American Journal of Epidemiology
Copyright © 2003 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved
Vol. 158, No. 3
Printed in U.S.A.
DOI: 10.1093/aje/kwg115

Comparison of Logistic Regression versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders

M. Soledad Cepeda[1,2], Ray Boston[3], John T. Farrar[2], and Brian L. Strom[2]

[1] School of Medicine, Javeriana University, Bogota, Colombia.
[2] Center for Clinical Epidemiology and Biostatistics and Department of Biostatistics and Epidemiology, School of Medicine, University of Pennsylvania, Philadelphia, PA.
[3] New Bolton Center, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA.

Received for publication March 26, 2002; accepted for publication January 16, 2003.

The aim of this study was to use Monte Carlo simulations to compare logistic regression with propensity scores in terms of bias, precision, empirical coverage probability, empirical power, and robustness when the number of events is low relative to the number of confounders. The authors simulated a cohort study and performed 252,480 trials. In the logistic regression, the bias decreased as the number of events per confounder increased. In the propensity score, the bias decreased as the strength of the association of the exposure with the outcome increased. Propensity scores produced estimates that were less biased, more robust, and more precise than the logistic regression estimates when there were seven or fewer events per confounder. The logistic regression empirical coverage probability increased as the number of events per confounder increased. The propensity score empirical coverage probability decreased after eight or more events per confounder. Overall, the propensity score exhibited more empirical power than logistic regression. Propensity scores are a good alternative to control for imbalances when there are seven or fewer events per confounder; however, empirical power could range from 35% to 60%. Logistic regression is the technique of choice when there are at least eight events per confounder.

bias (epidemiology); confounding factors (epidemiology); logistic models; models, statistical

In observational studies, the groups compared are often different because of lack of randomization. Subjects with specific characteristics may have been more likely to be exposed than other subjects. If these characteristics also affect the outcome, a direct comparison of the groups is likely to produce biased conclusions that may merely reflect the lack of initial comparability (1). These characteristics are called confounders.

Logistic regression is a commonly used method to control for imbalances between groups. Its primary advantage is the ability to control for many variables simultaneously. Although simultaneous adjustment is appealing, it can also be problematic. If too many variables need to be included in a model relative to the number of events, the estimates from these models can be incorrect (2, 3).

Another method to control for imbalances is the propensity score, which is the conditional probability of a subject's receiving a particular exposure given the set of confounders. For calculation of a propensity score, the confounders are used in a logistic regression to predict the *exposure* of interest, *without* including the outcome (4, 5). As a result, the collection of confounders is collapsed into a "single" variable, the probability (propensity) of being exposed.

The propensity score can be used as if it were the only confounder. When used as a stratifying variable, the propensity score should be divided into at least five strata. Within each stratum, the distribution of the confounders that went into its estimation should be similar in the exposed and the unexposed subjects (4–7).

The use of propensity scores is increasing (8, 9) and is appealing in the analysis of studies with rare events (outcomes) and multiple confounders (8). Creating a covariate that summarizes all the confounders could circumvent the problem of having too many variables in the model relative to the number of events. However, the utility of propensity scores has not been evaluated in this setting. The decision to use propensity scores should be based on their potential to reduce bias, their empirical coverage probability (the confidence interval of the result should include the true odds ratio), their empirical power (propensity scores should

Correspondence to Dr. Brian L. Strom, Center for Clinical Epidemiology and Biostatistics, School of Medicine, University of Pennsylvania, 824 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021 (e-mail: bstrom@cceb.med.upenn.edu).

Downloaded from http://aje.oxfordjournals.org/ by guest on June 28, 2012

detect an association if present), and their robustness (how sensitive the results are to errors in the estimation of the effect of the confounder on the outcome).

Propensity scores in a logistic model and the logistic regression estimate odds ratios (10). Propensity scores estimate the odds ratio given the propensity score categories, and logistic regression estimates the odds ratio given the confounders included in the model. These two odds ratios are often different from each other (10). This occurs because the average of the individual odds ratios is not the same as the total cohort odds ratio (11). This phenomenon does not occur with relative risks or risk differences (11, 12) and is the reason many researchers criticize the use of the odds ratio (11, 13). Still, the average of individual odds ratios and the total cohort odds ratio approximate each other when the incidence of the disease is low, all subjects have low risks (11), and both odds ratios are the same when there is no association between the exposure and the outcome (14). Therefore, to compare these techniques, we should replicate circumstances in which the odds ratio of each technique is close to the other.

The aim of this study was to compare logistic regression with the propensity score method in terms of bias, precision, empirical coverage probability, empirical power, and robustness when the numbers of events are low relative to the number of confounders.

MATERIALS AND METHODS

Data sets to be analyzed

Monte Carlo simulations were performed to simulate an observational cohort study in which the exposure of interest was based on the patients' characteristics, instead of randomly assigned (see the Appendix for the procedures used to generate the data).

Factors evaluated

Each factor of interest was varied systematically so that all possible combinations were evaluated.

Number of variables. We evaluated 2, 4, 6, 8, and 10 confounding variables in the model. Fifty percent of the confounders were continuous, normally distributed variables, and 50 percent were binary variables. Sixty percent increased the risk of developing the outcome, and 40 percent decreased the risk of developing the outcome. All the confounders were independent of each other. The strength of the association of the continuous confounders with the exposure and with the outcome ranged from an odds ratio of 0.97 to an odds ratio of 1.08. The strength of the association of the binary confounders with the exposure and with the outcome ranged from an odds ratio of 0.1 to an odds ratio of 4.0.

Number of events. A binary outcome was simulated, for example, dead or alive. We simulated 20, 50, 70, and 100 events. To accomplish this, we varied the sample size from 100 to 10,000 and the probability of the events from 1 percent to 20 percent. The range of the expected number of events permitted us to evaluate ratios of the number of events to the number of variables from the suboptimal rate of 2:1 (e.g., 20 events and 10 variables) to the rates advised in the literature of 10:1 (e.g., 100 events and 10 variables) and higher.

Strength of the exposure. A binary exposure was simulated, choosing odds ratios commonly found in epidemiology: 1.0, 1.5, 2.0, and 3.0.

Robustness. Researchers usually do not have access to the truth. Consequently, they can make mistakes when specifying the effect of a confounder on the outcome, and these mistakes could lead to wrong conclusions. The less sensitive a technique is to these errors, the better.

In the misspecified logistic regression model, the effect of one of the confounders on the outcome was assumed to be linear. In the correct model, this confounder had a quadratic effect. In the misspecified propensity score model, the effect of one of the confounders on the exposure was assumed to be linear. In the correct model, this confounder had a quadratic effect (15).

Analytical approaches

Once each data set was simulated, we analyzed it using the two approaches.

Logistic regression approach. To estimate the impact of the exposure on the outcome, we applied a logistic regression technique, using all the individual confounders and the exposure variable as independent variables and the clinical outcome as the dependent variable.

Propensity score approach. To estimate the propensity score, we used a logistic regression to obtain the predicted probability of exposure. In this case, the dependent variable was the exposure rather than the outcome, and the independent variables were the confounding variables. Note that the outcome variable was not used in this step. Once these probabilities (the propensity scores) were estimated, they were divided into five strata. The quintiles of the estimated propensity scores were used as the cutoffs for the different strata.

Then, to estimate the impact of the exposure on the outcome using the propensity score, we constructed a different logistic regression model. In this case, the dependent variable was the outcome, and the independent variables were the exposure variable and the categories of the propensity score.

Comparison of the two analytical approaches

In each Monte Carlo trial, we compared the results obtained with each of the two statistical techniques with the truth to determine the bias, precision, empirical coverage probability, empirical power, and robustness associated with the use of that technique.

To compare the techniques, we calculated the number of events per confounder considered in the model. It was obtained by dividing the number of observed outcomes by the number of confounders; the primary exposure under study was not included in the count. Then, it was divided into ten categories: 1–3, 4, 5, 6, 7, 8, 9, 10, 11–20, and more than 20 events per confounder. This categorization permits us to observe in detail situations in which there were few events

Downloaded from http://aje.oxfordjournals.org/ by guest on June 28, 2012

per confounder. The number of events per confounder was used to stratify estimates of bias, precision, empirical coverage probability, and empirical power.

In the case of logistic regression, the number of confounders included in the logistic model equals the number of confounders evaluated. In the case of the propensity score, it equals the number of confounders that were used to calculate the propensity score, not the number of terms in the propensity score model. Hence, the comparison of the propensity score and the logistic regression approaches involved the same number of confounders. However, because of the way these techniques deal with the confounders, at the end each approach had a different number of terms.

For the analyses, we used the coefficients obtained from the regression models instead of the odds ratios, because the distribution of the coefficients was less skewed than the distribution of the odds ratios. However, in this article, we refer to odds ratios since they are easier to interpret than coefficients.

We analyzed and reported separately the results of the comparisons between propensity score and logistic regression when there is and when there is no association of the exposure with the outcome.

Bias

Bias measures how different the estimated effect of the exposure is from the true effect. We expressed it as a percentage: Bias = ((estimated odds ratio/true odds ratio) – 1) × 100. We estimated the mean percentage of bias of the Monte Carlo trials and the standard deviation. We also report the median percentage of bias due to the skewness of the distribution of the estimated odds ratios. Values greater than zero indicate an overestimation of the effect of the exposure on the outcome, and negative values indicate an underestimation.

Empirical coverage probability

The confidence interval of the estimated odds ratio should include the true odds ratio. To estimate the empirical coverage probability, we tested whether the true odds ratio of 1, 1.5, 2, or 3 was included in the 95 percent confidence interval of the estimate of the association with each of the two analytical approaches (3, 16). In these circumstances, the expected value of the empirical coverage probability is 95 percent.

This way of evaluating empirical coverage probability is analogous to the Wald statistic, a test used to determine the statistical significance of a variable in a regression model (17).

Empirical power

Power is the probability of detecting an association when present. Empirical power is the power of the 95 percent confidence interval to reject a false null hypothesis (3, 16, 18). We estimated it by testing if an odds ratio of 1 was excluded in the 95 percent confidence interval of the estimate of the association with each of the two analytical approaches (16, 18). The higher the value of the empirical power is, the better. Because the empirical power depends on the magnitude of the difference being evaluated, we report the results in which the effect of the exposure on the outcome has an odds ratio of 3.

We did not evaluate empirical power when there was no association of the exposure with the outcome, because it is impossible to evaluate it in this particular setting.

Precision

Precision measures the degree of dispersion, the spread, of the estimates obtained with a particular technique. To measure it, we averaged the standard error of the odds ratios; the smaller the standard error is, the more precise a technique.

Number of Monte Carlo trials

To detect a difference in the odds ratio estimates of the two analytical techniques of 0.1 standard deviation with an alpha error of 0.05 and a beta error of 0.1, we estimated that 1,052 Monte Carlo trials were needed to evaluate each factor and its corresponding number of levels (19). We therefore performed a total of 252,480 Monte Carlo trials.

Monte Carlo simulations and analyses were performed with Stata version 7 SE statistical software (Stata Corporation, College Station, Texas).

RESULTS

Bias

In the logistic regression as the number of events per confounder increased, the magnitude of the bias decreased, from an underestimation or an overestimation of the effect of the exposure on the outcome of 16 times when there were four or fewer events per confounder to levels of bias close to zero when there were seven or more events per confounder (figure 1). In the logistic regression, the

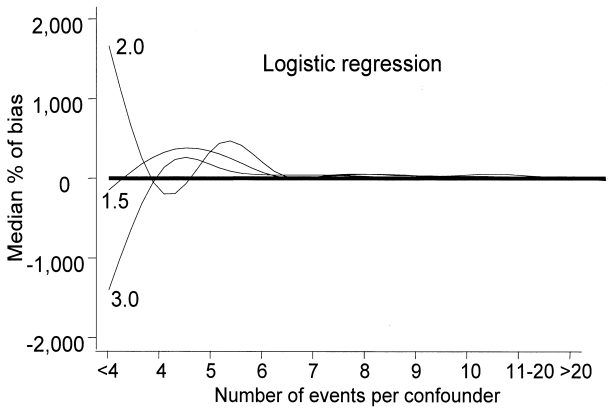

FIGURE 1. Median percentage of bias with the logistic regression, by strength of the exposure and number of events per confounder. In the logistic regression, the bias declines as the number of events per confounder increases. Values greater than zero indicate an overestimation of the effect of the exposure on the outcome. Negative values indicate an underestimation of the effect of the exposure on the outcome.

Downloaded from http://aje.oxfordjournals.org/ by guest on June 28, 2012

TABLE 1. Mean percentage of bias, by number of events per confounder, by technique, and by correct and misspecified model*

No. of events per confounder	Logistic regression		Propensity score	
	Correct model	Misspecified model	Correct model	Misspecified model
1–3	128.2 (1,205.7)†	1,064.8 (600.6)	81.4 (41.1)	81.4 (44.5)
4	290.1 (128.6)	435.4 (259.4)	82.8 (42.5)	81.3 (40.9)
5	184.7 (220.0)	404.0 (229.2)	71.4 (42.0)	70.0 (40.7)
6	72.1 (24.6)	104.0 (72.2)	66.5 (27.1)	67.6 (38.1)
7	54.4 (35.5)	76.1 (31.8)	59.0 (37.1)	60.3 (40.4)
8	34.2 (13.4)	78.0 (31.8)	63.6 (43.4)	64.8 (44.0)
9	23.5 (4.6)	43.5 (16.34)	74.7 (47.5)	78.1 (50.2)
10	27.6 (15.1)	48.2 (14.6)	69.4 (42.2)	71.9 (44.2)
11–20	4.2 (5.7)	22.0 (9.2)	77.3 (50.9)	78.8 (41.6)
>20	–4 .4 (5.0)	–20.2 (9.3)	81.7 (42.9)	82.8 (42.0)

* Percentage of bias = ((estimated odds ratio/true odds ratio) – 1) × 100. Values greater than zero indicate an overestimation of the effect of the exposure on the outcome. Negative values indicate an underestimation of the effect of the exposure on the outcome. Results are for odds ratios of 1.5, 2.0, and 3.0 for the exposure.

† Numbers in parentheses, standard deviation.

strength of the association of exposure with the outcome did not have a clear effect on the amount of bias (figure 1). To the contrary, in the propensity score the magnitude of the bias did not depend on the number of events per confounder variable (table 1), but it depended on the strength of the association of the exposure with the outcome (figure 2). The bias was less when the strength of the association increased; for example, the bias when the odds ratio was 3 was much lower (8 percent median of bias) than when the odds ratio was 1.5 (60 percent median of bias). When the odds ratio for the exposure was 3 and there were 10 or fewer events per confounder, propensity score estimates were less biased than the logistic regression. However, when the odds ratio for the exposure was 1.5 or 2 and there were eight or more events per confounder, logistic regression produced less biased estimates than the propensity score.

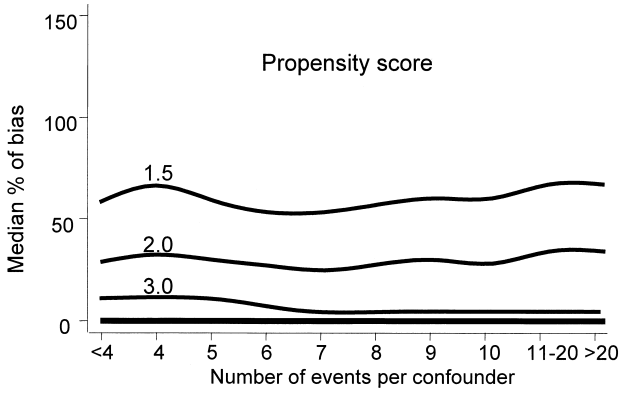

FIGURE 2. Median percentage of bias with the propensity score, by strength of the exposure and number of events per confounder. In the propensity score, the bias decreases as the strength of the association of the exposure with the outcome increases. Values greater than zero indicate an overestimation of the effect of the exposure on the outcome.

Robustness

Errors in the specification of the model led to an increase in the magnitude of the bias when the logistic regression was used. However, there was no increase in the magnitude of the bias when propensity scores were used (table 1).

Precision

In the logistic regression, the precision increased with the number of events per confounder in the model. The difference between the techniques was marked below six events per confounder. In these circumstances, the mean of the standard error was in the thousands in the logistic regression but only four in the propensity score. Logistic regression reached levels of precision similar to those of the propensity score when there were eight or more events per confounder in the model (figure 3).

Empirical coverage probability

The empirical coverage probability depended on the number of events per confounder in both techniques, but the pattern was different. In the logistic regression, the empirical coverage probability increased as the number of events per confounder increased (from 80 percent when there were 1–3 events per confounder to the expected value of 95 percent when there were eight or more events per confounder). In the propensity score, it was the opposite: the empirical coverage probability decreased from the expected value of 95 percent when there were seven or fewer events per confounder to 80

Downloaded from http://aje.oxfordjournals.org/ by guest on June 28, 2012

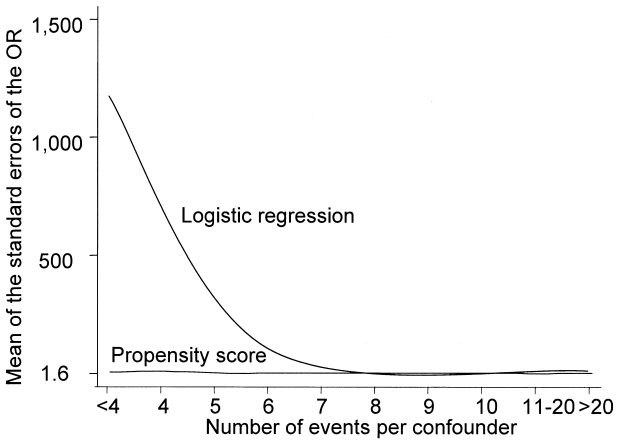

FIGURE 3. Mean of the standard errors of the odds ratio for the exposure, by number of events per confounder and by technique. The estimates of the propensity score are more precise (the standard errors are much smaller) than the estimates from logistic regression. As the number of events per confounder increases, the precision of the logistic regression increases. OR, odds ratio.

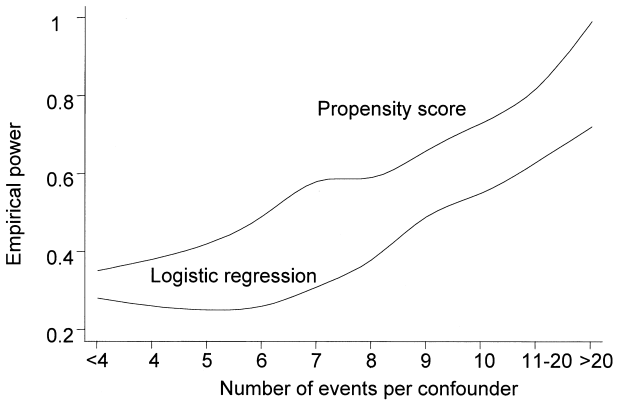

FIGURE 5. The empirical power, by number of events per confounder and by technique. This graph illustrates the empirical power in both techniques. The propensity score exhibits more empirical power than the logistic regression. The empirical power increases in both techniques as the number of events per confounder increases. We are evaluating an odds ratio of 3. The higher the value is, the better, that is, the more power a technique has.

percent when there were 21 or more events per confounder (figure 4).

Empirical power

The empirical power was higher with the propensity score than with the logistic regression. As the number of events per confounder increased, empirical power increased in both techniques, but the increase was more substantial in the propensity score. Propensity score empirical power reached almost 100 percent when there were 21 or more events per confounder (figure 5).

No association between the exposure and the outcome

The results for bias and empirical coverage probability when there was no association between the exposure and the outcome were similar to the ones obtained when there was an association between the exposure and the outcome (figures 6 and 7).

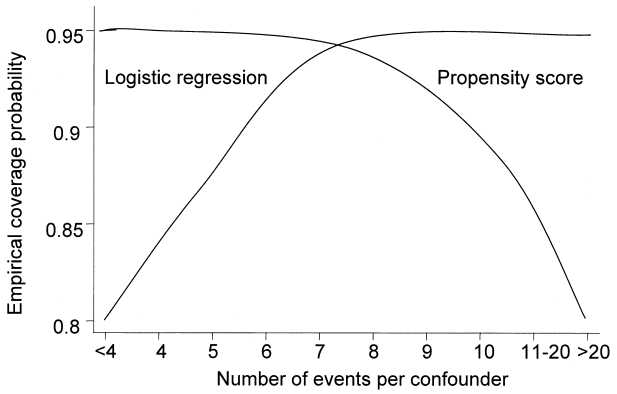

FIGURE 4. Empirical coverage probability, by number of events per confounder and by technique. This graph illustrates the empirical coverage probability in both techniques. The expected value of the empirical coverage probability is 0.95.

DISCUSSION

We found that propensity score estimates were less biased than the logistic regression estimates when there were six or fewer events per confounder. The amount of bias decreased as the strength of the association of the outcome with the exposure increased. Overall, the propensity score was more robust and more precise, and it had more power than logistic regression. The empirical coverage probability associated

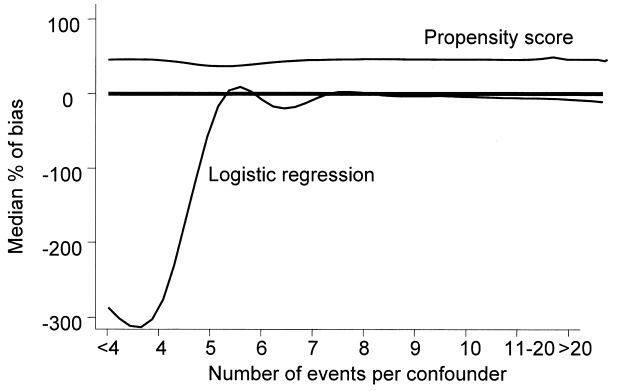

FIGURE 6. Median percentage of bias when there was no association between the exposure and the outcome, by number of events per confounder and by technique. In the logistic regression, the bias declines as the number of events per confounder increases. In the propensity score, the amount of bias did not depend on the number of events per confounder. Values greater than zero indicate an overestimation of the effect of the exposure on the outcome. Negative values indicate an underestimation of the effect of the exposure on the outcome.

Downloaded from http://aje.oxfordjournals.org/ by guest on June 28, 2012

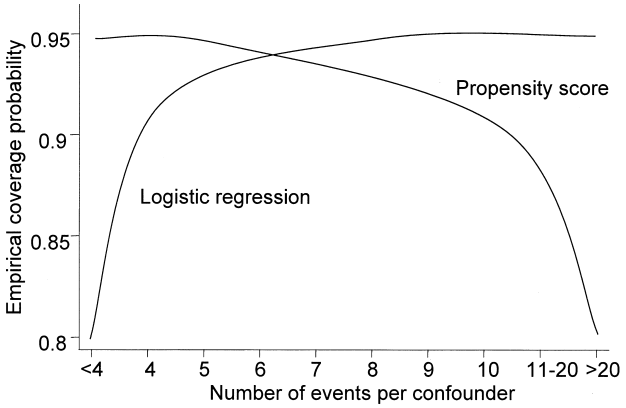

FIGURE 7. Empirical coverage probability when there was no association between the exposure and the outcome, by number of events per confounder and by technique. This graph illustrates the empirical coverage probability in both techniques. The expected value of the empirical coverage probability is 0.95.

with the propensity score was adequate until there were seven or less events per confounder, but it decreased substantially afterward.

Lack of bias and precision are desirable properties of a statistical technique. An unbiased estimate tells us that "on average" the estimate will be accurate (20). However, researchers usually have only one sample available, so besides lack of bias, the degree of precision of the estimates is important (20). The propensity score produced more precise estimates than the logistic regression, but when there were seven or more events per confounder the magnitude of the bias was larger than with logistic regression. The precision of logistic regression increased dramatically to that of the propensity score once there were eight or more events per confounder. Therefore, in terms of the tradeoff between bias and precision, once there are eight events per confounder, logistic regression is a better approach.

The strength of the association did have an effect on the amount of bias in the propensity score: the stronger the association, the smaller the magnitude of the bias. Others have reported similar findings (15). In contrast to the logistic regression approach, the number of events per confounder did not affect the magnitude of the bias in the propensity score. This finding could be explained by the fact that the propensity score always has the same number of terms independent of the number of confounders (the exposure plus the propensity categories).

We found that the propensity score was a robust technique; the magnitude of bias was similar when the model was correct or misspecified. Others have reported a similar result (15). The robustness of the propensity score makes it an attractive technique.

We found that the propensity score had the expected empirical coverage probability when there were fewer than eight events per confounder, but after this point it decreased. This means that the true test hypothesis is being incorrectly rejected, that a type I error is being made (21, 22). Others have described a high incidence of type I error (15 percent) with the propensity score when the exposure and the confounders were highly correlated (23). The empirical coverage probability in the logistic regression approach was inadequate until there were eight events per confounder. Others have found similar findings (3).

As expected, the empirical power increased in both techniques as the number of events per confounder increased. Overall, the propensity score had more empirical power.

The results of this study confirm that a low number of events per variable in a logistic regression analysis leads to biased results and estimates that are not precise. It has been suggested that one needs to have at least 10 events per variable included in a logistic model (3, 24, 25). However, the numbers of events per variable between five and 10 were not evaluated. We found that eight events per confounder in the model produced estimates that were similar to the results obtained when there were 21 or more events per confounder in terms of bias, degree of precision, and empirical coverage probability. The empirical power also increased substantially at this cutoff point. Therefore, eight events per confounder are probably a good cutoff for logistic regression.

This study has some limitations. For the calculation of empirical coverage probability and power we used the Wald statistic, but this test performs suboptimally in small sample sizes. In these circumstances, the likelihood ratio test would be likely to perform better (26). Nonetheless, we believe that our findings remain valid because both techniques were compared using the same test.

The way we carried out the simulations could have given logistic regression an advantage over propensity scores. The logistic regression model we used to analyze the data was in many scenarios the theoretically correct model (the model used to generate the data). Nonetheless, we introduced a random component when we generated the data and, in addition, we evaluated circumstances in which the logistic regression model used to analyze the data was not the theoretically correct model (when we misspecified the regression model). In these last circumstances, the magnitude of the bias with logistic regression increased.

In addition, in the simulations we did not check if the distributions of the confounders in the exposed and unexposed groups in each stratum of the propensity score were similar. This is a key step when using propensity scores in clinical research (5, 9, 27). However, this exercise is not possible during a simulation, because it requires a visual check of the distribution of each one of the confounders in the exposed and unexposed groups within each propensity score category. Consequently, it could be argued that propensity scores performed poorly because they did not reach the anticipated balance of the confounders. However, in our simulation study, the estimation of the propensity score was based on a correct model in many scenarios, and theory suggests that in these circumstances the propensity score balances the distribution of the confounders in the groups compared (4, 5, 8, 28–30).

While in theory propensity scores should balance the distribution of the confounders, in our study we observed that the propensity score method produced biased estimates even when there was no association between the exposure and the outcome. The discrepancies between theory and practice could be explained by residual bias due to the categorization of the propensity score. Although the use of five

Downloaded from http://aje.oxfordjournals.org/ by guest on June 28, 2012

subclasses generally removes 90 percent of the bias due to the subclassifying variable (6), this is a large-sample claim that assumes a normal distribution of the confounding variable and requires an adequate overlap of the distribution of the confounder in the groups compared within the five categories (6, 27). In practice, these conditions may not be fully met, which could lead to bias of greater magnitude than predicted by theory.

To avoid this problem, we could have included the propensity score as a continuous variable in the regression model. However, this would assume that the effect of the propensity scores on the outcome is linear. To address this, one can use more sophisticated methods such as including the propensity score in a semiparametric regression model (31). We did not evaluate these scenarios because they are not the traditional way to use propensity scores.

As discussed earlier, the use of a propensity score in a logistic model and of logistic regression itself estimates different odds ratios. However, we simulated circumstances in which the prevalence of events was low and all subjects had a low risk of developing the event. Therefore, the odds ratios produced by the two techniques should approximate each other (10) sufficiently closely to make the comparisons valid. Nonetheless, to confirm that the results were not an artifact due to the estimation of different odds ratios, we performed the same comparisons when there was no association of the exposure with the outcome; in this circumstance, the estimates produced by the two techniques should be equal to each other (10). The results were similar to those obtained in the other simulations, which substantiates the validity of our results.

In summary, the propensity score is a good alternative to control for imbalances between study groups when there are seven or fewer events per confounder variable. In these circumstances, analyses using propensity scores are more precise and more robust than the logistic regression estimates, the magnitude of bias is similar to the magnitude of bias obtained when the propensity score is used and there are plenty of events per variable, and the empirical coverage probability is adequate. Nonetheless, the empirical power could range from 35 percent to 60 percent. Logistic regression is the technique of choice when there are at least eight events per confounder. In these circumstances, analyses using logistic regression are precise and less biased than the propensity score estimates, and the empirical coverage probability and empirical power are adequate.

ACKNOWLEDGMENTS

The first author had a scholarship from the Colombian government.

The authors would like to thank Drs. Mona Baumgarten, Daniel B. Carr, and Paul R. Rosenbaum for their invaluable help in the design of this study. They would also like to express their appreciation to Drs. Jeffrey Carson, Scott Halpern, Sean Hennessy, David Margolis, and Daniel Mines for their input in previous drafts.

REFERENCES

1. Rosenbaum PR. Observational studies. In: Rosenbaum PR, ed. Observational studies. New York, NY: Springer-Verlag, 1995: 1–12.
2. Harrell FEJ, Lee KL, Califf RM, et al. Regression modelling strategies for improved prognostic prediction. Stat Med 1984;3: 143–52.
3. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable on logistic regression analysis. J Clin Epidemiol 1996;49:1373–9.
4. Rosenbaum PR, Rubin DR. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
5. Rosenbaum PR, Rubin DB. Reducing bias in observational studies. J Am Stat Assoc 1984;79:516–24.
6. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 1968;24:295–313.
7. Connors AFJ, Speroff T, Dawson NV, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. JAMA 1996;276:889–97.
8. Joffe MM, Rosenbaum PR. Propensity scores. Am J Epidemiol 1999;150:327–33.
9. Cepeda MS. The use of propensity scores in pharmacoepidemiologic research. (Editorial). Pharmacoepidemiol Drug Saf 2000; 9:103–4.
10. Rosenbaum PR. Propensity score. In: Armitage P, Colton T, eds. Encyclopedia of biostatistics. Chichester, United Kingdom: Wiley, 1998:3551–5.
11. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. Am J Epidemiol 1987;125:761–8.
12. Miettinen OS, Cook EF. Confounding: essence and detection. Am J Epidemiol 1981;114:593–603.
13. Greenland S, Robins JM. Identifiability, exchangeability and epidemiological confounding. Int J Epidemiol 1986;15:413–19.
14. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika 1984;71:431–44.
15. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. Biometrics 1993;49:1231–6.
16. Concato J, Peduzzi P, Holford TR, et al. Importance of events per independent variable in proportional hazards analysis. J Clin Epidemiol 1995;48:1495–501.
17. Kleinbaum DG, Kupper LL, Muller KE, et al. The method of maximum likelihood. In: Kleinbaum DG, Kupper LL, Muller KE, et al, eds. Applied regression analysis and other multivariable methods. 3rd ed. Pacific Grove, CA: Duxbury, 1998:639–55.
18. Mooney CZ. Using Monte Carlo simulation in the social sciences. In: Mooney CZ, ed. Monte Carlo simulation. Thousand Oaks, CA: Sage Publications, 1997:65–91.
19. Guenther WC. Sample size formulas for normal theory *t* tests. Am Stat 1981;35:243–4.
20. Berry WD, Feldman S. The multiple regression model: a review. In: Berry WD, Feldman S, eds. Multiple regression in practice. Newbury Park, CA: Sage University Paper, 1985:9–17.
21. Rothman KJ, Greenland S. Approaches to statistical analysis. In: Rothman KJ, Greenland S, eds. Modern epidemiology. 2nd ed. Philadelphia, PA: Lippincott-Raven Publishers, 1998:183–99.
22. Casella G, Berger RL. Hypothesis testing. In: Casella G, Berger RL, eds. Statistical inference. Belmont, CA: Duxbury Press, 1990:345–402.
23. Cook EF, Goldman L. Performance of tests of significance based on stratification by a multivariate confounder score or by propensity score. J Clin Epidemiol 1989;42:317–24.
24. Bergstralh EJ, Kosanke JL, Jacobsen SJ. Software for optimal

Downloaded from http://aje.oxfordjournals.org/ by guest on June 28, 2012

matching in observational studies. Epidemiology 1996;7:331–2.
25. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Intern Med 1993;118:201–10.
26. Agresti A. Generalized linear models. In: Agresti A, ed. An introduction to categorical data analysis. New York, NY: Wiley-Inter Science, 1996:71–102.
27. Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997;127:757–63.
28. Rubin DR, Thomas N. Characterizing the effect of matching using linear propensity score methods with normal distributions. Biometrika 1992;79:797–809.
29. Drake C, Fisher L. Prognostic models and the propensity score. Int J Epidemiol 1995;24:183–7.
30. Cook EF, Goldman L. Asymmetric stratification. An outline for an efficient method for controlling confounding in cohort studies. Am J Epidemiol 1988;127:626–39.
31. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. Biometrics 1992;48:479–95.
32. Hilbe J, Linde-Zwirble W. Random number generators. sg44.1. Stata Tech Bull 1995;28:20–1.

APPENDIX

Models for Generating the Data

Because we designed a simulation that resembled an observational study in which the likelihood of being exposed depended on baseline characteristics that were also related to the outcome, we used a random-number generator program that was developed by Hilbe and Linde-Zwirble (32) for Stata software. This program uses predetermined probabilities to generate the sample. In this case, it used the probability of being exposed and the probability of developing the outcome to determine who was going to be exposed and who was going to develop the outcome.

To obtain the probabilities that the random generator program utilized, we used linear logistic models. These are the most popular models to generate binary data (26). The probabilities were obtained as follows. The probability of exposure (t) is:

$$\Pr(t = 1 | x1, x2, \ldots xn) = 1/[1 + \exp(-(b0 + b1x1 + (b2x1)^2 + b3x2 + \ldots + bnxn))].$$

The probability of developing the outcome (y) is:

$$\Pr(y = 1 | t, x1, x2, \ldots xn) = 1/[1 + \exp(-(\gamma 0 + \gamma 1t + \gamma 2x1 + (\gamma 3x1)^2 + \gamma 4x2 + \ldots + \gamma nxn))],$$

where $b0$, $\gamma 0$ are the intercept terms, $x = (x1, \ldots xn)$ is the set of confounding variables, and b, $\gamma = (b2, \ldots bn$ and $\gamma 2, \ldots \gamma n)$ is the set of values of the regression coefficients.

Downloaded from http://aje.oxfordjournals.org/ by guest on June 28, 2012