Journal of Evaluation in Clinical Practice ISSN 1356-1294

Improving participant selection in disease management programmes: insights gained from propensity score stratification

Ariel Linden DrPH MS[1] and John L. Adams PhD[2]

[1]President, Linden Consulting Group, Portland, OR, USA and Clinical Assistant Professor, Oregon Health & Science University, School of Medicine, Portland, OR, USA
[2]Senior Statistician, RAND Corporation, Santa Monica, CA, USA

Keywords
disease management, observational study, design, propensity score, stratification

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

Accepted for publication: 31 July 2008

doi:10.1111/j.1365-2753.2008.01091.x

Abstract

While the randomized controlled trial (RCT) remains the gold-standard study design for evaluating treatment effect, outcomes researchers turn to powerful quasi-experimental designs when only observational studies can be conducted. Within these designs, propensity score matching is one of the most popular to evaluate disease management (DM) programme effectiveness. Given that DM programmes generally have a much smaller number of participants than non-participants in the population, propensity score matching will typically result in all or nearly all participants finding successful matches, while most of the non-participants in the population remain unmatched and thereby excluded from the analysis. By excluding data from the unmatched population, the effect of non-treatment in the remaining population with the disease is not captured. In the present study, we examine changes in hospitalization rates stratified by propensity score quintiles across the entire population allowing us to gain insight as to how well the programme chose its participants, or if the programme could have been effective on those individuals not explicitly targeted for the intervention. These data indicate the presence of regression to the mean, and suggest that the DM programme may be overly limited to only the highest strata when there is evidence of a potential benefit for those in all the lower strata as well.

Introduction

The strength of the randomized controlled trial (RCT) design lies in the ability to draw causal inferences about treatment effectiveness because individuals in treatment and control groups are assumed to be unconditionally *exchangeable* [1]. This is because random assignment distributes both known and unknown sources of variability equally between groups. While the RCT remains the gold-standard study design for evaluating treatment effect, rigorous evaluations of disease management (DM) programmes have historically been limited to academic settings or demonstration projects in the US Medicare/Medicaid populations. The typical study design used to evaluate commercial DM programmes in the United States (an industry with projected annual revenues of $1.8 billion in 2008) [2] is a simple pretest–posttest design that compares outcomes of the chronically diseased population in the baseline year to each year thereafter [3,4]. This study design suffers from several sources of bias and confounding factors that offer plausible alternative explanations for any observed treatment effect outside the DM programme intervention [5–7].

In DM, the primary selling point (and thus outcome of interest) is the return on investment (ROI). As a result, the approach to evaluation is often a controversial issue. While DM firms report ROIs from 2:1 to 8:1 based on the pre–post design [8–10], these conflict with reviews and meta-analyses in the peer-reviewed literature that are based on RCTs and have concluded that DM programmes do not reduce the overall cost of health care [11–14]. As purchasers of commercial DM services have realized this discrepancy, there has been an increased demand for rigorous evaluation of these programmes. As RCTs are often infeasible, research-trained programme evaluators turn to powerful quasi-experimental designs available to estimate treatment effects in observational studies. Several tutorials have been written on their application to DM [15–23].

Within these designs, propensity score matching is one of the most popular to evaluate DM programme effectiveness. The propensity score, the probability of assignment to the treatment group conditional on covariates [24] (i.e. independent variables), controls for pre-intervention differences between enrolled and non-enrolled groups. The underlying assumption for using the

propensity score in DM is that enrollment in the programme is associated with observable pre-programme variables (e.g. age, sex, utilization, cost). Propensity scores are derived from a logistic regression equation that reduces each participant's set of covariates to a single score. It has been demonstrated that, conditional on this score, all observed pretreatment covariates can be considered independent of group assignment, and in large samples, covariates will be distributed equally in both groups and will not confound estimated treatment effects [24]. After the propensity score is estimated, treatment effects can be modelled using matching, stratification, weighting and/or regression adjustment (see references 18, 25 and 26 for a comprehensive discussion on these methods).

While matching appears to be the most popular propensity scoring technique for evaluating DM programme effectiveness, there are some inherent limitations. DM programmes generally have a much smaller number of participants than non-participants in the population. Thus, propensity score matching will typically result in all or nearly all participants finding successful matches, while most of the non-participants in the population remain unmatched and thereby excluded from the analysis. There are two consequences of this: (1) treatment effects may be statistically insignificant because of the confluence of a small treatment group and a rare-event outcome (e.g. hospital admissions, emergency department visits); and (2) by excluding data from the unmatched population, the effect of non-treatment in the remaining population with the disease is not captured. Thus, we gain no insight as to how well the programme chose its participants, or if the programme could have been effective on those individuals not explicitly targeted for the intervention.

Stratification on the propensity score holds some advantages over matching or regression adjustments. It is a straightforward, transparent approach that is much easier to interpret than matching or regression techniques that require a deeper level of understanding of statistics. The approach generally taken to stratify the population under study is to simply arrange the outcomes into quintiles based on the range of propensity scores divided into treated and non-treated groups. This allows the evaluator to review outcomes between groups at each stratum, as well as to observe differences within groups between strata. This can be carried out using statistical tests, but often visual inspection alone can highlight important effects. Even when matching is used, the examination of the stratification levels provides valuable insight into the success and appropriateness of propensity scoring. It has been shown that stratification of the propensity score into quintiles (generally referred to as subclassification) can remove over 90% of the initial bias because of the choice of covariates used to create the propensity score [25,26]. If important within-subclass differences between cohorts are found on some covariates, it could be concluded that the covariate distributions did not overlap sufficiently to allow subclassification to adjust for these covariates, raising concern about the model's ability to draw valid conclusions about the results [26]. In such cases, alternate analytic adjustments should be considered.

In this paper, we demonstrate how a propensity score stratification table can provide valuable insights about the entire population from where programme participants are drawn. This approach can assist a programme determine if they are recruiting individuals for which the programme can be most effective, and how effective the intervention is across different strata. We estimate the change in hospital admission rates between participants and non-participants of a DM programme for congestive heart failure (CHF), and then assess these results relative to the distribution of hospitalization rate changes by quintile.

Methods

Study population and outcome measure

Our data come from a previously published study evaluating the effectiveness of a CHF DM programme in the first year of operation at a medium-sized health plan in Oregon [18]. The study population included health plan members with CHF who were continuously enrolled in the health plan for the year prior to programme initiation as well as for the entire programme 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. There were 94 participants and 4606 non-participants that met these criteria.

The outcome variable, hospital admission rate, is calculated as a difference-in-difference (DID) estimator. That is, we model the treatment effect by estimating the difference between hospitalization rate in the programme year minus the baseline year for both participants and non-participants (or matched controls, depending on the analysis) and then compare the difference between the two groups (treated as panel data) using an ordinary least squares regression model. A negative value for the DID estimate indicates that the programme participant group exhibited a reduction in admissions greater than the non-participant group, and a positive value indicates that the non-participants had a greater reduction in admissions than participants. Using this approach, the outcome measure can be described as the net change in admissions rate of participants over non-participants. The DID strategy ensures that any unobserved variables that remain constant over time, and are correlated with the participation decision and the outcome variable, will not bias the estimated effect [27].

Propensity score model

Observed covariates used to estimate the propensity score were baseline age, gender, health plan service area covered, number of hospitalizations in the prior 12 months, number of emergency department visits in the prior 12 months, total health care costs (in 2003 dollars) in the prior 12 months and health-risk level (see the study by Linden *et al*. [18] for a more comprehensive description of this propensity score model).

This score was then used to: (1) match each participant to a non-participant based on the nearest propensity score (thus creating 94 pairs) [18]; and (2) stratify the entire population ($n = 4700$) into propensity score quintiles.

Results

Table 1 presents both pre- and post-programme characteristics of the CHF programme participants ($n = 94$) compared with the entire CHF population ($n = 4606$) from which they were drawn as well as characteristics of the propensity score-matched controls ($n = 94$). There were significant differences between programme

Table 1 A comparison of pre- and post-first-year programme characteristics of the congestive heart failure (CHF) disease management intervention group, the CHF population from which they were drawn and propensity score-matched controls

	Intervention group ($n = 94$)	CHF population ($n = 4606$)	Matched controls ($n = 94$)	*P* value* (Int. vs. Pop.)	*P* value† (Int. vs. matches)
Age (years)	77.4 (0.96)	76.6 (0.19)	78.2 (0.98)	0.539	0.556
Gender‡	0.51 (0.05)	0.56 (0.01)	0.51 (0.05)	0.336	1.00
Resident of Portland, OR§	0.17 (0.04)	0.69 (0.01)	0.17 (0.04)	<0.0001	1.00
Health risk¶	0.54 (0.05)	0.40 (0.007)	0.60 (0.05)	<0.0001	0.379
Pre-programme (per member per year)					
Admission rate**	1.13 (0.15)	0.50 (0.02)	1.09 (0.15)	<0.0001	0.841
Emergency department rate**	0.70 (0.11)	0.40 (0.01)	0.67 (0.10)	0.003	0.832
Total costs††	$18 287 ($2053)	$8974 ($257)	$17 001 ($2449)	<0.0001	0.688
First-programme year (per member per year)					
Admission rate**	0.59 (0.10)	0.87 (0.02)	1.17 (0.18)	0.0008	0.005
Emergency department rate**	0.57 (0.08)	0.58 (0.02)	0.77 (0.10)	0.1874	0.048
Total costs††	$11 874 ($1408)	$16 036 ($370)	$24 085 ($3843)	0.005	0.003

Values are means (standard errors) ($US, 2003 values) [18].

*Two-tailed *t*-tests for independent samples (intervention group vs. CHF population).

†Two-tailed *t*-tests for dependent samples (intervention group vs. matched controls).

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

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

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

**Hospitalizations and emergency department 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.

participants and the CHF population in both their geographic location and pre-programme utilization and costs. The control group, established by the propensity score, matched up well with the programme participants, as indicated by the lack of significant differences between the groups on any baseline characteristics. The propensity score-matched pairs treatment effect estimate was –0.60 (95% CI = –1.11, –0.10, $P = 0.02$). In a DID model this estimate suggests that the programme group decreased participant admissions rate by 0.60 per person per year over the matched control group.

Table 2 illustrates the pre-programme admissions rates, the programme year admissions rates and the difference scores across the entire study population ($n = 4700$) stratified by propensity score quintiles. These data highlight several important points. First, the regression to the mean effect [6] is quite evident. Four of the five strata showed rather large increases in admissions for those individuals with CHF who did not participate in the programme (the exception was quintile IV where the admission rate was flat). Conversely, participants generally exhibited large decreases in admissions. Another important observation is that 90 of the 94 participants were located in the two highest quintiles (with no representation in the lowest quintile and only two participants appearing in quintile II). This indicates that those targeted for programme enrollment do not mirror the characteristics of the entire population, but only of a narrow subset. The ramifications of this will be discussed in the next section.

The data presented here are typical of a population in which a DM programme is implemented. That is, only a small subset of individuals who are considered 'high risk' was invited to participate in the programme, leaving the vast majority of the population untreated. Given that the classification of risk was based on prior acute utilization levels, the regression to the mean effect would suggest that those individuals who experienced a hospitalization in the prior period were not likely to experience another hospitalization in the following period, and vice versa [6]. As expressed in Table 2, this was indeed the case, with the 94 participants exhibiting decreased admissions and the 4606 non-participants showing increased admission rates across nearly all strata.

Discussion

The results of this study raise two different but interrelated issues. First, the current identification and enrollment strategy used by DM programmes may target the wrong individuals, given that the majority of future admissions come from individuals located in the lower propensity score quintiles. Rather than targeting individuals with high costs, it may be better to target individuals with high potential savings. Data presented here and elsewhere [27] clearly suggest that targeting high-cost individuals allows the programme to get a 'free ride' on the regression to the mean effect – those treated will show a natural decrease in admissions, while non-treated individuals in the lower strata will have higher acute utilization. This may help make the case for a treatment effect, but it does not help the non-participants who may truly benefit from the intervention. Similarly, in the bottom three quintiles there was a more marked increase in the comparison groups than for the fourth and fifth quintile. This suggests that DM programmes may be selecting participants based on their high rates rather than on their potential to decrease these rates. Our results suggest that there may be potential for targeting lower pre-programme admissions rate quintiles to produce savings as the highest rate patients may be near their maximum because of ceiling effects (while our focus in this paper has been on hospitalizations as the outcome, this logic can be readily applied to any outcome measure). Some

Table 2 Stratification by propensity score quintiles on pre- and post-programme variables for the congestive heart failure (CHF) population and disease management (DM) programme participants

	Quintile I		Quintile II		Quintile III		Quintile IV		Quintile V	
Variable	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 ($n = 925$)	DM group ($n = 15$)	CHF population ($n = 890$)	DM group ($n = 70$)
Pre-admission rate	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)
Post-admission rate	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)	1.09 (0.06)	0.67 (0.13)
Difference*	0.59 (0.04)	N/A	0.39 (0.04)	0 (0)	0.64 (0.06)	−0.14 (0.51)	0.01 (0.06)	−1.33 (0.44)	0.23 (0.07)	−0.43 (0.18)
DID†		N/A				−0.78		−1.34		−0.066
P Value					insufficient *n*	0.18		0.01		0.001

Values in parentheses are standard errors.
*first year – pre-programme year.
†DID = differences in differences = DM group (first year – pre-programme year) – CHF population (first year – pre-programme year).

experimentation with providing the DM programme to patients in these quintiles may reveal potential savings. However, success depends heavily on a programme's ability to identify these people.

Second, these data clearly indicate that the bias introduced in the DM identification and enrollment model may be difficult, if not impossible, to control. The basic assumption of conditional exchangeability is violated, and thus causal inferences cannot be made about the effectiveness of the intervention [1]. While this is generally true for any observational study, the purpose of analytic adjustment is to increase our confidence that these biases have been controlled for, and the results are likely to be causally associated with the intervention.

Outside of conducting an RCT, there are at least three areas where the DM programme design can be strengthened to reduce bias. First, it is evident that a better identification and enrollment model is needed [28]. Individuals thought to be 'low risk' based on little or no past acute utilization are in fact quite likely to be hospitalized in the next observation period. In improving this process, not only would the programme be providing beneficial care to the individuals who need it most, but the population of participants and non-participants will be more balanced across the spectrum of baseline characteristics. Thus, we would also expect to see more homogeneity in the treatment estimates across strata. Another strategy is to employ the regression-discontinuity design, which utilizes a 'cutoff' score on a pre-programme measure or test to determine who will be assigned to the intervention or control group. The defining characteristic of the RD design is in identifying whether a difference is found in the relationship between the assignment variable and outcome occurring exactly at that cutoff score, where individuals in the treatment and non-treatment groups are most similar [22]. Third, it is apparent that larger sample sizes are needed in the treatment group to make reasonable comparisons against a much larger untreated population, especially when the outcome is a rare event, as is the case with hospitalization rates [29]. This further supports a broader enrollment strategy targeting individuals in all strata that can benefit from the programme.

Absent randomization, the 'true' programme effect cannot be known. At best, every effort to minimize bias must be considered and the appropriate evaluation model be employed. Depending on the existing source of bias, various analytic techniques may elicit different results. Therefore, a reasonable recommendation is to run several different types of analyses on the data and then present the range of estimates and their respective confidence intervals. To this end, some form of propensity score stratification should be included. These data serve as a sensitivity analysis and test of plausibility for the study outcomes. In the present study, the subclassification method allowed for a more comprehensive view of the entire CHF population, and suggested that the DM intervention may be overly limiting the intervention to only the highest strata when there is evidence of a potential benefit for all the lower strata as well.

Conclusion

Disease management programmes are implemented in such a way that selection bias and regression to the mean are two major threats to the validity of study outcomes. Programme evaluators attempt to control for these biases by using various statistical adjustments. Stratification by propensity score quintiles is a simple technique that allows the evaluator to assess the impact of the intervention

across the entire population, not only in the narrowly populated highest stratum. For a DM programme, this view of the data may indicate whether the identification and enrollment strategy targeted the correct individuals and support or disprove a programme effect. Outside an RCT, causal inferences about DM programme effectiveness cannot be readily made. At best, every attempt must be made to reduce the introduction of bias into the programme design, a robust evaluation strategy employed, and outcome estimates be scrutinized accordingly.

Acknowledgement

The authors would like to thank Julia Adler-Milstein from the PhD programme in Health Policy at Harvard University for her invaluable assistance in editing the manuscript.

References

1. Hernán, M. A. & Robins, J. M. (2006) Estimating causal effects from epidemiologic data. *Journal of Epidemiology and Community Health*, 60, 578–586.
2. Matheson, D., Wilkins, A. & Psacharopoulos, D. (2006) Realizing the Promise of Disease Management: Payer Trends and Opportunities in the United States. Boston, MA: Boston Consulting Group.
3. Disease Management Association of America (2006) Outcomes Guidelines Report. Washington, DC: DMAA.
4. Disease Management Association of America (2007) Outcomes Guidelines Report – Volume II. Washington, DC: DMAA.
5. Linden, A., Adams, J. & Roberts, N. (2003) An assessment of the total population approach for evaluating disease management program effectiveness. *Disease Management*, 6 (2), 93–102.
6. Linden, A. (2007) Estimating the effect of regression to the mean in health management programs. *Disease Management and Health Outcomes*, 15 (1), 7–12.
7. Linden, A. (2007) Use of the total population approach to measure U.S. disease management industry's cost savings: issues and implications. *Disease Management and Health Outcomes*, 15 (1), 13–18.
8. Johnson, A. (2003) Disease management: the programs and the promise. Milliman USA Research Report. Available at: http://www.milliman.com/expertise/healthcare/publications/rr/pdfs/Disease-Mangement-Programs-Promise-RR05-01-03.pdf (last accessed 5 May 2008).
9. Shutan, B. (2004) The DM Rx: disease management programs producing fast and meaningful outcomes, impressive ROI. *Employee Benefit News*; 18 (13). Available at: http://www.matria.com/resources/articles/dm/EBNPetit.pdf (last accessed 5 May 2008).
10. Health Information Designs, Inc. (2008) *HID provides a disease management program to address the specialized needs of high risk patients*. Available at: http://www.hidinc.com/WhatWeDo/whatwedo_utilizingexperiences.htm (last accessed 18 August 2008).
11. Congressional Budget Office (2004) An Analysis of the Literature on Disease Management Programs. Washington, DC: Congressional Budget Office. Available at: http://www.cbo.gov/showdoc.cfm?index=5909&sequence=0 (last accessed 9 December 2006).
12. Ofman, J. J., Badamgarav, E., Henning, J. M., Knight, K., Gano, A. D. Jr, Levan, R. K., Gur-Arie, S., Richards, M. S., Hasselblad, V. & Weingarten, S. R. (2004) Does disease management improve clinical and economic outcomes in patients with chronic diseases? A systematic review. *American Journal of Medicine*, 117 (3), 182–192.
13. Goetzel, R. Z., Ozminkowski, R. J., Villagra, V. G. & Duffy, J. (2005) Return on investment on disease management: a review. *Health Care Finance Review*, 26, 1–19.
14. Mattke, S., Seid, M. & Ma, S. (2007) Evidence for the effect of disease management: is $1 billion a year a good investment? *American Journal of Managed Care*, 13 (12), 670–676.
15. Linden, A., Adams, J. & Roberts, N. (2003) Evaluating disease management program effectiveness: an introduction to time series analysis. *Disease Management*, 6 (4), 243–255.
16. Linden, A., Adams, J. & Roberts, N. (2004) Evaluating disease management program effectiveness: an introduction to survival analysis. *Disease Management*, 7 (3), 180–190.
17. Linden, A., Adams, J. & Roberts, N. (2004) Evaluating disease management program effectiveness adjusting for enrollment (tenure) and seasonality. *Research in Healthcare Financial Management*, 9 (1), 57–68.
18. Linden, A., Adams, J. & Roberts, N. (2005) Using propensity scores to construct comparable control groups for disease management program evaluation. *Disease Management and Health Outcomes*, 13 (2), 107–127.
19. Linden, A. & Adams, J. (2006) Evaluating disease management program effectiveness: an introduction to instrumental variables. *Journal of Evaluation in Clinical Practice*, 12 (2), 148–154.
20. Linden, A., Adams, J. & Roberts, N. (2006) Strengthening the case for disease management effectiveness: unhiding the hidden bias. *Journal of Evaluation in Clinical Practice*, 12 (2), 140–147.
21. Linden, A. (2006) Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *Journal of Evaluation in Clinical Practice*, 12 (2), 132–139.
22. Linden, A., Adams, J. & Roberts, N. (2006) Evaluating disease management program effectiveness: an introduction to the regression-discontinuity design. *Journal of Evaluation in Clinical Practice*, 12 (2), 124–131.
23. Linden, A., Trochim, W. M. K. & Adams, J. (2006) Evaluating program effectiveness using the regression point displacement design. *Evaluation in the Health Professions*, 29 (4), 1–17.
24. Rosenbaum, P. R. & Rubin, D. B. (1983) The central role of propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
25. Rosenbaum, P. & Rubin, D. (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistics*, 39, 33–38.
26. Rosenbaum, P. R. & Rubin, D. B. (1984) Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistics Association*, 79, 516–524.
27. Linden, A. & Goldberg, S. (2007) The case-mix of chronic illness hospitalization rates in a managed care population: implications for health management programs. *Journal of Evaluation in Clinical Practice*, 13 (6), 947–951.
28. Linden, A. & Adler-Milstein, J. (2008) Medicare disease management in policy context. *Health Care Finance Review*, 29 (3), 1–11.
29. Linden, A. (2008) Sample size in disease management program evaluation: the challenge of demonstrating a statistically significant reduction in admissions. *Disease Management*, 11 (2), 95–101.

