions, including the fully adjusted regression model, all suggesting a possible positive association between choice to provide maternal breast milk and subsequent IQ. However, it is important to bear in mind the likelihood of unmeasured confounding.

7 Discussion: Propensity score methods or traditional outcome regression?

We attempt to summarise the main points that have been made for and against propensity score methods, in comparison with outcome regression models.

7.1 Advantages and disadvantages of the propensity score approach

A key advantage of the propensity score approach is that model selection can be performed blind to outcome status, thereby minimising bias due to prior beliefs concerning the outcome–exposure association. Since only one outcome–exposure association is examined, the analysis and reporting of results is likely to be more focused than from an outcome modelling approach.

It has recently been stated that propensity score methods treat 'bias as paramount and variance as secondary'.[2] This is based on the observation that the variance of the exposure effect estimated from

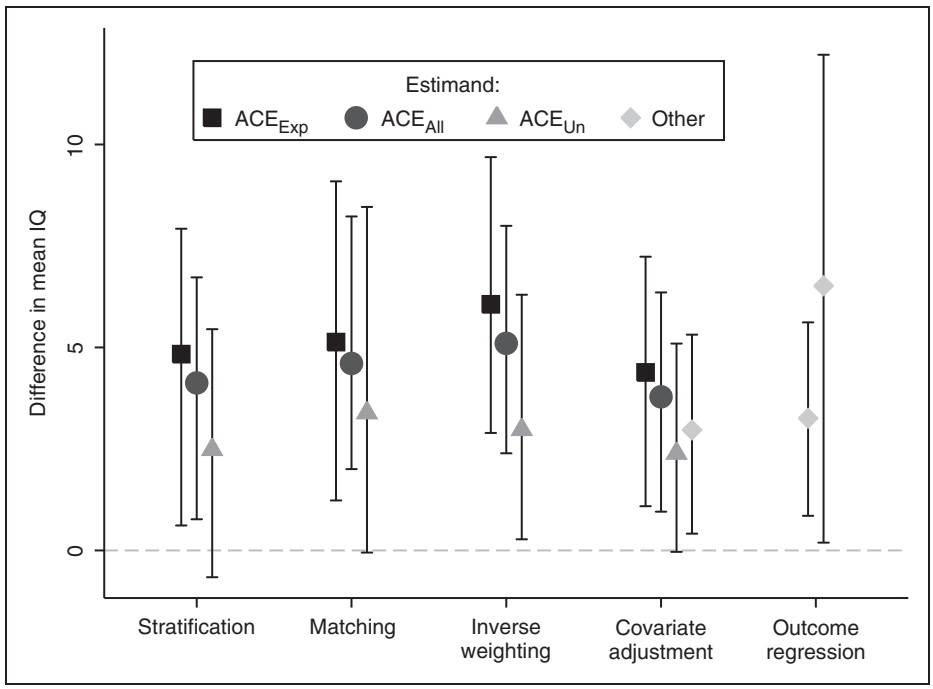

Figure 5. Estimates of the difference in mean IQ points associated with the choice to provide maternal breast milk with 95% CIs (see Table 3 for details).

a propensity score analysis (with an estimated propensity score) will be larger than that from a correctly specified outcome regression model. This has been demonstrated for both the conditional variance (i.e. the variance whilst holding the Xs fixed)[2] and the marginal variance (viewing the Xs as random variables).[53] Whichever variance is of interest, it is clear that purely on the basis of efficiency, outcome regression will always be preferable. However, if the outcome regression model is incorrectly specified, the exposure effect estimate will be biased. Propensity score methods are often seen as more robust to model misspecification than outcome regression models. Matching and stratification on the propensity score do, indeed, appear to be more robust to the mis-modelling of non-linearities than outcome regression models,[24] although it is not clear that this property is shared by the other propensity score methods.

An advantage of the propensity score approach is that a key assumption – the balance of covariates – can be easily checked, which provides a certain robustness.[54] Since the propensity score is a means to an end (balance), incorrect modelling does not matter as long as balance is achieved. However, the criteria for declaring covariate balance have not, as yet, been properly defined.[20]

The randomisation argument for the propensity score approach can be highly persuasive, perhaps too persuasive since there is a tendency to overlook the strong underlying assumption of no unobserved confounders. Balance of observed covariates does not guarantee balance of unobserved covariates. Omitting a confounder from the propensity score model produces biases similar to those produced by omitting a confounder from an outcome regression model.[24] Interactions, similarly, if omitted, can produce bias.

The balancing aspect of propensity scores is, furthermore, a large-sample property. In the absence of unmeasured confounding, propensity scores can be likened to a RCT at each value of the propensity score. Thus, a large sample is needed at each value of the propensity score to achieve balance.[21]

A common criticism of outcome regression models concerns insufficient overlap of the multivariate covariate distributions in the exposed and unexposed groups. In this situation, there may be exposed subjects who are intrinsically non-comparable with anyone in the unexposed group. Outcome regression responds to this problem with extrapolation, which may be inappropriate.[22] A histogram of the estimated propensity score in the exposed and unexposed groups would immediately draw attention to the problem and may lead to the valid conclusion that for those subjects it is not possible to estimate the exposure effect and to discard them from the analysis.

When the within-strata causal exposure effects differ, it has been suggested that propensity score stratification is inappropriate since these within-strata effects are meaningless.[2] However, whilst outcome regression offers the possibility of investigating interaction terms, these are not easily presented or interpreted. Arguably, it would be preferable in both cases to identify subgroups in which the exposure effect is constant.

7.2 Conclusion

In some situations – studying a rare outcome or a rare exposure – there may be clear advantages to modelling the propensity score in preference to the outcome or *vice versa*. At the other extreme, the two approaches sometimes coincide. For example, when all covariates are discrete, stratification on the propensity score produces exactly the same estimate as linear regression, if both the outcome model and the propensity score model contain all possible interaction terms.[2] More generally, there are likely to be advantages to both traditional outcome regression modelling and propensity score methods in most situations, although in practice these two approaches have been found to produce similar conclusions.[55] Combinations of the two approaches have been advocated by many including covariate-adjustment within propensity score strata or matched samples[17,56] and doubly robust inverse-weighting methods.[17] The latter are robust to misspecification of either the outcome regression model or the propensity score model.

In the light of our discussions above, we feel that that the key questions for future research with regard to propensity score methods are: improving the estimation of CIs for the propensity score, investigating the use of propensity score methods in generalised and other non-linear models, and establishing a consensus about 'best practice' in the use of propensity scores, particularly with regard to assessing balance. A further area that has been little explored is how best to handle missing data in propensity score analyses, although recent studies have begun to address this.[57,58]

In this article, we have given an intuitive review of the propensity score, addressing potential misunderstandings about the methods based on the propensity score, and in particular their use for estimating the average causal effect, the average causal effect on the exposed and the average causal effect on the unexposed. In practice, calculation and graphical exploration of the propensity score in the exposed and unexposed groups is an invaluable diagnostic tool, whether or not it is formally used in the estimation. Finally, it is important to keep a sense of perspective. In many analyses, biases due to unobserved confounding in observational studies are larger than possible differences between the propensity score methods and traditional outcome regression models.

Acknowledgements

JRC is funded by ESRC Research Fellowship RES-063-27-0257. This study was undertaken as part of EW's PhD funded by a GlaxoSmithKline studentship. We thank the reviewers for their comments which have led to a greatly improved manuscript.

References

1. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
2. Senn S, Graf E and Caputo A. Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. *Stat Med* 2007; **26**: 5529–5544.
3. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998; **17**: 2265–2281.
4. Joffe MM and Rosenbaum PR. Invited commentary: Propensity scores. *Am J Epidemiol* 1999; **150**: 327–333.
5. Austin PC and Mamdani MM. A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Stat Med* 2006; **25**: 2084–2106.
6. Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol* 2006; **163**: 262–270.
7. Austin PC. Discussion of 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003'. *Stat Med* 2008; **27**: 2066–2069.
8. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med* 2008; **27**: 2037–2049.
9. Rodgers B. Feeding in infancy and later ability and attainment: a longitudinal study. *Dev Med Child Neurol* 1978; **20**: 421–426.
10. Weschler D. *Weschler Intelligence Scale for Children, Anglicized revised edition*. Sidcup, Kent: The Psychological Corporation Ltd.; 1974.
11. Holland PW. Statistics and causal inference. *J Am Stat Assoc* 1986; **81**: 945–960.
12. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev Econ Stat* 2004; **86**: 4–29.
13. Cole SR and Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008; **168**: 656–664.
14. Rubin DB. Bias reduction using Mahalanobis-metric matching. *Biometrics* 1980; **36**: 293–298.
15. Little R and Rubin D. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Ann Rev Public Health* 2000; **21**: 121–145.
16. Hullsiek KH and Louis TA. Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics* 2002; **3**: 179–193.
17. 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.
18. Rosenbaum PR and Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 1984; **79**: 516–524.
19. Baser O. Too much ado about propensity score models? Comparing methods of propensity score matching. *Value Health* 2006; **9**: 377–385.
20. Hill J. Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Stat Med* 2008; **27**: 2055–2061.
21. Zhao Z. Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Rev Econ Stat* 2004; **86**: 91–107.
22. Rubin DB. On principles for modeling propensity scores in medical research. *Pharmacoepidemiol Drug Safe* 2004; **13**: 855–857.
23. Weitzen S, Lapane KL, Toledano AY, Hume AL and Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Safe* 2004; **13**: 841–853.
24. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 1993; **49**: 1231–1236.
25. Zhao Z. Sensitivity of propensity score methods to the specifications. *Econ Lett* 2008; **98**: 309–319.
26. McCaffrey DF, Ridgeway G and Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods* 2004; **9**: 403–425.
27. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ and Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Safe* 2008; **17**: 546–555.
28. Judkins DR, Morganstein D, Zador P, Piesse A, Barrett B and Mukhopadhyay P. Variable selection and raking in propensity scoring. *Stat Med* 2007; **26**: 1022–1033.
29. Gu XS and Rosenbaum PR. Comparison of multivariate matching methods: structures, distances and algorithms. *J Comput Graph Stat* 1993; **2**: 405–420.
30. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J and Stuürmer T. Variable selection for propensity score models. *Am J Epidemiol* 2006; **163**: 1149–1156.
31. Austin PC. Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. *Pharmacoepidemiol Drug Safe* 2008; **17**: 1218–1225.
32. Austin PC. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidemiol Drug Safe* 2008; **17**: 1202–1217.
33. Stuart EA. Developing practical recommendations for the use of propensity scores: Discussion of 'A critical appraisal of propensity score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Stat Med* 2008; **27**: 2062–2065.
34. Stuürmer T, Schneeweiss S, Avorn J and Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am J Epidemiol* 2005; **162**: 279–289.
35. Weitzen S, Lapane KL, Toledano AY, Hume AL and Mor V. Weaknesses of goodness-of-fit tests for evaluating

propensity score models: the case of the omitted confounder. *Pharmacoepidemiol Drug Safe* 2005; **14**: 227–238.
36. Rosenbaum PR. *Observational studies*, 2nd ed. New York: Springer-Verlag, 2002.
37. Heckman J, Ichimura H, Smith J and Todd P. Characterizing selection bias using experimental data. *Econometrica* 1998; **66**: 1017–1098.
38. Austin PC, Grootendorst P, Normand SLT and Anderson GM. 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.
39. Austin PC. The performance of different propensity-score methods for estimating relative risks. *J Clin Epidemiol* 2008; **61**: 537–545.
40. Greenland S, Robins JM and Pearl J. Confounding and Collapsibility in Causal Inference. *Stat Sci* 1999; **14**: 29–46.
41. Martens EP, Pestman WR and Klungel OH. Letter to the Editor: Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. *Stat Med* 2007; **26**: 3205–3212.
42. 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.
43. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med* 2007; **26**: 3078–3094.
44. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol* 1987; **125**: 761–768.
45. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika* 2000; **87**: 706–710.
46. Williamson E. Inference from estimates of exposure effects using stratification on the propensity score. PhD thesis, London School of Hygiene & Tropical Medicine, London, 2008.
47. Hill J and Reiter JP. Interval estimation for treatment effects using propensity score matching. *Stat Med* 2006; **25**: 2230–2256.
48. Tu WZ and Zhou XH. A bootstrap confidence interval procedure for the treatment effect using propensity score subclassification. *Health Serv Outcome Res Methodol* 2002; **3**: 135–147.
49. Lucas A, Morley R, Cole TJ, Lister G and Leeson-Payne C. Breast-milk and subsequent intelligence quotient in children born preterm. *Lancet* 1992; **339**: 261–264.
50. Lucas A, Gore SM, Cole TJ, et al. Multicenter trial on feeding low-birthweight infants — effects of diet on early growth. *Archives of Disease in Childhood* 1984; **59**: 722–730.
51. Lucas A, Morley R, Cole TJ, et al. Early diet in preterm babies and developmental status at 18 months. *Lancet* 1990; **335**: 1477–1481.
52. Carpenter J and Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med* 2000; **19**: 1141–1164.
53. Robins JM, Mark SD and Newey WK. Estimating exposure effects by modeling the expectation of exposure conditional on confounders. *Biometrics* 1992; **48**: 479–495.
54. Fitzmaurice G. Confounding: propensity score adjustment. *Nutrition* 2006; **22**: 1214–1216.
55. Shah BR, Laupacis A, Hux JE and Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol* 2005; **58**: 550–559.
56. Rubin DB and Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *J Am Stat Assoc* 2000; **95**: 573–585.
57. Mattei A. Estimating and using propensity score in presence of missing background data: an application to assess the impact of childbearing on wellbeing. *Stat Methods Appl* 2009; **18**: 257–273.
58. Qu Y and Lipkovich I. Propensity score estimation with missing values using a mutiple imputation missingness pattern (MIMP) approach. *Stat Med* 2009; **28**: 1402–1414.