Performance of Disease Risk Scores, Propensity Scores, and Traditional Multivariable Outcome Regression in the Presence of Multiple Confounders

Patrick G. Arbogast* and Wayne A. Ray

* Correspondence to Dr. Patrick G. Arbogast, Department of Biostatistics, S-2323 Medical Center North, Vanderbilt University, Nashville, TN 37232-2158 (e-mail: patrick.arbogast@vanderbilt.edu).

Initially submitted August 19, 2010; accepted for publication April 8, 2011.

Propensity scores are widely used in cohort studies to improve performance of regression models when considering large numbers of covariates. Another type of summary score, the disease risk score (DRS), which estimates disease probability conditional on nonexposure, has also been suggested. However, little is known about how it compares with propensity scores. Monte Carlo simulations were conducted comparing regression models using the DRS and the propensity score with models that directly adjust for all of the individual covariates. The DRS was calculated in 2 ways: from the unexposed population and from the full cohort. Compared with traditional multivariable outcome regression models, all 3 summary scores had comparable performance for moderate correlation between exposure and covariates and, for strong correlation, the full-cohort DRS and propensity score had comparable performance. When traditional methods had model misspecification, propensity scores and the full-cohort DRS had superior performance. All 4 models were affected by the number of events per covariate, with propensity scores and traditional multivariable outcome regression least affected. These data suggest that, for cohort studies for which covariates are not highly correlated with exposure, the DRS, particularly that calculated from the full cohort, is a useful tool.

confounding factors (epidemiology); epidemiologic methods; observational study; pharmacoepidemiology

Abbreviations: DRS, disease risk score; NSAID, nonsteroidal antiinflammatory drug.

In contemporary cohort studies, there frequently are a large number of covariates that must be considered as potential confounders. The increasing use of automated databases, in which past medical encounters (e.g., a prescription for insulin) provide surrogates for important patient comorbidities (e.g., diabetes), has contributed to this trend. Indeed, it is not uncommon to have 100 or more covariates available.

The standard method to control for such confounding is to fit multiple regression models with terms for the exposure variable of primary interest and each of the potential confounders. However, this approach poses 4 important challenges for statistical analysis. The first is computational complexity, dependent upon the number of covariates and exposure categories. Consider a study of nonsteroidal antiinflammatory drugs (NSAIDs) with 100 covariates. If there are 9 individual drugs of interest (6 traditional NSAIDs and 3 newer selective inhibitors of cyclooxygenase-2), each with 3 dose categories, the exposure has 30 levels (9 × 3 levels of current use, 1 for recent use, 1 for former use, and 1 for nonuse). Even with modern computing capacity, models with 130 or more covariates will require substantial computing power and analyses may be cumbersome, particularly if there are dependencies (e.g., allowing an individual to enter the cohort multiple times) that require more complex variance estimation. This, in turn, may inhibit performance of important sensitivity analyses.

Second is the question of model specification. Because some of the covariates may not be confounders, variable selection procedures may be considered to improve exposure estimate precision and to reduce computational complexity. However, this may involve subjective decisions such as the type of variable selection procedure, whether to base selection upon $P$ values or change in exposure parameter estimates, and numeric cutoffs (e.g., $P = 0.05, 0.10, 0.20$).
for variable inclusion. For covariates that confer relatively modest increases in disease risk, some variable selection procedures, such as stepwise regression, may exclude important covariates from the final model. Furthermore, such techniques as stepwise regression have limitations that can lead to underestimation of standard errors for exposure estimates (1).

Third, if the number of outcome events under study is small, parameter estimates based upon large sample theory, such as those provided by widely used regression programs, may be inappropriate. Peduzzi et al. (2) conducted simulation studies evaluating events per variable in logistic regression; they observed that, for fewer than 10 events per variable, regression coefficients could be biased and standard errors could be incorrect (2). Harrell et al. (3) reported similar findings.

A final challenge is the question of effect modification. Investigators are nearly always interested in whether or not the effect of the exposure varies for patients according to baseline risk of the study outcome. However, with more than 100 variables that measure this factor, the effect modification analyses may be cumbersome.

An increasingly popular approach to these challenges is to calculate a single summary measure from the covariates and then to use this summary measure in subsequent statistical analysis. At present, the most common approach to reducing the dimensionality of models is propensity scores (4). For dichotomous exposure variables, this score is calculated as the probability that each subject is exposed, as a function of his/her observed covariates. Rosenbaum and Rubin (4) demonstrated that the propensity score is a balancing score, meaning that, conditional on the propensity score, the distribution of the covariates used to derive the propensity score is the same for exposed and unexposed persons. Cepeda et al. (5) conducted simulation studies comparing multiple logistic regression models adjusting for the individual covariates with propensity score analyses. They reported that the propensity score was a good alternative when there were 7 or fewer events per covariate; exposure estimates had less bias, were more robust, and were more precise. However, if there were at least 8 events per covariate, logistic regression was preferable; in the corresponding propensity score analyses, coverage probabilities decreased and exposure effect estimates had greater bias.

The disease risk score is an alternative approach to the propensity score (6, 7). Like the propensity score, it is a summary measure derived from the observed values of the covariates. However, the disease risk score estimates the probability or rate of disease occurrence as a function of the covariates. The disease risk score may be estimated in 2 ways. First, it can be calculated as a “full-cohort” disease risk score (DRS), which is the multivariate confounder score originally proposed by Miettinen (8) in 1976. This score was constructed from a regression model relating the study outcome to the exposure of interest and the covariates for the entire study population. The score was then computed as the fitted value from that regression model for each study subject, setting the exposure status to unexposed. The subjects were then grouped into strata according to the score, and a stratified estimate of the exposure effect was calculated. The disease risk score also may be estimated as an “unexposed-only” DRS, from a regression model fit only for the unexposed population, with the fitted values then computed for the entire cohort.

However, use of the disease risk score was inhibited by an early study of its statistical properties. Pike et al. (9) examined the full-cohort DRS theoretically and in simulation studies and demonstrated an exaggeration of the statistical significance of the exposure estimate. Cook and Goldman (10) reexamined the score and, in particular, revisited the simulation study conducted by Pike et al. Cook and Goldman agreed with the theoretical arguments of Pike et al. and expanded their simulation study to include a propensity score analysis. They observed an exaggeration of statistical significance, similar to that reported by Pike et al. However, this exaggeration was most pronounced when the squared multiple correlation coefficient between exposure and confounders exceeded 90%, which is unlikely to occur in practice. When the correlation between exposure and confounders was lower, there was little exaggeration of statistical significance. The propensity score analysis was more robust in the presence of large exposure-confounder correlations.

In recent years, interest in disease risk scores has increased. We have used these scores in several pharmaco-epidemiologic studies (11–17), motivated by the need to reduce computational complexity in large cohort studies with numerous covariates, as well as to assess effect modification. In 2005, Stürmer et al. (18) compared several approaches to controlling confounding in a “real world” example that assessed NSAIDs and 1-year all-cause mortality in 103,133 elderly Medicaid beneficiaries, using 71 covariates to control for confounding. They compared several propensity score analyses, the disease risk score, and traditional multiple regression models and observed no major difference in any of these methods. We recently reported simulation studies comparing regression models using the unexposed-only DRS with models adjusting for the individual covariates (6). These were designed to resemble a recently published study of NSAIDs and risk of acute myocardial infarction/sudden cardiac death, with the exception that we simulated cohort data instead of case-control samples (16). So long as there was not a high degree of intercorrelation between confounders and exposure, the unexposed-only DRS was a reasonable approach for summarizing risk factors from large cohort studies.

Hansen (19) recently examined the theoretical properties of the disease risk score (termed the prognostic score), demonstrating that it is a balancing score. Hansen noted some practical difficulties and additional assumptions required with use of the estimated prognostic score, particularly the possible need to estimate the score in the unexposed group and the difficulty of assessing prognostic balance between exposed and unexposed. He noted that, if the prognostic score is estimated in the full cohort, with exposure in the model, and the exposure increases the outcome, then the estimated prognostic score will tend to be a mixture of the true propensity and prognostic scores that can obscure the exposure effect. Hansen also noted that if a covariate is strongly associated with exposure, then applying a prognostic score to the same sample used to estimate it could result...
in an increased type I error rate that did not occur when estimating the prognostic score in a separate sample and applying it.

In this paper, we present simulation studies comparing the unexposed-only DRS, full-cohort DRS, and propensity score analyses with traditional regression models that include all of the individual covariates used to construct these summary scores. We sought to assess performance for 3 types of scenarios: 1) those typical of practical applications with an adequate number of events per confounder, 2) those in which traditional methods might be subject to the effects of model misspecification, and 3) those with small numbers of events per confounder.

MATERIALS AND METHODS

Scenarios and simulations

All simulations assume a binary exposure, a binary disease outcome, and a fixed period of follow-up with no censoring. The values of the covariates were first determined to achieve the expected prevalence specified for each scenario. Once the covariates were specified, the exposure status was simulated by using a logistic model to determine exposure probability, with the baseline exposure prevalence (all covariate values zero) specified according to that scenario. Once both the covariates and exposure status were determined, the disease outcome was simulated via a Poisson model, by use of the baseline disease rate per person (covariates and exposure zero) for that scenario. The reason for using a Poisson model instead of a logistic model is due to the noncollapsibility of the odds ratio, which is an issue for the propensity score (20, 21). Unless otherwise specified, the baseline exposure prevalence was 10%, and the baseline disease rate was 0.01 per person.

For each simulation setting, 1,000 samples were generated, each consisting of 10,000 study subjects. Simulations were conducted by using SAS, version 9.1, software (SAS Institute, Inc., Cary, North Carolina).

Adequate events per covariate, traditional model correctly specified. The first scenario compared performance of the summary scores under favorable conditions for traditional regression methods: an adequate number of events per covariate and the traditional regression model correctly specified. There were 10 independent binary covariates (reflecting applications in which these are indicators of medications or comorbidities), each with a prevalence of 10%. Prior work suggests that the performance of the disease risk score is affected by both the correlation between exposure and confounders and the degree of confounding present (10). The simulations thus varied both of these factors over a broad range. In one setting, we set the covariate-exposure odds ratios and covariate-disease rate ratios to 1.25 for all 10 covariates and simulated data with the exposure-disease rate ratio of 1.25. We then considered exposure-disease rate ratios of 1.5, 2.0, and 3.0. We then increased the covariate-exposure rate ratio to 2.0 and simulated data for the 4 different exposure-disease rate ratios. We repeated this with the covariate-exposure rate ratio set to 1.25 and the covariate-disease rate ratio set to 2.0. We repeated this again with both the covariate-exposure and covariate-disease rate ratios set to 2.0.

Limited number of events per covariate. The second scenario assessed the effect of model misspecification that can occur when regression models are constructed with traditional variable selection methods based upon statistical significance. These simulations compared this method with the alternative approach of using all measured covariates to calculate the summary scores.

The covariates were simulated by using the same set-up as described in the previous section, except that the covariate-disease rate ratios were set to 1.0 for 2 covariates, 1.25 for 6 covariates, and 1.5 for 2 covariates, and the covariate-exposure odds ratios were set to 1.25 for all 10 covariates. These values reflect the commonly encountered circumstance in which many of the covariates are modestly or not associated with the disease. We further assumed that the traditional variable selection methods resulted in models that included only the 2 covariates with covariate-disease rate ratios of 1.5. The summary scores were calculated from all of the potential covariates. We then repeated this setting the covariate-exposure odds ratio to 2.0 for all 10 covariates.

Propensity and disease risk scores

The propensity score was estimated by using logistic regression with the exposure as the dependent variable and the covariates as the independent variables. The estimated propensity score is then the fitted value from this regression model, that is, the estimated probability of being exposed conditional on each study subject’s individual covariate values.

The unexposed-only DRS, or probability of disease occurrence in the absence of the exposure, was estimated with Poisson regression in the unexposed group. The dependent variable was the disease, and the independent variables were the covariates. We estimated the full-cohort DRS using the entire cohort with exposure status set to unexposed for all simulated subjects.

Model-fitting approaches

For each simulated sample, the data were fit by using 4 different Poisson regression models. We first fit the exposure and all of the individual covariates in a single regression model, which was designated the “traditional” model. The remaining models fit the exposure and propensity score, the exposure and unexposed-only DRS, and the exposure and full-cohort DRS. The scores were expressed as linear terms. We also ran simulations expressing the summary scores as deciles, and the findings were similar.
Evaluation strategies

The primary measures of performance were the percent bias and coverage probabilities. Percent bias was the average difference between the estimated and true regression coefficients for the exposure divided by the true regression coefficient times 100%. Coverage probabilities were the proportion of the 95% confidence intervals that include the true regression coefficient. If a method has proper coverage, then for a 95% confidence interval, the coverage should be very close to 0.95. Secondary measures included the mean standard error of the estimated regression coefficient for the exposure, as well as the standard error estimate from the simulated samples. The standard error estimate is the standard deviation of the estimated regression coefficients from the simulated samples, and the mean standard errors should approximate the standard error estimates. Finally, the empirical power was estimated by dividing the exposure’s estimated regression coefficient by its standard error and tabulating the proportion of simulations in which this ratio exceeded 1.96, the upper 97.5 percentile for a 2-tailed 0.05-level test. Given that there is a true association between the exposure and disease, values closer to 1 are desirable.

RESULTS

Adequate events per covariate, traditional model correctly specified

Figures 1 and 2 graphically depict percent bias and coverage probabilities, respectively, under the different degrees of confounding likely to be encountered in practice for scenarios with an adequate number of events per confounder. When there was a moderate association between covariates...
and exposure (i.e., covariate-exposure odds ratio of 1.25), all 4 models had little bias. Performance was comparable except when there was a moderate association between covariates and disease (i.e., covariate-disease rate ratio of 1.25) in which the unexposed-only DRS model consistently had slightly greater bias. However, this bias never exceeded 3%. Coverage was consistent across all 4 models. When there were strong correlation between covariates and exposure and moderate association between covariates and disease (i.e., covariate-exposure odds ratio of 2.0 and covariate-disease rate ratio of 1.25), all models except for the unexposed-only DRS had little bias. The unexposed-only DRS model consistently had greater bias with range 6%–10%. All models had proper and consistent coverage except for the unexposed-only DRS. When the covariates were strongly associated with both exposure and disease (i.e., rate ratios of 2.0), the traditional regression and full-cohort DRS models had little bias. The propensity score model had greater bias, but it was consistently less than 3%. The unexposed-only DRS model had the greatest bias, but it was still consistently less than 5%. All 4 models had proper coverage.

**Adequate events per covariate, traditional model misspecified**

Figure 3 graphically depicts percent bias and coverage probabilities when the traditional model was misspecified. For both moderate and strong associations between covariates and exposure, the propensity score and full-cohort DRS models had little bias. The unexposed-only DRS models had greater bias, although for moderate associations between covariates and exposure, bias was consistently less than 3%. Not surprisingly, the misspecified traditional model had the greatest bias, although this bias was consistently less than 5% when the covariates and exposure were moderately correlated. All 4 models had comparable coverage for moderate associations between covariates and exposure. For strong associations between covariates and exposure,
the propensity score and full-cohort DRS models had proper coverage, whereas the unexposed-only DRS and misspeci-
fied traditional models had poorer coverage.

**Limited number of events per covariate**

Figure 4 graphically depicts percent bias and coverage probabilities for scenarios with limited numbers of events per covariate. All models performed well so long as there were more than 5 events per covariate. For smaller numbers of events, bias increased for all 4 models, with the unexposed-only DRS model having the greatest increase in bias. The other models were remarkably comparable. Coverage was comparable across all 4 models.

**DISCUSSION**

We assessed the performance of the increasingly popular practice of replacing individual covariates with summary scores in regression modeling for observational cohort studies. We thus ran simulations comparing traditional models with all of the covariates versus those that included either the propensity score or disease risk score derived from either the full cohort or from only unexposed patients.

When there was an adequate number of events per con-founder and when the covariates were moderately associated with the exposure, all 4 models were comparable and performed well. When covariates were strongly associated with both exposure and disease, the full-cohort DRS and traditional models performed best, followed by the propensity score model, and lastly by the unexposed-only DRS model. However, the unexposed-only DRS model still had reasonable performance. When the covariates were strongly associated with exposure and moderately with disease, all of the models except the unexposed-only DRS model performed well. In this setting, the unexposed-only DRS model had greater bias and poor coverage. The sensitivity of the performance of the unexposed-only DRS to covariates strongly
All 4 models were affected by the numbers of events per confounder. Bias increased for all 4 models as the number of events per confounder decreased, with the DRS models most affected by the number of events per confounder. The traditional regression and propensity score models were remarkably consistent, which could reflect the robustness of traditional regression (directly adjusting for all covariates in the exposure-outcome model) and/or the limited advantages of propensity score models in the scenarios we studied. Although the bias was small, 10%, for propensity scores at 1–2 events per confounder, this may seem counterintuitive and differs from the findings of Cepeda et al. (5). However, that study used logistic regression for the exposure-outcome model. Changing our simulations to a logistic rather than a log-linear model yielded findings similar to those of Cepeda et al. Further work that assessed such performance for much larger numbers of covariates, such as those utilized in high-dimensional propensity scores (22), would be useful.

Our findings thus suggest that, so long as the association between covariates and exposure is moderate, disease risk scores are an acceptable alternative to traditional models and propensity scores. Our simulations further suggest that, if the assumptions underlying the full-cohort disease risk score are met (absence of effect modifiers among the covariates), the full-cohort method is generally superior to the unexposed-only method, with performance generally equivalent to that of the propensity score. The superiority of the full-cohort DRS in our simulations is consistent with findings reported by Cadarette et al. (23).

Thus, the disease risk score may be particularly useful in circumstances in which propensity scores are not appropriate or have poor performance (5, 24, 25). These include exposures that are rare or that have a large number of categories. Because the disease risk score directly estimates each cohort member’s risk of developing the study outcome, it also provides a single summary measure of disease risk, particularly useful for assessing effect modification.

A limitation of simulation studies is whether they adequately cover the range of possible scenarios that are likely to be encountered with real-life data. We designed our simulations to reflect realistic covariate-disease and covariate-exposure rate ratios, as well as to reflect both previously published simulation and epidemiologic studies in terms of events per covariate, numbers of covariates, incidence of outcome, prevalences of exposure and covariates, measures of association, and sample size. However, it is likely that we have not considered some important settings; further work for these would be useful. In most of our simulation settings, all of the covariates were confounders. However, with real data, some covariates may only be related to the outcome and some only to the exposure. This has been examined for propensity scores (26). We plan to investigate this for disease risk scores in the future.

There are many unanswered questions regarding the use of both propensity and disease risk scores in practice. These include the effects of model misspecification on the scores (i.e., whether it is better to include all available variables or to apply variable selection procedures) and the use of either type of score for case-control studies (16). In addition, Hansen (19) noted the potential benefit of using a separate

Figure 4. Percent bias (A) and coverage probabilities (B) for limited numbers of events per covariate. Solid curve, traditional multiple outcome regression; short-dash curve, disease risk score (DRS)—unexposed only; long-dash curve, DRS—full cohort; dash-3 dots-dash curve, propensity score. The scenario included 10 binary covariates (prevalence of 10%), with covariate-exposure odds ratios and covariate-disease rate ratios of 1.25. Baseline exposure prevalence was 10%, and the exposure-disease rate ratio was 2.0. Baseline disease probability varied so that the expected number of events per covariate ranged from 1–2 to greater than 20.


associated with the exposure is consistent with what has been reported elsewhere (7). Hence, for adequate numbers of events per confounder, the full-cohort DRS and traditional regression models consistently performed the best. The propensity score models had similar performance or were at least comparable depending on the associations among covariates, exposure, and disease, and the unexposed-only DRS models were comparable when the covariates were moderately associated with exposure.

As expected, traditional models performed poorly when the model was misspecified because of exclusion of covariates only weakly associated with the disease outcome. The magnitude of the underperformance was affected by the strength of the association between covariates and exposure. Also, as expected, the performance of the unexposed-only DRS was affected by the covariate-exposure association. In contrast, both the propensity score and full-cohort DRS performed well, with little bias and good coverage probabilities.
sample to estimate the disease risk score, which is a very intriguing notion but may be difficult to implement in practice because it may be problematic to find a separate-yet-comparable cohort. Given the widespread and seemingly growing use of summary scores for epidemiologic research, further research to address these questions is needed.

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics, Vanderbilt University, Nashville, Tennessee (Patrick G. Arbogast); Department of Preventive Medicine, Vanderbilt University, Nashville, Tennessee (Patrick G. Arbogast, Wayne A. Ray); and Geriatric Research, Education, and Clinical Center, Nashville Veterans Affairs Medical Center, Nashville, Tennessee (Wayne A. Ray).

This work was supported in part by the Centers for Education and Research on Therapeutics, Agency for Healthcare Research and Quality (HS1-0384, HS016974).

The authors thank Dr. M. Alan Brookhart for his valuable discussions and suggestions.

Conflict of interest: none declared.

REFERENCES