Dis Manage Health Outcomes 2005; 13 (2): 107-115
1173-8790/05/0002-0107/$34.95/0

Using Propensity Scores to Construct Comparable Control Groups for Disease Management Program Evaluation

Ariel Linden,[1] *John L. Adams*[2] and *Nancy Roberts*[3]

1 Linden Consulting Group, Portland, Oregon, USA
2 RAND Corporation, Santa Monica, California, USA
3 Integrated Performance/Six Sigma Champion, Providence Health System, Portland, Oregon, USA

Abstract

Introduction: The ability of observational studies to draw conclusions on causal relationships between covariates and outcomes can be improved by incorporating randomly matched controls using the propensity scoring method. This procedure controls for pre-program differences between the enrolled and non-enrolled groups by reducing each participant's set of covariates into a single score, which makes it feasible to match on what are essentially multiple variables simultaneously. This paper introduces this concept using the first year results of a congestive heart failure (CHF) disease management (DM) program as an example.

Methods: This study employed a case-control pre-post study design with controls randomly matched to patients based on the propensity score. There were 94 patients with CHF enrolled in a DM program for at least 1 year (cases), who were matched to 94 patients with CHF drawn from a health plan's CHF population (controls). Independent variables that estimated the propensity score were pre-program: hospital admissions, emergency department (ED) visits, total costs, and risk level. Baseline (1 year prior to program commencement) and 1-year outcome variables were compared for the two groups.

Results: The results indicated that, at post-program, program participants had significantly lower hospitalization rates (p = 0.005), ED visit rates (p = 0.048), and total costs (p = 0.003) than their matched controls drawn from the CHF population.

Conclusions: Because of its simplicity and utility, propensity scoring should be considered as an alternative procedure for use with current non-experimental designs in evaluating DM program effectiveness.

The randomized controlled trial is considered the gold-standard research and program evaluation design.[1,2] Randomization reduces the threat of selection bias by giving each member of the population an equal opportunity to be chosen for inclusion in the study, and adding a control group reduces many threats to the validity of the study's findings (e.g. time trends, regression to the mean). As desirable as the randomized controlled trial design may be, this model is not suited for many research endeavors unless the study is being conducted in a tightly controlled environment.

Disease management (DM) by its very nature is population-based and, thus, cannot be tightly controlled. Therefore, other more appropriate observational study designs must be sought.[3-6] That said, the most common method currently used in the DM industry for evaluating program outcomes is referred to as a 'total population approach'.[3] This model is a pre-test/post-test design, which is a relatively weak research and evaluation technique.[7-9] The most basic limitation of this design is that there is no randomized control group for which comparisons of outcomes can be made, thereby allowing several sources of bias and/or competing extraneous confounding factors to offer plausible alternative explanations for any change from baseline.[3,7-9] Advocates of this approach argue that most threats to validity are nullified by using the entire population in the analysis.[10] However, unless some basic factors are controlled for, such as the purchaser's case mix and the membership turnover rate, bias still remains a significant concern.[3] Even with these controlling variables in place, this

Table I. A comparison of pre- and post-first-year program characteristics of the congestive heart failure (CHF) disease management (DM) intervention group and the CHF population from which they were drawn. Values are means (standard errors) [$US, 2003 values]

Variable	DM intervention group (n = 94)	CHF population (n = 4606)	p-Value[a]
Age (years)	77.4 (0.96)	76.6 (0.19)	0.539
Sex[b]	0.51 (0.05)	0.56 (0.01)	0.336
Resident of Portland, OR, USA[c]	0.17 (0.04)	0.69 (0.01)	<0.0001
Health risk[d]	0.54 (0.05)	0.40 (0.007)	<0.0001
Pre-program (per member per year)			
Admission rate[e]	1.13 (0.15)	0.50 (0.02)	<0.0001
ED visit rate[e]	0.70 (0.11)	0.40 (0.01)	0.003
Total costs[f]	$18 287 ($2053)	$8974 ($257)	<0.0001
Post-program (per member per year)			
Admission rate[e]	0.59 (0.10)	0.87 (0.02)	0.0008
ED visit rate[e]	0.57 (0.08)	0.58 (0.02)	0.1874
Total costs[f]	$11 874 ($1408)	$16 036 ($370)	0.005

a p-Values are two-tailed t-tests for independent samples.

b A score of 1 indicates women and 0 indicates men.

c A score of 1 indicates resident within Portland and 0 indicates resident outside of Portland.

d A score of 1 indicates high risk and 0 indicates low risk for future CHF-related claims.

e Hospitalizations and ED visits were included only if they were CHF specific.

f Total costs included all associated costs per member, disease and non-disease related, excluding pharmacy costs (this exclusion was a result of a pharmacy benefit not being available to all members, making comparative analyses difficult).

ED = emergency department.

method can be confounded with environmental changes unrelated to the DM program interventions.

If the results of a DM program intervention may be suspect, why then do DM program evaluations continue to shun the use of randomized control groups? As a practical matter, DM programs currently do not use randomized control groups under the belief that: (i) it would be costly and difficult to track longitudinally behavioral change and outcomes for a group not under their purview; (ii) the organization may be hesitant to offer services to one subset of the population while withholding that same 'value-added' benefit to others; and (iii) there is a need to treat all members with the disease (if these interventions are indeed clinically effective) because each member receiving the intervention has the potential of adding to the medical cost savings and positive clinical outcomes promised by the program.

With these concerns in mind, an alternative procedure that should be considered for use with current non-experimental designs in evaluating DM program effectiveness will be presented in this paper. This method, termed 'propensity scoring',[11-15] utilizes existing data sources to create randomly matched controls, which are conditional on having an adequate set of observable characteristics for both DM program participants and non-participants. We will illustrate the use of a propensity scoring technique on observational data obtained from the first year of a congestive heart failure (CHF) DM program in evaluating its outcomes. However, the reader should keep in mind that the focus of this paper is on the propensity scoring methodology and not a comprehensive evaluation to determine the effectiveness of this particular CHF program. As such, other aspects of the DM program that would typically be examined and described will not be introduced in this paper because it would unnecessarily draw attention away from the propensity scoring method as the focal point.

Principles of Propensity Scoring

In general, DM programs provide high-intensity interventions (such as telephonic nurse/coaching services) to only a small number of participants out of a much larger population of patients with similar disease. Matched sampling techniques attempt to choose members from the untreated population so that they are similar to the program participants with respect to one or more pre-program variables. Controlling for differences in pre-intervention characteristics is extremely important in DM because program participants are typically dissimilar to non-participants (as a rule, DM programs strive to enroll those individuals at highest risk for incurring higher costs or higher healthcare utilization during the program term, thereby creating an unbalanced case mix between enrolled and non-enrolled groups).

The propensity score, defined as the probability of assignment to the treatment group, conditional on covariates[11] (i.e. independent variables), can control for pre-intervention differences between the enrolled and non-enrolled groups. The underlying assumption for using the propensity score in DM is that enrollment in the program is associated with observable pre-program variables (e.g. age, sex, utilization, and cost).[14,15] Propensity scores are derived from a logistic regression equation,[16,17] which reduces each participant's set of covariates into a single score, making it feasible to match on what are essentially multiple variables simultaneously.[9] There are several ways in which program participants can be matched to non-enrolled members by their propensity score, such as pair-wise matching, matching with and without replacement, matching using propensity score categories,[18] matching based on the Mahalanobis distance,[13] and kernel-density matching.[19] However, at present there is no thorough review available of the advantages and disadvantages of all these matching options.[9] The technique employed for evaluating the CHF program data will be matching based on the nearest propensity score. This method was chosen because of its straightforward implementation and conceptual simplicity.

Methods

Preliminary Data Analysis

The data evaluated in this paper come from an internally built patient-focused CHF program at a medium-sized health plan in Oregon, USA. These data represent the first year's experience of the program for those members with CHF who were continuously enrolled in the health plan for the year prior to commencement of the program as well as for the entire program year. Continuously enrolled populations were used for both the intervention and control groups to allow equal opportunity to experience the utilization-outcome events of interest (e.g. emergency department [ED] visits, and hospitalizations). The terms 'baseline' and 'pre-program' both refer to the time period 1 year prior to the start of the CHF program interventions.

Table I presents both pre-and post-program characteristics of the CHF program participants compared with the entire CHF population from which they were drawn. Figure 1 provides an illustration of this comparison for hospital admissions. As mentioned in this section, criteria for inclusion in either group required continuous enrollment for 1 year prior to commencement of the CHF program. Additionally, the 94 program enrollees were participants in the telephonic nursing intervention component of the program for the entire first program year, while the 4606 non-program members were continuously enrolled in the health plan during that same period and were not exposed to any intervention. A small number of members who began the CHF program but did not complete the full year were not included in either the intervention or control groups. As shown in table I, there were significant differences between program participants and the CHF population in both their geographic location and pre-program utilization and costs. These results are not surprising since the program specifically targeted the highest-risk CHF members primarily outside of the Portland, USA, service area for enrollment.

Pre-post outcomes between the two groups display a similar trend across all measures. With the exception of ED visits, the DM program cohort showed a significant reduction in hospitalizations and costs, compared with the general CHF population. This counters the significant increases in all these measures for the general CHF population. While these data are compelling, there are two major issues with these results that bear further consideration. First, as indicated earlier and illustrated in figure 1, the DM program cohort was very different with respect to baseline variables from the CHF population from which they were drawn. This raises the concern of selection bias and clouds the interpretation of the outcomes. Second, since DM program participants were specifically chosen based on their past high level of utilization and cost, there is a concern that the reduction in utilization and cost during the program year is indicative of regression to the mean. Even though the DM cohort appeared to demonstrate reduced utilization and costs (whereas the CHF population's outcomes appeared to worsen over the year), we may not be able to attribute the observed results solely to the program.

Estimating the Propensity Score

Logistic regression[16,17] is the most widely used method for estimating the propensity score. The form of the equation used here is shown in equation 1.

$$\Pr(Y_i = 1) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}\right)\right]}$$

where $\Pr(Y_i = 1)$ is the propensity score (the probability of being enrolled in the DM program), the βs are parameters to be estimated and the Xs are the independent variables. Thus, propensity scores range from 0 to 1, with 1 indicating a perfect probability of being enrolled and 0 indicating a perfect probability of not being enrolled.

Seven individual level independent variables were used for estimating the model including baseline age, sex (a score of 1 indicates women and 0 indicates men), service area, number of hospitalizations, number of ED visits, total costs, and health-risk level. These variables were readily available through health plan claims data. The addition of clinical variables indicating the severity of disease (e.g. New York Heart Association classification or ejection fraction) might have strengthened the model; however, these types of data were not generally available for members not participating in the DM program. Additionally, since this health plan's Medicare product does not include a pharmacy benefit, drug claim data were also not available for most members with CHF. One of the most appealing traits of the propensity score is that any pre-program variables that differentiate between the groups should be considered, whether they are linked to the outcome or not.[15,20-24] Therefore, the DM program evaluator should not hesitate to try any variables at their disposal, especially given the limited number of data types readily available.

A dummy variable was created for service area indicating whether the individual resided within or outside of Portland, OR, USA (a score of 1 indicates Portland and 0 indicates outside of Portland). Other CHF initiatives were being implemented by the health plan for its members in Portland who were served by one large physician organization during the baseline period. Thus, the DM program specifically targeted members outside of Portland and a small subset of Portland members who were not served by the large physician organization for enrollment in the program. To reduce the potential for contamination bias, patients with CHF who enrolled in the DM program and lived in the Portland area did not participate in the other Portland initiatives.

Health risk was a dichotomous variable indicating whether the individual was initially classified as low or high risk (a score of 1 indicates high risk and 0 indicates low risk) for future CHF-related claims. This classification was based on a claims data algorithm (using baseline period data), which stratified risk according to the amount and type of health services used for a given member. For example, a prior disease-specific hospitalization or ED visit would place a member into the high-risk category whereas regular scheduled office visits with no acute exacerbations may classify a member as low risk.

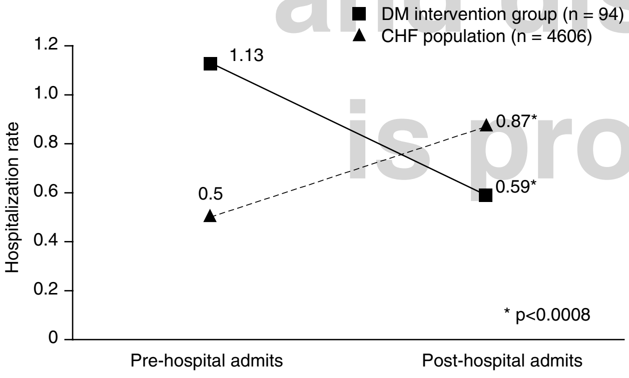

Fig. 1. A comparison of hospital admission rates between the congestive heart failure (CHF) disease management (DM) program and the CHF population from which they were drawn. All patients were continuously enrolled with the health plan for at least 2 years (1 year prior to program commencement and the duration of the first program year). p-Values are two-tailed t-tests for dependent samples. * p < 0.0008.

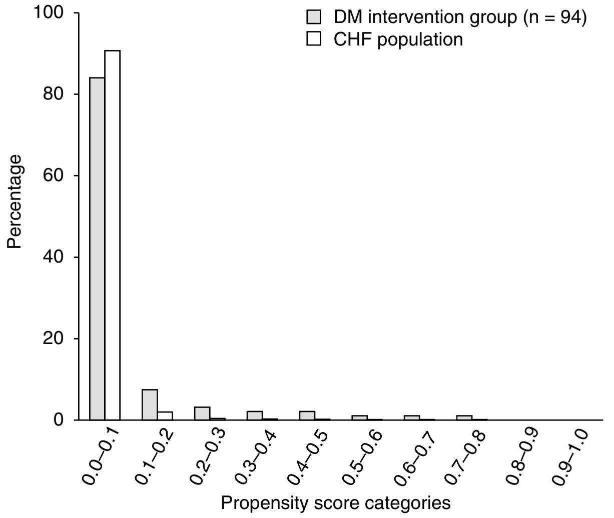

Fig. 2. Propensity score distribution for the congestive heart failure (CHF) disease management (DM) program participants and the CHF population from which they were drawn.

Hospitalizations, ED visits, and total costs were expressed as per member per year rates. Hospitalizations and ED visits were included only if they were CHF specific. Total costs included all associated costs per member, disease and non-disease related, excluding pharmacy costs (this exclusion was a result of a pharmacy benefit not being available to all members, making comparative analyses difficult).

Construction of a Comparison Group

Once propensity scores are estimated by logistic regression, it is helpful to review the overlap in the distribution of scores for both cohorts. Figure 2 illustrates that there was an overlap in propensity scores between the DM program cohort and the total CHF population, suggesting that for every DM program participant there was a non-program participant who was comparable and could be matched to. As shown, the majority of scores fell into the lowest category (0.0–0.1). This is not surprising as the intervention group is expected to mirror the population from whence they were drawn, assuming a random selection from that population. However, in some settings the analyst may discover that the DM program participants and non-participants are so different that no sensible matching is possible. This would be illustrated by a less pronounced distribution of values spanning over more of the scoring categories.

The process used for nearest available matching is straightforward. Each DM program participant is matched with a control having the nearest propensity score and then both subjects are removed from the data set. This procedure is repeated until all DM program participants have been matched. To reduce the possibility of selection bias, all controls with identical propensity scores were assigned a random order within their numeric frequency strata. Thus, each DM program participant was matched with a randomly determined control having the nearest propensity score.

As illustrated in figure 3, each matched pair fit very closely, with little or no departure at any point along the continuum. This visual display is a good confirmation of the viability of this method for this particular data set. One very important point to remember is that logistic regression modeling is used as a means for creating the propensity score (i.e. to reduce the multiple covariates into one score for ultimate pairing with a control, while adjusting for all confounding variables).[12-15,20,21] As such, the correct modeling of the regression is less important than including all the relevant predictors of group membership.[22-24] In fact, a regression model that achieves a high goodness of fit (i.e. correctly identifies program participants and non-participants) has a lower likelihood of finding close matches between the two groups, since non-participants will score close to 0 and program participants will score close to 1. Therefore, the evaluator should assess goodness of fit by reviewing graphic displays (i.e. figure 2 and figure 3), and statistical comparisons between cases and controls on baseline characteristics (table II).

Results

Table II presents the pre- and post-program characteristics of the DM program participants and their matched controls, drawn from the CHF population. As indicated by the non-significant p-values, both cohorts were closely matched on their pre-program characteristics at baseline. Conversely, as pointed out in the last three rows of table II and illustrated in figure 4, the DM program participants showed a significant reduction in utilization and costs as compared with their controls who, in fact, demonstrated an increase in healthcare utilization and cost during the program year.

How did the propensity scoring method change our estimate of the program's effectiveness? A naive calculation might take the difference between the post-program total costs for the DM group (n = 94) and all the non-DM members (n = 4606) and estimate an average cost savings of $US4162 (standard error [SE] = $US1456). However, most analysts would be concerned with the large pre-period differences between the DM and non-DM groups. Noting that the costs in the DM group decreased over time and the costs in the non-DM group increased, summing the pre-post changes of each group to get an effect estimate of $US13 475 would be another approach. The propensity scoring method adjusts for most of the pre-period differences between DM and matched comparison groups; therefore, in the CHF example, the program effectiveness estimate would be $US12 221 (SE = $US4093) [the difference between the total costs for matched controls and cases]. This figure appropriately rewards the program for treating more expensive cases and the SE gives a more honest assessment of the uncertainties. The results for admissions and ED visits are similar.

Subclassification on the Propensity Score

The propensity score was used to stratify the 4700 individuals comprising the total data set into quintiles. The lowest quintile (I) represented the subclass least likely to be enrolled in the DM program and the highest quintile (V) represented the subclass most likely to be enrolled in the DM program. This stratification is often used as another approach to developing program effect estimates. But even when the method of choice is matching, the examination of the quintiles provides valuable diagnostic information as to the success and appropriateness of propensity scoring. As illustrated in table III, the distribution of covariates was similar between both cohorts within each subclass (as indicated by non-significant differences of the covariates in each stratum). It has been shown that this method of subclassification can remove >90% of the initial bias as a result of the covariates used to create the propensity score.[25,26] If important within-subclass differences between cohorts had been found on some covariates, it could have been concluded that the covariate distributions did not overlap sufficiently to allow subclassification to adjust for these covariates, thereby raising concern about the model's ability to draw valid conclusions about the results.[26]

In reviewing the outcomes based on propensity score subclasses (the post-program results in table III), it is evident that members of quintiles IV and V (whose characteristics make them more likely to be enrolled in the DM program) have lower hospitalization rates, ED visit rates, and costs than their counterparts not enrolled in the program. This supports the data presented in table II and figure 4.

Limitations of the Propensity Scoring Method

There are several limitations to the propensity scoring method. Propensity scores are based solely on observable confounding variables and not for unknown or 'hidden' sources of variation. Randomization controls for selection bias, by distributing the

Table III. Subclassification by propensity score quintiles on pre- and post-program variables for the congestive heart failure (CHF) disease management (DM) intervention group and the CHF population from which they were drawn. Values are means (standard errors) [$US, 2003 values]

Variable	Quintile I		Quintile II		Quintile III		Quintile IV		Quintile V	
	CHF population (n = 940)	DM group (n = 0)	CHF population (n = 938)	DM group (n = 2)	CHF population (n = 933)	DM group (n = 7)	CHF population (n = 905)	DM group (n = 15)	CHF population (n = 890)	DM group (n = 70)
Age (years)	73.0 (0.47)	N/A	77.3 (0.43)	73.5 (4.5)	78.0 (0.37)	78.2 (3.1)	76.4 (0.41)	75.1 (2.0)	78.3 (0.37)	78.0 (1.2)
Sex[a]	0.81 (0.01)	N/A	0.39 (0.02)	0.50 (0.50)	0.57 (0.02)	0.58 (0.20)	0.53 (0.02)	0.60 (0.13)	0.49 (0.02)	0.49 (0.06)
Resident of Portland, OR, USA[b]	0.93 (0.01)	N/A	0.99 (0.002)	1.0 (0)	0.99 (0.30)	1.0 (0)	0.46 (0.02)	0.47 (0.13)	0.02 (0.004)	0 (0)
Health risk[c]	0.25 (0.01)	N/A	0.02 (0.004)	0 (0)	0.74 (0.01)	0.43 (0.20)	0.40 (0.02)	0.47 (0.13)	0.60 (0.02)	0.59 (0.06)
Pre-program (per member per year)										
Admission rate[d]	0.10 (0.01)	N/A	0.17 (0.01)	0.50 (0.50)	0.41 (0.02)	0.71 (0.29)	0.98 (0.05)	1.53 (0.49)	0.86 (0.05)	1.10 (0.16)
ED visit rate[d]	0.11 (0.01)	N/A	0.33 (0.02)	0 (0)	0.41 (0.03)	0.86 (0.46)	0.56 (0.04)	0.40 (0.21)	0.59 (0.04)	0.77 (0.13)
Total costs[e]	2851 (279)	N/A	4414 (228)	10 052 (36)	8880 (359)	16 885 (6426)	14 438 (708)	24 832 (9695)	14 798 (918)	17 280 (1732)
Post-program (per member per year)										
Admission rate[d]	0.69 (0.04)	N/A	0.56 (0.04)	0.50 (0.50)	1.05 (0.37)	0.57 (0.05)	0.97 (0.05)	0.20 (0.14)[f]	1.09 (0.06)	0.67 (0.13)[g]
ED visit rate[d]	0.60 (0.04)	N/A	0.43 (0.03)	0 (0)	0.59 (0.04)	1.29 (0.71)	0.64 (0.04)	0.60 (0.21)	0.86 (0.04)	0.51 (0.11)[h]
Total costs[e]	13 298 (655)	N/A	10 608 (563)	9689 (6816)	19 561 (899)	8573 (2904)	17 380 (865)	7344 (1802)[f]	19 637 (1060)	13 237 (1802)[g]

a A score of 1 indicates women and 0 indicates men.

b A score of 1 indicates resident within Portland and 0 indicates resident outside of Portland.

c A score of 1 indicates high risk and 0 indicates low risk for future CHF-related claims.

d Hospitalizations and ED visits were included only if they were CHF specific.

e Total costs included all associated costs per member, disease and non-disease related, excluding pharmacy costs (this exclusion was a result of a pharmacy benefit not being available to all members, making comparative analyses difficult).

f $p < 0.0001$, using two-tailed t-test of independent samples (standard error in parentheses).

g $p < 0.005$.

h $p < 0.05$.

ED = emergency department.; **N/A** = not applicable.

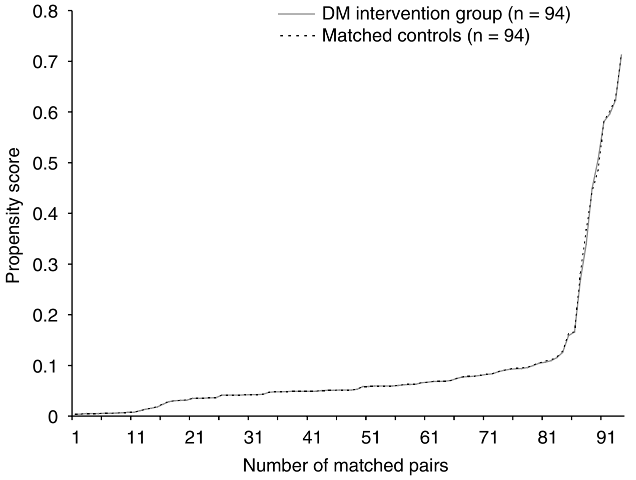

Fig. 3. Comparison of propensity scores for congestive heart failure disease management (DM) program participants and their matched controls (n = 94 matched pairs).

unobserved variation equally among the groups. Since observational studies cannot control for hidden bias, there is limited confidence in drawing causal inferences about conclusions reached in the study. This may be significant in a DM program evaluation where many unobservable confounders may impact the use of services,[27,28] which are the outcomes most typically analyzed in DM programs.

In light of these concerns, methods have been developed to estimate the magnitude of a hidden bias that would be needed to invalidate the study findings.[29,30] In case-control studies in which patients are matched on the propensity score, the sensitivity analysis provides the estimated odds of patients assigned to the program intervention having this hidden bias. Using the present data as an illustration, the results of the sensitivity analysis suggest that DM program participants would need to be 1.58 times more likely to possess hidden traits or factors than their matched controls to change our conclusion that the program intervention led to significantly lower hospitalizations. This value indicates that the study results are relatively insensitive to small amounts of bias and require moderately high levels of bias to alter our conclusions that the reduction in hospitalizations was indeed an outcome of the program intervention and not a function of hidden bias. Similar levels of insensitivity to hidden bias were estimated for total costs (odds ratio = 1.28), whereas ED visits showed a higher level of sensitivity to hidden bias (odds ratio = 1.15). These results suggest that we can be somewhat more confident that the resultant lower hospitalization rates and decreased total costs were caused by the program intervention than the lowered ED visit rate. Therefore, ED visit data should be further scrutinized to assess whether a causal link with the intervention can be identified. For a more comprehensive discussion on the topic of estimating hidden bias, the reader is referred to Linden et al.[29]

Adequacy of the non-participant pool available for matching can be a second limitation of the propensity scoring method. Many DM programs have moved to an ‘opt-out’ method of program enrollment, meaning that identified members are considered participants unless they specifically request to be excluded from the program. This method typically yields very high enrollment levels, typically >95% (or a ≤5% opt-out rate). This limitation can be overcome in several ways. Firstly, only members actively participating in program interventions (such as those engaged in the telephonic nurse coaching component) could be included in the participant group. Others ‘enrolled’ in the program but not actively engaged in high-intensity interventions are considered non-participants. Alternatively, a historical (rather than concurrent) control group could be developed. If using a historical matching method, outcomes related to cost must be appropriately adjusted for inflation. Finally, non-participants could be drawn from population subsets to which the DM program was not offered (e.g. because of geographic location, physician affiliation, and/or benefits design).

The third limitation to the propensity score technique for use in DM program evaluation is that sufficiently large samples are required. This is especially true when using subclassification. As noted in table III, most subclasses had extremely small numbers of DM program participants (the lowest quintile had no DM program participants at all). This leads to great variability in the covariate distribution, thus limiting the ability to draw conclusion about the results using the subclass method.

A fourth limitation has to do with the use of administrative claims data for the determination and estimation of covariates and propensity scores. While claims data are notorious for their lack of

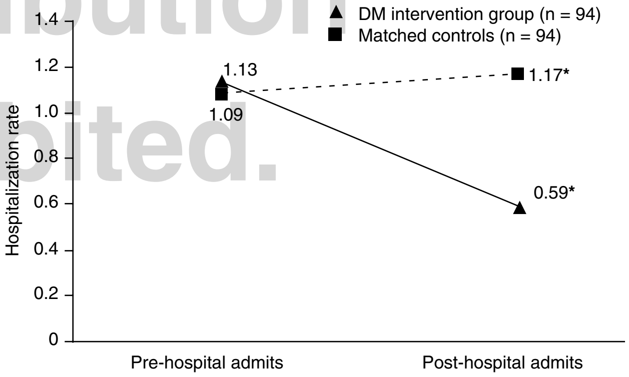

Fig. 4. A comparison of admission rates between the congestive heart failure disease management (DM) program participants and their matched controls. All patients were continuously enrolled with the health plan for at least 2 years (1 year prior to program commencement and the duration of the first program year). p-Values are two-tailed t-tests for dependent samples. * p = 0.005.

Table II. A comparison of pre- and post-first-year program characteristics of the congestive heart failure (CHF) disease management (DM) intervention group and matched controls drawn from the CHF population. Values are means (standard errors) [$US, 2003 values]

Variable	DM intervention group (n = 94)	Matched controls (n = 94)	p-Value[a]
Age (years)	77.4 (0.96)	78.2 (0.98)	0.556
Sex[b]	0.51 (0.05)	0.51 (0.05)	1.000
Resident of Portland, OR, USA[c]	0.17 (0.04)	0.17 (0.04)	1.000
Health risk[d]	0.54 (0.05)	0.60 (0.05)	0.379
Pre-program (per member per year)			
Admission rate[e]	1.13 (0.15)	1.09 (0.15)	0.841
ED visit rate[e]	0.70 (0.11)	0.67 (0.10)	0.832
Total costs[f]	$18 287 ($2053)	$17 001 ($2449)	0.688
Post-program (per member per year)			
Admission rate[e]	0.59 (0.10)	1.17 (0.18)	0.005
ED visit rate[e]	0.57 (0.08)	0.77 (0.10)	0.048
Total costs[f]	$11 874 ($1408)	$24 085 ($3843)	0.003

a p-Values are two-tailed t-tests for independent samples.

b A score of 1 indicates women and 0 indicates men.

c A score of 1 indicates resident within Portland and 0 indicates resident outside of Portland.

d A score of 1 indicates high risk and 0 indicates low risk for future CHF-related claims.

e Hospitalizations and ED visits were included only if they were CHF specific.

f Total costs included all associated costs per member, disease and non-disease related, excluding pharmacy costs (this exclusion was a result of a pharmacy benefit not being available to all members, making comparative analyses difficult).

ED = emergency department.

accuracy,[31-33] currently this is the most widely available source of data at our disposal. Therefore, special precautions must be taken to ensure data integrity and accuracy, such as treatment of missing values, accuracy of coding, identification of patients across databases, etc.

A final constraint, somewhat related to the previous one, pertains to the limited number of variables actually available to DM program evaluators for estimating propensity scores and to assess outcomes. The more variables available for use in estimating the propensity score, the more likely that a good fitting model can be developed while concomitantly reducing the number of unobserved covariates.

Conclusions

Using the first-year results of a CHF DM program as an example, this paper has described in some detail the application of propensity scoring as a technique to assist in the evaluation of DM program effectiveness. This method is particularly suitable to DM program analysis (where use of randomized control groups are generally not practical) because adding matched controls may reduce many of the biases typically inherent to observational studies, most notably selection bias and regression to the mean. This is especially important because DM program participants are generally chosen based on high cost and utilization in the baseline period. Therefore, they are fundamentally different in their baseline characteristics from the population from which they were drawn.

One significant advantage of the propensity scoring method is that it can identify whether a particular data set can address the question of whether a causal relationship exists between program participation (e.g. vis-à-vis the covariates predicting program participation) and outcomes. Subclassification is an additional tool to validate the model by eliminating up to 90% of the bias associated with the covariates used to estimate the propensity score. An important fact to keep in mind when using propensity scoring is that it can only adjust for the observed variation. Therefore, any discussion on the results achieved through the analysis must note that the magnitude of the unobserved variation remains a potential threat to validity of the findings. Nonetheless, because of its simplicity and utility, propensity scoring should be considered as an alternative procedure for use with current non-experimental designs in evaluating DM program effectiveness.

Acknowledgments

No sources of funding were used to assist in the preparation of this study. The authors have no conflicts of interest that are directly relevant to the content of this study.

References

1. Kleijen J, Gotzche P, Kunz RA, et al. So what's so special about randomization? In: Maynard A, Chalmers I, editors. Non-random reflections on health services research. London: BMJ Publishing, 1997: 93-106
2. D'Arcy Hart P. Early controlled clinical trials. BMJ 1996; 312 (2): 769-82
3. Linden A, Adams J, Roberts N. An assessment of the total population approach for evaluating disease management program effectiveness. Dis Manag 2003; 6 (2): 93-102
4. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: an introduction to time series analysis. Dis Manag 2003; 6 (4): 243-55
5. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: an introduction to survival analysis. Dis Manag. 2004; 7 (3): 180-190
6. Linden A, Adams J, Roberts N. Evaluation methods in disease management: determining program effectiveness. Position Paper for the Disease Management Association of America (DMAA). 2003 Oct
7. Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research. Chicago (IL): Rand McNally, 1966
8. Cook TD, Campbell DT. Quasi-experimentation: design and analysis issues for field settings. Chicago (IL): Rand McNally College Publishing Company, 1979
9. Shadish SR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston (MA): Houghton Mifflin, 2002
10. American Healthways and the John Hopkins Consensus Conference. Consensus report: standard outcome metrics and evaluation methodology for disease management programs. Dis Manag 2003; 6 (3): 121-38
11. Dehejia RH, Wahba S. Propensity score-matching methods for non-experimental causal studies. Rev Econ and Stats 2002; 84: 151-61
12. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70: 41-55
13. Rosenbaum P, Rubin D. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 1985; 39: 33-8
14. Rubin D. Estimating causal effects of treatments in randomized and non-randomized studies. J Educ Psychol 1974; 66: 688-701
15. Rubin D. Assignment to treatment group on the basis of a covariate. J Educ Stats 1977; 2: 1-26
16. Cox DR. The analysis of binary data. London: Methuen, 1970
17. Cox DR. The analysis of multivariate binary data. Appl Stat 1972; 21: 113-20
18. Dehejia RH, Wahba S. Causal effects in nonexperimental studies: reevaluating the evaluation of training studies. J Am Stat Assoc 1999; 94: 1053-62
19. Heckman J, Ichimura J, Todd P. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Rev Econ Stud 1997; 64: 605-54
20. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. Biometrics 1996; 52: 249-64
21. Rubin DB. Estimating causal effect from large data sets using propensity scores. Ann Intern Med 1997; 127: 757-63
22. Canner P. How much data should be collected in clinical trials? Stat Med 1984; 3: 423-32
23. Canner P. Covariate adjustment of treatment effects in clinical trials. Control Clin Trials 1991; 12: 359-66
24. Drake C. Effects of misspecification of the propensity score on estimators of treatment effects. Biometrics 1993; 49: 1231-6
25. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 1968; 24: 205-13
26. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 1984; 79: 516-24
27. Aday L, Andersen RM. Equity in access to medical care: realized and potential. Med Care 1981; 19 (12 Suppl.): 4-27
28. Andersen RM. Behavioral model of families: use of health services. Research Series No. 25. Chicago (IL): Center for Health Administration Studies, University of Chicago, 1968
29. Linden A, Adams J, Roberts N. Strengthening the case for disease management effectiveness: unhiding the hidden bias. Am J Eval. In press
30. Rosenbaum P. Sensitivity analysis for certain permutation tests in matched observational studies. Biometrika 1987; 74: 13-26
31. Fisher ES, Barton JA, Malenka DJ, et al. Overcoming potential pitfalls in the Medicare data for epidemiologic research. Am J Public Health 1990; 80: 533-46
32. Romano PS, Mark DH. Bias in the coding of hospital discharge data and its implication for quality assessment. Med Care 1994; 32: 81-90
33. Roos LL, Sharp SM, Cohen MM. Comparing clinical information with claims data: some similarities and differences. J Clin Epidemiol 1991; 44: 881-8

About the Author: Dr Linden is a health services researcher whose focus is on evaluating disease management program effectiveness. In this area alone, he has recently had 17 manuscripts published, including a position paper for the Disease Management Association of America. He is President of the Linden Consulting Group and Clinical Associate Professor in the School of Medicine at Oregon Health and Science University, USA.

Correspondence and offprints: Dr *Ariel Linden*, Linden Consulting Group, Hillsboro, 6208 NE Chestnut Street, OR 97124, USA.
E-mail: alinden@lindenconsulting.org