Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care

I-Chan Huang[†], Constantine Frangakis[†], Francesca Dominici[†], Gregory B. Diette[†‡], Albert W. Wu[†‡]

Bloomberg School of Public Health[†], School of Medicine[‡], Johns Hopkins University, Baltimore, Maryland

KEYWORDS: Physician group, profiling, propensity score, risk adjustment

Introduction

For quality assessment, random assignment of patients to different health care providers would be the ideal method to balance the distributions of patient characteristics among providers, thus removing confounding by factors related to selection and provider performance. However, it is not practical to randomly assign patients to different providers. Therefore, in observational studies statistical risk-adjustment techniques are used to remove confounding effects [1]. The most common method for risk adjustment is regression modeling [1-3]. However, the standard regression-based risk adjustment is limited because it does not explicitly assure balance in the distributions of covariates among providers [4]. The importance of explicit balancing increases with the number of covariates [5].

The propensity score introduced by Rosenbaum and Rubin is a method for producing balance of many covariates between two groups [6;7]. This method can explicitly balance a set of many covariates by estimating the probability (propensity) of assignment to a specific provider given those covariates. For observed covariates, theory assures that given any value of the propensity score, the subgroups of patients who enroll with different providers will have the same joint distribution in all the covariates that were used to estimate that propensity score [5-7]. This is a main advantage of propensity score methods, because they allow a straightforward check for whether the adjustment has made providers comparable with respect to the observed covariates [5-7].

Without a controlled design, the true unconfounded differences are not known, and indirect evidence is used to judge whether or not the propensity score method is proper. Indirect evidence is provided when (a) there is explicit validation that the propensity score has balanced all important observed covariates between the comparison groups; and (b) the results from the propensity score method differ importantly from those of the methods that do not use propensity scores.

Propensity score techniques were originally designed for two-group comparisons [6;7], and have been used in observational studies with cohort or case-control designs to reduce the bias from estimated effects of treatment programs, and social or health services programs. Imbens developed a modified method for comparison of multiple groups [8]. To our knowledge, such a method has not been used in health services research for profiling multiple providers. In addition, with multiple providers, provider-specific estimates of performance are subject to regression-to-the-mean due to small case numbers within provider [9]; this issue has not been addressed using propensity scores.

The goals of this study were (a) to develop and validate a propensity score-based risk adjustment method to estimate performance of multiple providers, in order to simultaneously balance all observed covariates, as well as to address regression-to-the-mean, and (b) to compare our propensity score-based method vs. more conventional regression-based risk adjustment methods of evaluating and ranking performance in 20 California physician groups. Satisfaction with asthma care was used as the performance indicator. The regression-based method adopted in this study is a hierarchical model that adjusts for the regression-to-the-mean inaccuracies of standard methods, but without using the propensity score [2;9-13].

We hypothesized that the propensity score-based risk adjustment method would balance all observed covariates. If, in addition, the propensity score method also results in substantial ranking differences in physician group performance compared to the standard method, this finding will provide indirect evidence for greater usefulness of the propensity score method.

Methods

This study was conducted in conjunction with 20 California physician groups that participated in the 1998 Asthma Outcomes Survey (AOS). The AOS was initiated by the Pacific Business Group on Health (PGBH) and HealthNet to evaluate, improve and report on the quality of asthma care at physician group level. The 20 participating physician groups were instructed to use administrative materials to identify all managed care patients with at least one asthma-related encounter in the outpatient, emergency or inpatient settings (identified by ICD-9 code 493.xx) between January 1, 1997 and December 31, 1997. Patients had to be continuously enrolled in the physician group for that calendar year. From eligible patients, this study randomly selected a sample of 650 patients from each physician group. If a physician group had fewer than 650 eligible patients, then all eligible patients were sampled. A total of 7,820 patients had usable addresses and met the study eligibility criteria.

Patient data were collected by self-administered mailed survey. The survey instrument was largely based on the "Health Survey for Asthma Patients" [14-16]. The survey period began in July 1998 and ended in February 1999. The survey was fielded by PBGH and HealthNet using an identical methodology. A total of 2,515 responses were obtained.

All of the variables used in this study, including risk adjustors and outcome indicators, were from the patient survey. Satisfaction with asthma care was used as the performance indicator. In the survey instrument, patient satisfaction was rated on a 5-point Likert-type scale (i.e. Poor/ Fair / Good/ Very Good/ Excellent). We dichotomized this variable into "greater satisfaction (Very Good/ Excellent)" vs. "less satisfaction (Poor/ Fair/ Good)".

We adjusted for exogenous factors, that is, factors for which providers have no influence (mainly patient characteristics, such as age, sex, education, baseline severity, etc). We did not adjust for endogenous factors, that is, factors that providers can affect (mainly physician group characteristics, such as physician group specialty, number of supplementary staff, etc) [17]. Adjusting for endogenous factors may mask true performance of physician groups because these factors can influence the quality of care.

Two analytic methods were used to evaluate the impact of different risk-adjustment methods on physician group performance. We chose the first physician group as the reference group for comparisons among different methods.

For method 1, we implemented a hierarchical model-based risk adjustment without propensity score. At the first stage (patient level) we used a logistic regression model for estimating the group-specific log-odds ratio of patient satisfaction as a function of patient characteristics, including age, sex, education level, type of insurance, prescription drug coverage, asthma severity, number of comorbidities, and health status. At the second stage (group level), we modeled the variation of the log-odds ratio across 20 physician groups. The hierarchical regression approach takes into account clustering of patients within physician groups and the different number of patients within each physician group (reliability). Under the hierarchical model the group-specific estimates of performance are shrunk toward the average performance common to all physician groups to address the regression-to-the mean that arises with comparison of multiple groups [9;10;12;18].

Based on this method, the relative performance of physician groups was assessed by estimating the risk-adjusted odds ratio (OR) of satisfaction with care (greater satisfaction vs. less satisfaction) attributable to the j^{th} physician group relative to the first physician group (reference group) by exponentiating the difference between the estimated provider-specific random intercept of the j^{th} (j=2, …, 20) and the first physician groups.

For Method 2, we implemented a propensity score-based risk adjustment. Under this method, the main goal was to estimate the proportions, P_j^+, of satisfied patients in the hypothetical scenario under which all patients would have been enrolled in the j^{th} group (j=1,…,20). Then, performance was compared among physician groups by comparing the proportions P_j^+ (j=1,…,20). We estimated these proportions by adapting Imbens's propensity score method for multiple groups, [8] with a method for accounting for regression-to-the-mean. Specifically, we developed five major steps for propensity score-based risk adjustment to compare performance of physician groups: (1) calculation, for each patient, of twenty estimated propensity scores for enrolling in the twenty physician groups , (2) sub-classification of patients

into quintiles based on the propensity scores, (3) validation of estimation of the propensity scores, (4) estimation of the adjusted proportion of satisfaction, P_j^+ , of each physician group by combining across the five propensity strata, and (5) estimation of the relative performance of each physician group using shrinkage techniques.

In this study, the impact of different risk-adjustment methods on physician group profiling was measured in terms of differences in estimated performance ranking between the two methods. Rankings of physician groups were compared based on the odds ratio (OR) of greater satisfaction vs. less satisfaction for the j^{th} physician group vs. the reference group. Two methods were used to demonstrate the changes in ranking: percentage changes in absolute ranking (AR) and percentage changes in quintile ranking (QR). Percentage changes in AR represented the portion of physician groups that changed in ranking. The QR represented the portion of physicians groups that moved into a different quintile of ranking, which was evaluated using a weighted-kappa statistic. We used quadratic-weighted kappa rather than standard kappa (no weight) to reflect the ordinal nature (quintile) of the ranking scale [19].

We used SAS 8.1 with Glimmix Macro for hierarchical modeling analysis, STATA 7.0, existing routines for shrinkage [20], and S-plus 2000 for developing and validating the propensity score method (available from the authors).

Results

Of the 20 participating physician groups, 8 were located in Northern California and 12 in Southern California. Patients ranged in age from 18-56 years with a mean age of 39.9 years [SD: 9.5]. 71.2% were female; 70.3% were white, and 5.1% were African-American; 81.6% had at least some college education; 69.1% obtained health insurance through their employer, and 24.8% by themselves; and 96.5% had prescription drug coverage. In terms of clinical characteristics, 14.4% had mild intermittent asthma, 19.2% had mild persistent asthma, 49.3% had moderate persistent asthma, and 17.1% had severe persistent asthma. The mean number of comorbidities was 2.1 [SD: 1.4]. For general health status, the mean SF-36 physical component score (PCS) was 45.7 [SD: 10.3], and the mean SF-36 mental component score (MCS) was 47.4 [SD: 10.7].

Before applying the propensity score method, there was imbalance in each covariate across 20 physician groups. For example, the range of mean patient age among the 20 physician groups was 35.6 to 43.4 [SD: 1.9] ($p<0.01$); the range of mean severity was 2.5 to 3.0 [SD: 0.49] ($p<0.01$); the range of mean SF-36 PCS was 41.6 to 52.7 [SD: 2.19] ($p<0.01$). The distributions of sex, level of education, type of health insurance, prescription drug coverage, and number of comorbidities were also significantly unbalanced across the 20 physician groups (all $p<0.01$). The difference in distribution of SF-36 MCS was marginally significant ($p=0.05$).

After applying the propensity score method, the balance of each covariate across the 20 physician groups improved substantially. The ranges of physician-specific distributions of the important covariates of age, asthma severity, and SF-36 PCS across the 20 physician groups before and after adjustment using the propensity score techniques. After propensity score adjustment, the standard deviation for age was reduced from 1.9 to 0.93 (51.1% reduction) and the range was reduced from 7.8 to 4.7. For asthma severity, the standard deviation was reduced from 0.14 to 0.08 (42.9% reduction) and the range was reduced from 0.49 to 0.34. For SF-36 PCS, the standard deviation was reduced from 2.19 to 0.67 (69.4% reduction) and the range was reduced from 11.04 to 3.23. For the other covariates, the ranges of distributions were also significantly reduced. After adjustment, only 3.9% of all comparisons for balance status (7 out of 180 comparisons) were statistically significant at the level of $p<0.05$, indicating that the propensity score adjustment produced balance, in the observed covariates, similar to that which would be expected by randomization of these covariates across the physician groups.

When comparing the propensity score-based method (Method 2) to the hierarchical model-based method (Method 1), there was a 75% difference in absolute ranking (AR) and 50% difference in quintile ranking (QR), with a weighted kappa of 0.69. More specifically, the differences in rankings fell into two clusters. For absolute rankings (AR), most of the shifts occurred within the 80% middle ranks of physician groups. For quintile rankings (QR), five physician groups (ID number 4, 6, 7, 18, and 19) shifted their quintile ranking into a lower quintile after propensity score adjustment. Also, five physician groups (ID number 2, 8, 15, 17, and 20) shifted their quintile ranking into a higher quintile.

However, physician groups with the best (ID number 9) or worst (ID number 11, 12, and 14) performance in the hierarchical model-based method did not shift in ranking after applying the propensity score. Compared to the hierarchical model-based method, the propensity score-based method produced only slightly larger standard errors of the odds ratios.

Discussion

To accurately compare provider performance, it is critical to control for differences in the characteristics of patients treated by different providers. Owing to the difficulty of designing a randomized experiment to compare provider group performance, risk adjustment is used to account for background differences. In this study, we applied a propensity score method for multiple physician groups to explicitly balance the covariates and thus produce more proper adjustment.

Our results showed that the propensity score-based risk adjustment method improved the balance of covariates among physician groups to a degree similar to what would be expected by randomization of these covariates. Demonstration of this balance also showed that the propensity score method was satisfactorily specified [7;8]. Moreover, the propensity score-based method produced substantially different ranking results from the regression-based method, suggesting that provider performance profiling is very sensitive to different statistical approaches. In the absence of a controlled experiment, we cannot have direct evidence that the propensity score-based method is superior to the regression-based method. However, the above two results taken together do provide indirect evidence that the propensity score method is more reliable than the regression-based method, because the latter method cannot explicitly assure good comparability of the distribution of patient covariates across providers and provides different results from the method (using propensity scores) that can assure such comparability.

To date, the development of risk-adjustment methods for provider profiling has been limited. Most efforts have emphasized careful selection of risk-adjustors [1], while few have focused on the impact of statistical approaches [2;3;9;11;21], and none has underscored the importance of explicitly balance of covariates.

From a methodological point of view, there are several advantages of applying propensity score method over regression-based methods in provider profiling. First, propensity score method allows researchers to compare provider performance with similar patient characteristics without specification (or assumption) of linear relationship between the profiling indicator and risk adjustors as is required by a regression-based method. Some risk adjustors, such as patient's age, may not fit this linear assumption. The application of regression-based risk adjustment to balance the distributions of covariates across providers is particularly limited when some providers have more skewed patient characteristics than others, since it involves extrapolation where there is little overlap of the covariate distributions [6]. Propensity score methods model the assignment of patients (rather than the outcome or performance indicators) to a specific provider based on patient characteristics. Therefore, propensity score methods can be more robust to model misspecification than regression-based methods. Second, propensity score methods may avoid the loss of degrees of freedom seen in statistical modeling. Rather than accounting for risk adjustors individually as in the regression-based method, the propensity score method reduces the numbers of risk adjustors into a composite variable, avoiding the loss of statistical power due to a large number of risk adjustors.

Risk-adjustment techniques have been used in health policy studies to minimize the impact of biased selection in setting capitation payment and comparing provider performance. It is important to clarify that different risk-adjustment methods may be appropriate depending on their purposes. For setting premium rates, it is desirable to develop models that can predict individual patients' future costs based on specific individuals' values of a group of risk adjustors. For that purpose, a regression-based model is a practical prediction tool. If, however, the purpose is to compare overall provider performance, it is important to properly balance many covariates with propensity score methods.

For practical use in provider profiling, we would recommend using a propensity score-based method to refine and complement regression-based risk adjustments. A general regression-based method can be used first to select a subset of risk adjustors, followed by application of propensity-score techniques to explicitly balance those risk

adjustors among providers. To identify the best and worst providers for benchmarking or quality management, it may be useful to plot ranking shifts based on different risk-adjustment methods as demonstrated in Figure 2. For rankings for which both methods agree, we can be more confident in the results. For the rankings for which the methods do not agree, and so long as the standard errors are comparable between the two methods, explicit balance of the covariates resulting from the propensity score method are likely to be more trustworthy.

In interpreting our findings, several limitations should be noted. First, the patient survey had a low response rate. A lower response rate (35-50%) is a common phenomenon on satisfaction survey [22]. However, this seems unlikely to affect the comparison among regression-based and propensity score-based methods. Second, the set of risk adjustors included in our risk-adjustment models may not be optimal. In this study, all of the risk adjustors were collected from the patient survey. We did not collect clinical assessments, some patient characteristics (e.g. personal income or family size), and other non-patient characteristics that providers cannot influence (e.g. health plan or physician group penetration rate), all of which could lead to confounding and hidden biases in provider performance comparison [23]. Thus, as with regression methods, our propensity score method can balance unobserved covariates only to the extent that those are correlated with the observed covariates [5]. Finally, in this study, we only examined patient satisfaction as the performance indicator. Further studies need to examine the impact of other indicators such as process or outcome to reflect the impact on provider profiling.

In conclusion, in this paper we proposed a novel propensity score method to compare performance across multiple physician groups by properly balance patient-specific covariates across physician groups, by taking into account the clustered nature of the data, and the different number of patients enrolled in the different physician groups. Further improvements are needed in risk-adjustment methodology for provider profiling, especially in methods that combine propensity score techniques together with other statistical approaches to address these issues.

Appendix: Description of the propensity scores method to compare multiple physician groups

The goal was to estimate the proportions, P_s^+, of satisfied patients, if all patients had been enrolled in group j (for j=1,…,20). There were five major steps to use propensity scores method to estimate these proportions.

Step 1: Calculation of the propensity score, e_{ij}, of patient i enrolling in the j^{th} physician group (j=1, …,20)

Each patient has 20 propensity scores, each of them representing the probability of enrollment in each of 20 physician groups. The propensity score e_{ij} was estimated as the conditional probability of patient i to have been enrolled in the j^{th} physician group (j=1,…,20) as a function of the patient's specific covariates. The preliminary propensity scores estimates were obtained by using a multinomial logistic regression model.

Step 2: Subclassification of patients into quintiles based on the propensity scores

For evaluating the j^{th} physician group the propensity scores of all 2,515 patients, as estimated from step 1, were ranked and then sub-classified into five strata based upon quintiles. Evidence showed that such sub-classification based on the quintiles of the scores generally reduces bias due to unbalanced covariates by 90% [7;24]. Patients in these five strata (low to high propensity scores) were then classified by whether or not they actually belonged to the j^{th} group, and then further classified by whether or not they were satisfied with asthma care.

Step 3: Validation of propensity scores estimates

The theory described in Rosenbaum and Rubin [6;7] and Imbens [8] shows that the estimated propensity score in Step 1 has the correct balancing properties, stated in paragraph 3 of the introduction, if, for each three-way classification as shown in Table 1, the distribution of covariates was the same for patients actually in the j^{th} group (column 3) and in the other 19 groups (column 4) within sub-classes of the propensity score. We therefore validated

the estimation in Step 1 by testing equality (or balance), for each of the physician groups from 20 Tables (not shown), of the distribution for each covariate between columns 3 and 4 within the propensity strata. For these diagnostics, we used two-way ANOVA and logistic regression.

If the variables were not well balanced, as judged by comparison to the expected imbalance merely due to chance, interaction terms of that variable with other variables were estimated into a logistic model along with all previous variables, and a new propensity score was calculated [7].

Table 1: Comparing the j^{th} physician group vs. other 19 physician groups

	Patients actually in the j^{th} physician group (j=1,…,20)		Patients actually in the other 19 physician groups	Patients in 20 physician groups
Propensity stratum, s	Patients who are satisfied with care in strata s	Total patients in strata s	Total patients in strata s	Total patients in strata s
(1)	(2)	(3)	(4)	(5)
1	$C_{j,1}$	$N_{j,1}$	$M_{j,1}$	$N_{j,1} + M_{j,1}$
2	$C_{j,2}$	$N_{j,2}$	$M_{j,2}$	$N_{j,2} + M_{j,2}$
3	$C_{j,3}$	$N_{j,3}$	$M_{j,3}$	$N_{j,3} + M_{j,3}$
4	$C_{j,4}$	$N_{j,4}$	$M_{j,4}$	$N_{j,4} + M_{j,4}$
5	$C_{j,5}$	$N_{j,5}$	$M_{j,5}$	$N_{j,5} + M_{j,5}$
Overall	$\Sigma C_{j,s}$	$\Sigma N_{j,s}$	$\Sigma M_{j,s}$	$\Sigma (N_{j,s} + M_{j,s})$

Step 4: Estimation of overall risk-adjusted proportion of satisfied patients

For each physician group j, we first estimated the proportion of patients who were satisfied with asthma care within each stratum P_j^s. Then, we estimated the overall risk-adjusted proportion of satisfied patients, P_j^+, by averaging the P_j^s across strata with weights equal to the relative frequency of the patients belonging to each stratum.

(1) Estimate the proportion of patients who were satisfied with asthma care in the j^{th} physician group (j=1,…,20) and in stratum s (P_j^s):

$$P_j^s = C_{j,s} / N_{j,s}$$

(2) Estimate the overall risk-adjusted proportion of satisfied patients using weighted average for the j^{th} physician group across the five strata by P_j^+, where

$$P_j^+ = \Sigma (P_j^s * W_{j,s}), \text{ with weights}$$

$$W_{j,s} = (N_{j,s} + M_{j,s}) / \Sigma (N_{j,s} + M_{j,s})$$

Step 5: Estimation of the relative performance of each physician group using shrinkage techniques

We estimated the log-odds, $\log [P_s^+ / (1- P_s^+)]$, of satisfaction with asthma care for each physician group by using of the estimated overall risk-adjusted proportions P_j^+ in step 4. To address regression-to-the-mean associated with comparing multiple physician groups, we then adjusted these preliminary log-odds estimates of physician group performance towards the grand mean using the shrinkage technique developed by Morris [10]. Finally, we estimated the overall risk-adjusted odds ratio (OR) of physician group performance in comparing the j^{th} physician group (j=, 2…, 20) vs. physician group 1 by exponentiating the difference of the corresponding log odds.

References

1. Iezzoni LI. Risk adjustment for measuring healthcare outcomes. 2nd ed ed. Chicago, Illonis: Health Administration Press, 1997.

2. DeLong ER, Peterson ED, DeLong DM, Muhlbaier LH, Hackett S, Mark DB. Comparing risk-adjustment methods for provider profiling. *Stat Med* 1997; 16(23):2645-2664.

3. Shahian DM, Normand SL, Torchiana DF, Lewis SM, Pastore JO, Kuntz RE et al. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg* 2001; 72(6):2155-2168.

4. Dehejia RH, Wahba S. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *J Am Stat Assoc* 1999; 94(448):1053-1062.

5. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997; 127(8 Pt 2):757-763.

6. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70(1):41-55.

7. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 1984; 79(387):516-524.

8. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika* 2000; 87(3):706-710.

9. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. *Ann Intern Med* 1997; 127(8 Pt 2):764-768.

10. Morris CN. Parametric empirical Bayes inference: theory and applications. *J Am Stat Assoc* 1983; 78(381):47-55.

11. Sullivan LM, Dukes KA, Losina E. Tutorial in biostatistics. An introduction to hierarchical linear modelling. *Stat Med* 1999; 18(7):855-888.

12. Diez-Roux AV. Multilevel analysis in public health research. *Annu Rev Public Health* 2000; 21:171-192.

13. Katon W, Rutter CM, Lin E, Simon G, Von Korff M, Bush T et al. Are there detectable differences in quality of care or outcome of depression across primary care providers? *Med Care* 2000; 38(6):552-561.

14. Steinwachs DM, Wu AW, Skinner EA. How will outcomes management work? *Health Aff (Millwood)* 1994; 13(4):153-162.

15. Diette GB, Wu AW, Skinner EA, Markson L, Clark RD, McDonald RC et al. Treatment patterns among adult patients with asthma: factors associated with overuse of inhaled beta-agonists and underuse of inhaled corticosteroids. *Arch Intern Med* 1999; 159(22):2697-2704.

16. Wu AW, Young Y, Skinner EA, Diette GB, Huber M, Peres A et al. Quality of care and outcomes of adults with asthma treated by specialists and generalists in managed care. *Arch Intern Med* 2001; 161(21):2554-2560.

17. Welch HG, Black WC, Fisher ES. Case-mix adjustment: making bad apples look good. *JAMA* 1995; 273(10):772-773.

18. Raudenbush SW, Bryk AS. Hierarchical linear models applications and data analysis methods. 2nd ed ed. Thousand Oaks, California: Sage Publications, 2002.

19. Streiner DL, Norman GR. Health measurement scales a practical guide to their development and use. 2nd ed ed. Oxford, UK: Oxford University Press, 1995.

20. Everson PJ, Morris CN. Splus software: hierarchical normal regression model. Boston, Massachusetts: 1993.

21. Zaslavsky AM. Statistical issues in reporting quality data: small samples and casemix variation. *Int J Qual Health Care* 2001; 13(6):481-488.

22. Fowler FJ, Jr., Gallagher PM, Stringfellow VL, Zaslavsky AM, Thompson JW, Cleary PD. Using telephone interviews to reduce nonresponse bias to mail surveys of health plan members. *Med Care* 2002; 40(3):190-200.

23. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann Intern Med* 2002; 137(8):693-695.

24. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968; 24(2):295-313.