American Journal of Epidemiology

Vol. 169, No. 7
DOI: 10.1093/aje/kwn391
Advance Access publication January 19, 2009

Practice of Epidemiology

Different Methods of Balancing Covariates Leading to Different Effect Estimates in the Presence of Effect Modification

Mark Lunt, Daniel Solomon, Kenneth Rothman, Robert Glynn, Kimme Hyrich, Deborah P. M. Symmons, Til Stürmer, the British Society for Rheumatology Biologics Register, and the British Society for Rheumatology Biologics Register Control Centre Consortium

Initially submitted May 9, 2008; accepted for publication November 20, 2008.

A number of covariate-balancing methods, based on the propensity score, are widely used to estimate treatment effects in observational studies. If the treatment effect varies with the propensity score, however, different methods can give very different answers. The authors illustrate this effect by using data from a United Kingdom–based registry of subjects treated with anti–tumor necrosis factor drugs for rheumatoid arthritis. Estimates of the effect of these drugs on mortality varied from a relative risk of 0.4 (95% confidence interval: 0.16, 0.91) to a relative risk of 1.3 (95% confidence interval: 0.8, 2.25), depending on the balancing method chosen. The authors show that these differences were due to a combination of an interaction between propensity score and treatment effect and to differences in weighting subjects with different propensity scores. Thus, the methods are being used to calculate average treatment effects in populations with very different distributions of effect-modifying variables, resulting in different overall estimates. This phenomenon highlights the importance of careful selection of the covariate-balancing method so that the overall estimate has a meaningful interpretation.

covariate balance; effect modification; observational study; propensity score; weighting

Abbreviation: TNF, tumor necrosis factor.

It is widely accepted that randomized controlled trials provide the best evidence for the effect of drug treatments. Nevertheless, in situations in which a randomized trial is impractical or would take a long time to complete, valuable information can be gathered from an observational study, provided that the study is designed and analyzed appropriately. In general, exposed and unexposed subjects in an observational study will differ regarding a number of variables related to the outcome under study, and balancing these variables is required. Many of the balancing methods proposed involve the propensity score (1), used for either stratifying subjects (2), matching (3), or weighting (4).

Drug registries are becoming more widely used to assess the long-term effectiveness and safety of drug treatments. Such registries commonly aim to include all subjects administered a particular treatment at a given point in time and subjects unexposed to the drug under study to act as a comparison group. The subjects in a registry, however, may not be representative of all those using a particular drug. For example, the drug may be used initially by subjects with severe disease; later, access is widened to include subjects with milder disease. Propensity models generally assume either that the effect of treatment is the same for all subjects or that a mean treatment effect for some particular population is the object of inference. If the effect of treatment depends on patient-specific factors, then 2 populations with different distributions of these factors will produce different estimates of the treatment effect.

In this paper, we present an example of using propensity methods to balance covariates when assessing the rate of adverse events in drug registry data. The treatment effect measure varied considerably with the propensity score, causing different propensity-based balancing methods to produce very different effect estimates. Thus, the expected

Correspondence to Dr. Mark Lunt, arc Epidemiology Unit, University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, United Kingdom (e-mail: mark.lunt@manchester.ac.uk).

effect of treatment would depend on the characteristics of the population to which it was given.

MATERIALS AND METHODS

Subjects

The subjects in this analysis were enrolled in the British Society for Rheumatology Biologics Register. The primary aim of the register is to examine the medium-term to long-term safety of biologic drugs in patients with rheumatic disease. The methodology has been described in detail elsewhere (5). In brief, the register consists of a cohort of rheumatoid arthritis patients treated with anti–tumor necrosis factor (TNF)-α therapy (the exposed) and an anti-TNF-α-naïve comparison cohort of rheumatoid arthritis patients treated with non-anti-TNF-α therapy (the unexposed). The United Kingdom national guidelines recommend that anti-TNF-α drugs be reserved for patients with active rheumatoid arthritis (defined as a 28-joint-count Disease Activity Score (6) of >5.1) despite previous therapy with at least 2 disease-modifying antirheumatic drugs, one of which should be methotrexate (7), and that "any clinician prescribing these medications must (with the patient's permission) undertake to register the patient with the BSRBR and forward information on dosage, outcome and toxicity on a six-monthly basis" (8, p. 4, section 1.5). The unexposed patients should have a physician diagnosis of rheumatoid arthritis with active disease (guideline 28-joint-count Disease Activity Score of ≥4.2) despite current treatment with a disease-modifying antirheumatic drug and be biologically naïve. Unexposed subjects were recruited during a visit to their physician, when their treatment could have been changed, so that their baseline was comparable with that for exposed subjects.

For both exposed and unexposed subjects, data were collected at baseline on demographics, disease duration, 28-joint-count Disease Activity Score, Health Assessment Questionnaire score (adapted for British use (9)), body mass index, smoking history, previous and current drug use, and a number of comorbidities (listed in Table 1). In this analysis, we used death as an endpoint: all patients in the register were "flagged" with the Office for National Statistics, so we were informed of all deaths and received copies of the death certificates.

Covariate-balancing methods

A number of methods of balancing baseline covariates between exposed and unexposed subjects are available. The methods examined in this analysis all utilize a propensity score. The propensity score, introduced by Rosenbaum and Rubin (1), reflects the probability that a patient with a given set of covariate values would receive treatment and is most commonly estimated by using logistic regression. It is sometimes more convenient to work with the linear predictor from a logistic regression, that is, the log-odds of treatment or logit, since its distribution is likely to be closer to normal.

The variables used to define the propensity score for this example are listed in Table 1. To allow for nonlinearity in the association between the continuous predictors and the log-odds of treatment, powers of the 6 continuous variables up to the sixth were included. All 2-way product terms for which the P value was less than 0.05 were included in the propensity score model.

Stratifying. The first, and most common, balancing method is stratification on the propensity score. The population is divided into subgroups based on estimated propensity score, and the exposed and unexposed subjects are compared within strata of propensity score. In this way, exposed subjects are compared with unexposed subjects whose propensity scores are similar. It has been shown that using 5 strata can be expected to remove about 90% of the confounding bias introduced by a continuous confounder (10). However, the remaining confounding bias means that this method is not asymptotically unbiased.

Weighting. Alternatively, a weighting scheme can be used to balance the covariates (4). With this method, observations are reweighted to form a larger *pseudopopulation* in which the covariates are no longer associated with treatment. Let the probability of receiving treatment at a given level of the covariates, x, be $p_1(x)$, and let $p_0(x) = 1 - p_1(x)$ be the probability of not receiving treatment. Then, sampling weights for the treated subjects, $w_1(x)$, and untreated subjects, $w_0(x)$, are selected so that the odds of receiving treatment in the pseudopopulation, $\frac{w_1(x)p_1(x)}{w_0(x)p_0(x)}$, do not depend on x. For example, if $w_1(x) = 1/p_1(x)$ and $w_0(x) = 1/p_0(x) = 1/(1 - p_1(x))$, then the odds of receiving treatment are given by

$$\begin{aligned}\text{Odds} &= \frac{w_1(x)p_1(x)}{w_0(x)p_0(x)}\\ &= \frac{1/p_1(x) \times p_1(x)}{1/p_0(x) \times p_0(x)}\\ &= \frac{1}{1}\\ &= 1\end{aligned}$$

for all values of x. Since the covariates are no longer related to the probability of receiving treatment, they are no longer confounders.

Different definitions of $w_1(x)$ and $w_0(x)$ lead to different distributions of the covariates in the pseudopopulation (11). The weighting scheme described above is referred to as inverse probability of treatment weighting, and it creates the same distribution of covariates in the pseudopopulation as in the entire sample. An alternative weighting scheme, referred to as standardized mortality ratio weighting, has $w_1 = 1$ and $w_0 = p_1(x)/p_0(x)$. This scheme does not alter the distribution of covariates in the treated subjects but rather reweights the distribution in the untreated subjects to make it the same as for the treated. It therefore provides an estimate of the effect of treatment for those who are treated.

Matching. The final method considered was matching. With matching, each exposed subject is paired with the unexposed subject whose propensity score is the closest. The usual procedure involves setting a limit on how close a match needs to be before it can be considered appropriate: this limit is referred to as a caliper.

Matching may be performed either with or without replacement. In matching without replacement, once an unexposed subject has been selected as a match for an exposed one, he or she is removed from the list of potential matches. With matching, it is possible that not every exposed subject

Table 1. Distribution of Baseline Covariates Among Subjects Exposed and Unexposed to Treatment for Rheumatoid Arthritis, and Impact of Adjustment on the Estimated Relative Rate Among Biologic Users vs. Nonusers[a]

Variable	Exposed (n = 8,437)	Unexposed (n = 1,497)	Change in Rate Ratio,[b] %
Age, years	57 (48–65)	61 (53–69)	41
Disease duration, years	12 (6–19)	7 (2–16)	2
Disease Activity Score	6.6 (5.9–7.3)	5.0 (4.1–6.0)	−26
Health Assessment Questionnaire score	2.1 (1.8–2.5)	1.6 (1–2.1)	−22
Systolic blood pressure, mm Hg	135 (120–150)	136 (122–150)	2
Diastolic blood pressure, mm Hg	80 (71–88)	80 (71–86)	1
Gender: female	76.5	72.9	3
No. of previous DMARDs			
1	1.7	29.6	
2	16.1	26.7	
3	21.5	19.2	
4	19.2	12.4	
5	16.1	6.5	
≥6	25.4	5.7	−10
Body mass index[c] group			
<20	8.6	6.1	
≥20–<25	34.3	34.1	
≥25–<30	33.1	35.2	
≥30	24.0	24.6	−1
Smoking			
Never	22.1	21.6	
Former	38.4	39.9	
Current	39.5	38.5	1
History of			
Hypertension	30.4	33.9	3
Angina	4.7	7.7	4
Myocardial infarction	3.1	5.5	4
Stroke	2.1	3.6	3
Epilepsy	1.2	1.6	0
Asthma	9.9	13.2	1
Chronic obstructive pulmonary disease	5.2	9.2	6
Peptic ulcer disease	8.7	8.3	1
Liver disease	2.3	1.9	0
Renal disease	2.9	3.4	0
Tuberculosis	2.0	2.7	0
Demyelin	0.2	0.5	1
Diabetes	5.6	6.0	0
Hyperthyroidism	3.4	4.1	0
Depression	20.0	16.4	1
Cancer	3.3	6.7	3

Abbreviation: DMARDs, disease-modifying antirheumatic drugs.

[a] Values are expressed as median (interquartile range) or percentage.

[b] This column gives $\frac{RR_A - RR_C}{RR_C}$, where RR_C is the crude rate ratio and RR_A is the rate ratio after adjusting for the variable in question. It measures the extent to which the variable is confounding the rate ratio.

[c] Weight (kg)/height $(m)^2$.

will be matched if there are insufficient suitable unexposed subjects. In matching with replacement, each exposed subject is matched to the nearest unexposed subject. In this way, more exposed subjects can be included in the analysis; all will be included provided there is at least one unexposed subject within the caliper, but an unexposed subject may be matched to several exposed subjects.

Assessing balance. The aim of balancing covariates is to create exchangeability of the exposed and unexposed subjects and thus eliminate the confounding effect of variables associated with both the treatment and the outcome. Therefore, both the difference in the distribution of the variable between the exposed and unexposed subjects and the association between the variable and the outcome are important: variables not associated with the outcome do not lead to confounding and therefore do not need to be balanced.

The degree to which an individual variable confounds the association between outcome and exposure after balancing can, for example, be assessed by measuring the propensity-adjusted effect of treatment and then repeating the analysis, further adjusting for the variable of interest to the extent possible. If the estimate of the treatment effect changes, this change implies residual confounding by the covariate and that the balancing was incomplete. When we performed this additional adjustment, continuous variables were fitted after categorizing them into quintiles to avoid assuming a linear association between covariate and outcome.

Relative contribution to rate ratio based on covariate balancing method. The process of matching subjects can be thought of as a form of weighting. In matching with replacement, each exposed subject receives a weight of 1, and each unexposed subject receives a weight equal to the number of exposed subjects to whom he or she is matched (0 for those unexposed subjects not matched to any exposed subject). In matching without replacement, each subject used in the analysis is given a weight of 1, and each subject not used is given a weight of 0.

For stratification, the weighting works slightly differently. Individuals are not weighted, but the maximum likelihood estimate is a weighted mean of the stratum-specific estimates:

$$\theta = \frac{\sum_s W_s \theta_s}{\sum_s W_s},$$

where

$$W_s = \frac{D_{1s} Y_{0s}}{Y_{1s} + \theta Y_{0s}}$$

D_{1s} = number of deaths among exposed subjects in stratum s

Y_{0s} = person-years of follow-up among unexposed subjects in stratum s

Y_{1s} = person-years of follow-up among exposed subjects in stratum s

θ_s = Rate ratio in stratum s

provided that there is at least one death among the unexposed and one death among the exposed within each stratum. We can therefore think of all subjects in stratum s being assigned a weight W_s.

In this analysis, stratifying, weighting, and matching were all used to balance the baseline covariates. In the stratified analysis, 5 strata were defined as the quintiles of the propensity score distribution, with stratum 1 being the least likely to receive treatment and stratum 5 the most likely. Weighting was used to match the distribution of covariates in both the exposed and unexposed groups to the overall distribution in the entire sample (inverse probability of treatment weights) and to the distribution in the exposed subjects (standardized mortality ratio weights). Matching was performed both with and without replacement. For both types of matching, we used the linear predictor of the propensity score (3) as the matching variable, with a caliper of 0.01.

Estimation of rates and rate ratios

The mortality rates and rate ratios for the exposed and unexposed subjects were calculated by using Poisson regression in Stata, version 9.2 software (12). The P values for the differences in rate ratio between propensity quintiles were calculated by fitting indicator variables for the quintiles and calculating a Wald test of the hypothesis that the parameter was constant over all quintiles. To test for a trend across quintiles, the quintile number was fitted as a continuous variable and the coefficient of that variable compared with 0 by using a Wald test. In the weighted analyses, the weights were fitted as probability (sampling) weights so that the standard errors were not artificially decreased by the apparent increase in sample size due to the weighting.

RESULTS

Initial distribution of covariates

This analysis considered 8,437 exposed subjects and 1,497 unexposed subjects recruited between December 2001 and June 2006. Recruitment of unexposed subjects began in January 2003: mean (and median) follow-up was 2.5 years for the exposed (interquartile range: 1.7–3.3) and 1.5 years for the unexposed (interquartile range: 0.9–2.1). A total of 622 physicians identified exposed subjects, while 112 of these physicians identified unexposed subjects.

Table 1 shows the distribution of baseline covariates among subjects treated with anti-TNF and those not treated with anti-TNF. Age, Disease Activity Score, and Health Assessment Questionnaire score had the greatest potential for confounding, as seen in the last column of Table 1, because of their large differences between exposed and unexposed subjects and strong association with mortality. Unexposed subjects were older and had more comorbidities, increasing their mortality risk. Conversely, they had less active disease, which would tend to decrease mortality. The net effect of these differences in baseline covariates was not obvious a priori.

Balancing of covariates

The effects of the different methods of balancing covariates on actual balance of covariates, as assessed by the change in estimated rate ratio after controlling for each variable

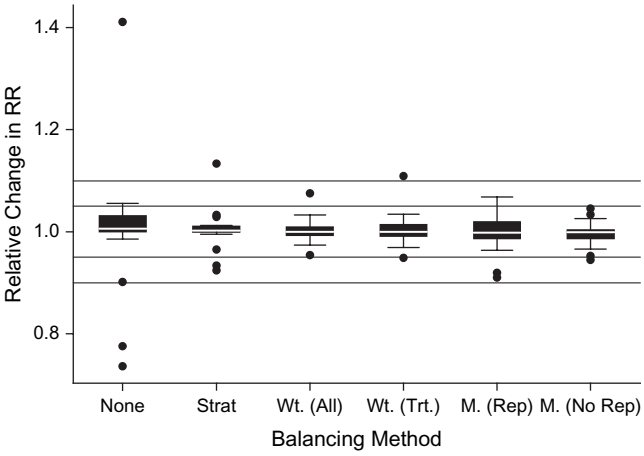

Figure 1. Effect of different methods of balancing covariates on actual balance of covariates between exposed and unexposed subjects. None, no balancing; Strat, stratification (5 strata); Wt. (All), weighting to the distribution of covariates in the entire sample; Wt. (Trt.), weighting to the distribution of covariates in the exposed subsample; M. (Rep), matching with replacement; M. (No Rep), matching without replacement; RR, rate ratio. Shaded boxes, interquartile range; white bars, median; whiskers, the most extreme observation less than 1.5 times the length of the shaded box beyond the shaded box; dots, individual observations beyond the whiskers; horizontal lines, relative change of ±5%, ±10%.

individually, are shown in Figure 1. Each box-and-whiskers plot shows the distribution of the changes in the rate ratio after adjusting for each of the 26 variables in Table 1 using one of the balancing methods: the first plot corresponds to the last column of Table 1. All methods demonstrated a marked improvement in balance, with the rate ratio changing by less than 5% for most variables with all methods and by less than 10% for all variables and all methods with the exception of age when stratifying. Matching without replacement seemed to perform best, followed by weighting to the overall distribution.

Table 2. Rate Ratio Estimates Using Different Methods of Balancing

Method	Mortality Rate Ratio	95% Confidence Interval
Unadjusted	1.21	0.82, 1.79
5 Strata	1.33	0.79, 2.25
Weighted (to the whole sample)	0.47	0.21, 1.06
Weighted (to the exposed subsample)	0.43	0.19, 1.01
Matched (with replacement)[a]	0.39	0.16, 0.91
Matched (without replacement)[b]	1.24	0.69, 2.25

[a] 8,437 exposed subjects were matched to 682 unexposed subjects.

[b] 712 exposed subjects were matched to 712 unexposed subjects.

Rate ratio estimates

The rate ratio estimates produced by the different balancing methods are given in Table 2. There were considerable differences between the different estimates. These differences are unlikely to be due to residual confounding; matching without replacement and weighting to the entire population were the most effective in removing imbalance in the covariates and yet produced very different estimates of the rate ratio.

Quintile-specific rate ratios

One possible explanation for these differences in estimates would be differences in effect depending on one or more covariates. To explore this possibility, we calculated the rate ratios in each quintile of propensity score separately (Table 3). We found large differences between quintiles of the rate ratio when all methods were used, but, within quintiles, the estimates were broadly similar across the different methods (except for matching without replacement, which used only a small proportion of the data in the upper quintiles and hence had very wide confidence intervals for these estimates). For all methods, the relative risk for exposure to biologics was high for subjects with a low propensity to receive them and was much lower for subjects with a high propensity to receive them.

Quintile-specific rates

Table 4 shows the number of person-years of follow-up, the number of deaths, and the mortality rate for each quintile. The mortality rates were similar across quintiles for the exposed subjects but increased markedly over the quintiles for the unexposed subjects. The relative risk of mortality for the exposed compared with the unexposed decreased as the propensity to be treated increased. It ranged from almost 2 for those least likely to receive treatment to 0.2 for those most likely to receive treatment, a trend largely due to increasing mortality among the unexposed subjects.

Quintile-specific weights

We have shown that the quintile-specific rate ratios were similar between methods but that the rate ratios varied between quintiles from 0.2 to 2 and that the overall rate ratios differed. Table 5 shows the relative weight given to each quintile in the calculation of the rate ratio by each method. The relative weight is the sum of the weights given to each subject in the quintile, divided by the sum of the weights given to all subjects in that treatment group. Stratifying and matching without replacement gave considerable weight to quintile 1, where the rate ratio was highest; as a consequence, these methods produced the highest overall rate ratios. The other methods gave equal weight to all quintiles (weighting to the entire sample) or less weight to quintile 1, producing smaller rate ratios, that is, less risk associated with treatment compared with nontreatment.

The expected rate ratios were similar to the observed rate ratios with each method. This observation implies that the differences in weighting within strata are not having a major effect because the expected rate ratios assume equal weights for all subjects within a stratum.

Table 3. Propensity Quintile-Specific Rate Ratios

Quintile[a]	5 Strata		Weighted to the Entire Sample		Weighted to the Exposed Subsample		Matched With Replacement		Matched Without Replacement	
	Mortality Rate Ratio	95% Confidence Interval	Mortality Rate Ratio	95% Confidence Interval	Mortality Rate Ratio	95% Confidence Interval	Mortality Rate Ratio	95% Confidence Interval	Mortality Rate Ratio	95% Confidence Interval
1	2.18	1.21, 3.93	1.56	0.74, 3.27	1.77	0.75, 4.19	2.04	0.72, 5.81	2.23	0.89, 5.62
2	0.71	0.28, 1.78	0.72	0.28, 1.85	0.73	0.28, 1.87	1.27	0.47, 3.38	0.64	0.18, 2.20
3	0.85	0.21, 3.48	0.90	0.21, 3.80	0.90	0.21, 3.83	1.20	0.28, 5.11	1.93	0.39, 9.54
4	0.41	0.10, 1.67	0.43	0.10, 1.78	0.43	0.10, 1.78	0.21	0.05, 0.94	0.33	0.03, 3.75
5	0.24	0.06, 0.90	0.23	0.06, 0.91	0.23	0.06, 0.91	0.19	0.05, 0.76	0.20	0.02, 2.18
P value for differences		0.01		0.13		0.12		0.02		0.18
P value for trend		0.01		0.07		0.09		0.01		0.05

[a] Quintile 1 is least likely to receive treatment; quintile 5 is most likely to receive treatment.

Distribution of the linear predictor among the exposed and unexposed

The distribution of the linear predictor of the propensity score among exposed and unexposed subjects is shown in Figure 2. Although there was a clear difference between the exposed and unexposed, very few exposed subjects had scores higher than the highest score among the unexposed, and few unexposed subjects had scores lower than the lowest score among the exposed. Thus, at least one untreated subject was similar to most treated subjects, although there were far more treated than untreated at the higher ranges.

Figure 3 shows the distribution of the linear predictor of the propensity score among the exposed and unexposed subjects after matching, both with and without replacement. When matching without replacement was used, the distribution was effectively that of the overlap between the exposed and unexposed subjects shown in Figure 2. This distribution did not differ much between exposed and unexposed subjects, so there should be no confounding by the variables included in the propensity score when matching without replacement is used.

When matching with replacement was used, the distributions were again the same among the exposed and unexposed subjects, but this time the distribution was that of the exposed subjects shown in Figure 2. Thus, the distribution of confounders was quite different from that achieved by matching without replacement, with more weight in the higher propensity scores, as seen in Table 5. If the effect of treatment differs with the propensity score (or with the potential confounders), these 2 methods will both eliminate confounding but will result in very different estimates of the overall rate ratio.

DISCUSSION

If the effect of treatment varies with the propensity score, different approaches to estimating it will give different results. This phenomenon has been observed by Kurth et al. (13) and was commented on by Stürmer et al. (14). The reason is that the different methods are effectively estimating the effect in different populations, with different distributions of covariates. So, which is the best estimate to use?

In fact, every estimate has the drawback that it applies to only one particular distribution of covariates. A population with a different distribution of covariates will have a different treatment effect. This property implies that no single estimate could be guaranteed to apply to any subsequent population. In the presence of effect modification, however, the same would be true of the effect estimate from a randomized trial: if the treatment were subsequently used in a population with a different distribution of covariates, the expected treatment effect would differ from that seen in the trial.

Our analysis shows how a specific treatment effect varies across propensity quintiles. In practice, it would be important to show how the treatment effect differed according to the variables that made up the propensity score: doing so would enable clinicians to judge the likely effect of that treatment in a given patient.

Table 4. Quintile-Specific Mortality Rates

Quintile[a]	Exposed			Unexposed			Rate Ratio
	Deaths	Person-Years	Rate	Deaths	Person-Years	Rate	
1	36	1,819	19.8	16	1,762	9.1	2.18
2	52	4,168	12.5	5	285	17.6	0.71
3	57	4,704	12.1	2	141	14.2	0.85
4	66	5,080	13.0	2	63	31.9	0.41
5	93	5,467	17.0	2	28	71.5	0.24
Overall	304	21,237	14.3	27	2,279	11.8	1.21

[a] Quintile 1 is least likely to receive treatment; quintile 5 is most likely to receive treatment.

There are drawbacks to presenting an overall effect of treatment when the effect in fact varies between different types of individuals. Still, each overall effect estimate can be interpreted as the effect of treatment on a particular population. Both the standardized mortality ratio–weighted estimator and that derived from matching with replacement measure the average effect of treatment in those who are treated, that is

$$\frac{\text{Observed mortality in exposed subjects}}{\text{Expected mortality in exposed subjects had they not been exposed}}.$$

This causal parameter may be of interest as a measure of how the exposure affected the subjects who were exposed.

The inverse probability of treatment–weighting estimator measures

$$\frac{\text{Expected mortality in all subjects if they are exposed}}{\text{Expected mortality in all subjects if they are not exposed}},$$

which may again be of some interest in a particular population.

In contrast, the matching without replacement estimator gives a measure of the average effect of treatment in those exposed subjects for whom a match can be found, which is a subpopulation of the exposed unlikely to be of any interest. Equally, the weights used in the maximum likelihood stratified estimator are chosen to maximize the precision of the estimate, assuming that the effect is constant across strata, and there is no reason why the average effect of treatment in this particular population would be of interest. An alternative would be to standardize the stratum-specific estimates to a specified distribution of propensity scores. For example, one could calculate a standardized mortality ratio from the stratum-specific estimates (14).

All of the estimates have a valid interpretation as the overall benefit of treatment in a given population. However, they are valid only if the assumptions of the propensity score methodology are satisfied, that is, if there are no unmeasured confounders, the propensity score model is correctly specified, the study size is sufficiently large to make the asymptotically unbiased estimator in fact unbiased, and the standard errors are reliable.

The increasing mortality rate among the unexposed subjects is extremely unusual. Although it is theoretically possible that increasing disease severity is associated with increased mortality in the unexposed but not in the exposed, a more likely explanation of this almost-10-fold increase in rate is unmeasured confounding in the upper quintiles. Unexposed subjects in the top propensity quintiles were very good candidates to receive anti-TNF-α, but they did not. This situation may have arisen because they were too frail or ill to receive treatment or had a contraindication, and this condition was not fully captured by the comorbidity and Health Assessment Questionnaire measures. It may, therefore, be that the effect of treatment does not vary across the quintiles as much as Table 3 suggests and that the benefit of anti-TNF-α in the upper quintiles is overestimated.

The problem was exacerbated in this instance because of the limited amount of data on unexposed subjects with high propensity scores. Only 2 deaths among the unexposed occurred in each of the 3 highest propensity quintiles. Therefore, few excess deaths occurred in these quintiles compared with the expected number based on the rates among the exposed subjects. It would take only a small number of subjects who were excluded from the exposed group owing to their ill health, despite being good candidates for treatment, and who subsequently died, to explain all of the excess deaths among the unexposed subjects.

Although, in this instance, the most likely explanation for the differences between quintiles in the apparent effect of treatment is unmeasured confounding, the possibility of a genuine effect-measure modification should be considered. In many cases, it is reasonable to assume that treatment is most beneficial for those subjects most suitable for treatment, and this phenomenon may cause different treatment effects across the range of the propensity score. This analysis supports the suggestion of Glynn et al. (15) that the interaction between treatment effect and propensity quintile should routinely be examined to evaluate effect-measure modification.

Balance between exposed and unexposed subjects in propensity models is commonly assessed by using the standardized difference (16). This statistic, however, measures the association between only the covariate and the exposure: a large difference in a variable weakly associated with the outcome may cause less bias than a smaller difference in a variable strongly associated with the outcome. So, a method of assessing balance that depends on the strength of the associations of each variable with both the exposure and the outcome is preferable.

In observational studies, it is common practice to identify a confounding variable by examining the extent to which the

Table 5. Quintile-Specific Weights

Quintile[a]	No Balancing		5 Strata		Weighted to the Entire Sample		Weighted to the Exposed Subsample		Matched With Replacement		Matched Without Replacement	
	Exposed, %	Unexposed, %	Exposed, %	Unexposed, %	Exposed, %	Unexposed, %	Exposed, %	Unexposed, %	Exposed, %	Unexposed, %	Exposed, %	Unexposed, %
1	8.6	77.3	46.4	46.4	18.3	20.0	8.6	10.1	8.6	10.9	48.8	53.2
2	19.6	12.5	23.8	23.8	18.7	17.8	19.6	18.7	19.6	20.6	27.6	25.8
3	22.1	6.2	9.8	9.8	19.8	21.0	22.1	23.5	22.1	24.1	13.5	12.8
4	23.9	2.8	9.9	9.9	20.9	19.5	23.9	22.4	23.9	27.1	5.8	5.7
5	25.7	1.2	10.0	10.0	22.2	21.7	25.7	25.2	25.7	17.2	4.2	2.5
	Expected	Observed	Expected	Observed	Expected	Observed	Expected	Observed	Expected	Observed	Expected	Observed
Rate ratio		1.21		1.33	0.50	0.47	0.44	0.43	0.49	0.39	1.10	1.24

[a] Quintile 1 is least likely to receive treatment; quintile 5 is most likely to receive treatment.

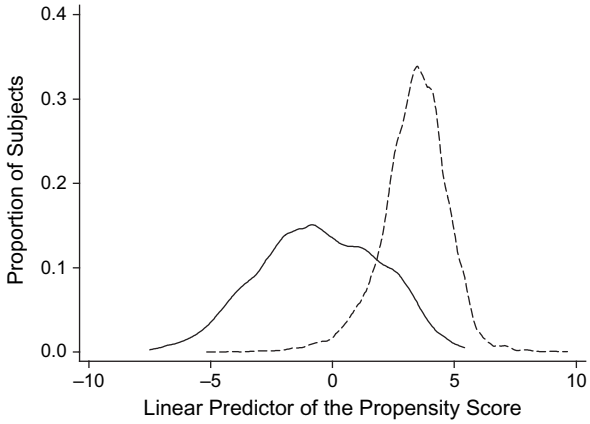

Figure 2. Distribution of the linear predictor of the propensity score among exposed and unexposed subjects, without adjustment. ———, untreated; – – – , treated.

effect estimate changes after adjusting for that variable (17). We used this method to assess whether the baseline covariates continued to confound the estimate of the treatment effect after balancing by using one of the propensity score methods: we are not aware of any previous use of this method to evaluate balance.

It should be pointed out that we are not relying on adjusting for a variable to remove confounding. We hope that the variable will not be a confounder in the balanced data and that the effect estimate will not change. If it does change, we have to conclude that the balancing has not succeeded and refine the propensity score until the effect estimate does not change upon adjustment. Hence, although simply fitting the variable as 5 categories could be criticized as being too crude to remove all confounding, it is sufficient for our purposes of showing the lack of balance in that variable.

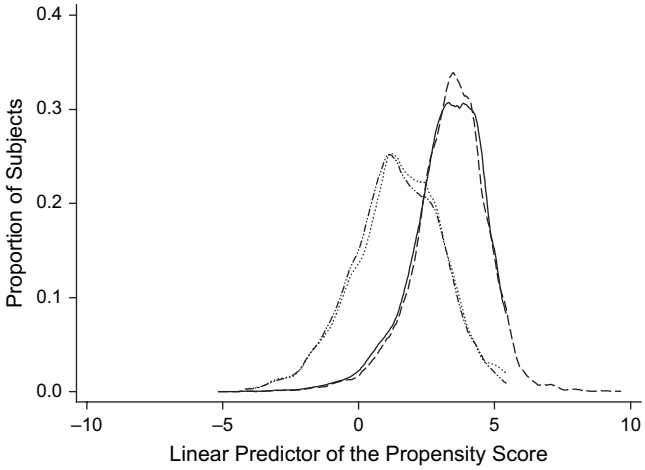

Figure 3. Distribution of the linear predictor of the propensity score among exposed and unexposed subjects, after adjustment. ———, untreated, matched with replacement; – – – –, treated, matched with replacement; – ·· – ··, untreated, matched without replacement; ·····, treated, matched without replacement.

To conclude, if the effect of treatment varies between individuals, different propensity-based methods of balancing covariates may give different answers in a given population. Each estimate may reflect a parameter of interest in that population. However, none of the estimates will reflect the effect of treatment in a different population. It is therefore essential when using propensity-based methods to test whether the treatment effect varies between individuals.

ACKNOWLEDGMENTS

Author affiliations: arc Epidemiology Unit, University of Manchester, Manchester, United Kingdom (Mark Lunt, Kimme Hyrich, Deborah P. M. Symmons); and Division of Pharmocoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Boston, Massachussetts (Mark Lunt, Daniel H. Solomon, Kenneth Rothman, Robert Glynn, Til Stürmer).

This work was partially funded by the Arthritis Research Campaign.

For more information about the British Society for Rheumatology Biologics Register, refer to the following website: http://www.medicine.manchester.ac.uk/epidemiology/research/arc/inflammatorymusculoskeletal/pharmacoepidemiology/bsrbr/.

The members of the British Society for Rheumatology Biologics Register Control Centre Consortium: Musgrave Park Hospital, Belfast (Dr. Allister Taggart); Cannock Chase Hospital, Cannock Chase (Dr. Tom Price); Christchurch Hospital, Christchurch (Dr. Neil Hopkinson); Derbyshire Royal Infirmary, Derby (Dr. Sheila O'Reilly); Russells Hall Hospital, Dudley (Dr. George Kitas); Gartnavel General Hospital, Glasgow (Dr. Duncan Porter); Glasgow Royal Infirmary, Glasgow (Dr. Hilary Capell); Leeds General Infirmary, Leeds (Prof. Paul Emery); King's College Hospital, London (Dr. Ernest Choy); Macclesfield District General Hospital, Macclesfield (Prof. Deborah Symmons); Manchester Royal Infirmary, Manchester (Dr. Ian Bruce); Freeman Hospital, Newcastle-upon-Tyne (Dr. Ian Griffiths); Norfolk and Norwich University Hospital, Norwich (Prof. David Scott); Poole General Hospital, Poole (Dr. Paul Thompson); Queen Alexandra Hospital, Portsmouth (Dr. Fiona McCrae); Hope Hospital, Salford (Dr. Romela Benitha); Selly Oak Hospital, Selly Oak (Dr. Ronald Jubb); St. Helens Hospital, St. Helens (Dr. Rikki Abernethy); Haywood Hospital, Stoke-on-Trent (Dr. Andy Hassell); and Kings Mill Centre, Sutton-In Ashfield (Dr. David Walsh).

Conflict of interest: none declared.

REFERENCES

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
2. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc.* 1984;79(387):516–524.
3. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat*. 1985;39(1):33–38.
4. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11(5):550–560.
5. Griffiths I, Silman AJ, Symmons DPM, et al. BSR Biologics Registry. *Rheumatology (Oxford)*. 2004;43(12):1463–1464.
6. Prevoo ML, van't Hof MA, Kuper HH, et al. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. *Arthritis Rheum*. 1995;38(1):44–48.
7. Ledingham J, Deighton C. Update on the British Society for Rheumatology guidelines for prescribing TNF-α blockers in adults with rheumatoid arthritis (update of previous guidelines of April 2001). *Rheumatology (Oxford)*. 2005;44(2):157–163.
8. National Institute for Clinical Excellence. Guidance on the use of etanercept and infliximab for the treatment of rheumatoid arthritis. London, United Kingdom: National Institute for Clinical Excellence; 2002. (Technical appraisal no. 36). (http://www.nice.org.uk/nicemedia/pdf/RA-PDF.pdf).
9. Kirwan JR, Reeback JS. Stanford Health Assessment Questionnaire modified to assess disability in British patients with rheumatoid arthritis. *Br J Rheumatol*. 1986;25(2):206–209.
10. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 1968;24(2):295–313.
11. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology.* 2003;14(6):680–686.
12. Stata Corporation. Stata statistical software, release 9.2. College Station, TX: Stata Corporation, 2006.
13. 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(3):262–270.
14. Stürmer T, Rothman KJ, Glynn RJ. Insights into different results from different causal contrasts in the presence of effect-measure modification. *Pharmacoepidemiol Drug Saf*. 2006; 15(10):698–709.
15. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol*. 2006;98(3):253–259.
16. Baser O. Too much ado about propensity score models? Comparing methods of propensity score matching. *Value Health*. 2006;9(6):377–385.
17. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. *Am J Epidemiol*. 1993;138(11):923–936.