POPULATION HEALTH MANAGEMENT
Volume xx, Number xx, 2012

DOI: 10.1089/pop.2011.0104

Original Article

Exploring Robust Methods for Evaluating Treatment and Comparison Groups in Chronic Care Management Programs

Aaron R. Wells, PhD, Brent Hamar, DDS, MPH, Chastity Bradley, PhD, William M. Gandy, EdD, Patricia L. Harrison, MPH, James A. Sidney, MA, Carter R. Coberley, PhD, Elizabeth Y. Rula, PhD, and James E. Pope, MD

Abstract

Evaluation of chronic care management (CCM) programs is necessary to determine the behavioral, clinical, and financial value of the programs. Financial outcomes of members who are exposed to interventions (treatment group) typically are compared to those not exposed (comparison group) in a quasi-experimental study design. However, because member assignment is not randomized, outcomes reported from these designs may be biased or inefficient if study groups are not comparable or balanced prior to analysis. Two matching techniques used to achieve balanced groups are Propensity Score Matching (PSM) and Coarsened Exact Matching (CEM). Unlike PSM, CEM has been shown to yield estimates of causal (program) effects that are lowest in variance and bias for any given sample size. The objective of this case study was to provide a comprehensive comparison of these 2 matching methods within an evaluation of a CCM program administered to a large health plan during a 2-year time period. Descriptive and statistical methods were used to assess the level of balance between comparison and treatment members pre matching. Compared with PSM, CEM retained more members, achieved better balance between matched members, and resulted in a statistically insignificant Wald test statistic for group aggregation. In terms of program performance, the results showed an overall higher medical cost savings among treatment members matched using CEM compared with those matched using PSM (-$25.57 versus -$19.78, respectively). Collectively, the results suggest CEM is a viable alternative, if not the most appropriate matching method, to apply when evaluating CCM program performance. (*Population Health Management* 20xx;xx:xx–xx)

Introduction

Effective chronic care management (CCM) programs identify individuals diagnosed with a disease and help them to manage their disease(s) using interventions that monitor progression and provide education and coaching to encourage healthy behaviors. Evaluation of behavioral, clinical, and financial outcomes is necessary to determine the true value and effectiveness of a CCM program. Unfortunately, the accuracy of outcome evaluation is limited by inherent selection bias in CCM programs, and difficulties with implementing randomized controlled studies have resulted in a complex environment in which to apply statistical methods to assess CCM programs.[1] In order to gain a better understanding of how evaluation methods within the CCM field have evolved over time and, in the interim, accommodated technical setbacks, a brief history of CCM research designs and recent methodological applications are presented.

Historical CCM study designs

Based on the Care Continuum Alliance (CCA) guidelines, incorporation of a valid study design is essential to (1) ascertain the value of disease management (DM) programs (referenced as CCM in this article) in achieving favorable outcomes for populations with chronic disease, (2) further improve the delivery of DM services and foster competition among industry participants based on objective criteria, (3) meaningfully advance the delivery of health care services through the reporting and dissemination of interventions

Center for Health Research, Healthways, Inc., Franklin, Tennessee.

that reduce the burden of disease, and (4) help assess which health plans or delivery systems are providing higher quality care at a more affordable cost.[2]

The CCA recommends the use of a randomized controlled study design to assess causality of DM in achieving outcomes.[1] Randomized controlled trials (RCT), when properly designed, offer one of the most rigorous study designs and are considered to be the "gold standard" in other industries. When correctly performed, the random assignment of members to the treatment or control group will provide an equal distribution of the unmeasured confounding variables, limiting study bias. Unlike other fields of study, however, RCTs in the CCM field are often difficult to implement. Ethical issues regarding random assignment of individuals to a CCM intervention group or control group remain a primary concern, particularly because purchasers rarely wish to deny treatment interventions to a subgroup of eligible members when they believe the treatment intervention is beneficial. Additionally, random assignment does not necessarily guarantee equivalent groups, as evidenced in the randomized block design employed in the recent Medicare Health Support pilots.[3,4] In these cases, the designs did not achieve equal distribution of members with similar characteristics of age, sex, race, mortality likelihood, and medical expenditures. Therefore, it is possible to randomly assign individuals, whether using complete or block randomization, to treatment and control and yet still have groups that are not equivalent or comparable.[5]

Quasi-experimental study designs offer a practical alternative to RCTs; however, the groups typically are selected without using randomization or other efforts to ensure member- and group-level comparability. Unfortunately, nonrandomly selected groups may not be proportionally allocated (ie, imbalanced and heterogeneous) and may be evidenced to have observed and unobserved differences associated with the outcome of interest (ie, statistical bias). If either or both of these concerns are confirmed in the data, subsequent analyses and estimates of program effectiveness (causal effects) will be limited and potentially inaccurate. However, quasi-experimental designs that use effective matching on critical variables related to the outcome of interest have the capability to minimize potential bias, imbalance, and inefficiency inherent to "after the fact" (ex post facto) group identification.

Matching is a nonrandomized, quasi-experimental approach commonly used to test CCM intervention effectiveness. These methods generally involve comparing members who are enrolled in the CCM program (ie, treatment or intervened group members) with members who have comparable attributes but are not enrolled in a CCM program (ie, comparison or nonintervened group members), with both sets of members matched on an ex post facto basis using a common set of factors. The direct comparison between matched comparison and treatment groups has allowed researchers to quantify in a more robust and, potentially, accurate manner the true value of CCM programs,[6–8] confirm associations linking CCM participation with outcomes,[6,7,9–11] and even offer insight into long-term outcomes associated with program involvement.[12] Similar to other retrospective study designs, matching is not without its limitations. For instance, the comparison group frequently is created after the treatment group has been defined such that only the remaining nonparticipating members comprise the comparison group; this type of result can lead to significant selection bias regardless of the statistical means used to adjust for the intergroup differences. It may be impossible to overcome such a strong selection bias, even with robust matching techniques, if the groups do not contain a sufficient proportion of similar individuals. Therefore, caution should be used when implying causality of the CCM intervention when evaluating study groups created using samples of convenience (eg, participant, nonparticipant), as failure to account for the outlined limitations could lead to erroneous conclusions.

In this study, the authors present their experience evaluating outcomes from a CCM program over a 2-year period using observed administrative claims data for ex post facto generated comparison and treatment groups. The goals of this case study were 3-fold: (1) evaluate comparability between a comparison and treatment group based on objective metrics, (2) test the effectiveness of 2 matching methods in creating valid study groups, thus improving the accuracy of the measurement of CCM programs, and (3) evaluate CCM program performance by determining the medical cost savings during the evaluation period.

Methods

Study participants

This case study evaluated CCM program performance of health plan members with 6 continuous months or more of plan eligibility in each evaluation year, who were between the ages of 18 and 64.9 years, and identified as having 1 or more chronic conditions: coronary artery disease, congestive heart failure, chronic obstructive pulmonary disease (COPD), and diabetes. The treatment group was defined from this set of members as those who were fully insured and enrolled in the CCM program ($n = 12{,}202$); enrollment did not necessitate the presence of intervention, which included telephonic interactions with clinicians and/or written material delivered to the members during their enrollment. Conversely, the comparison group ($n = 7914$) was defined as those members selected ex post facto from the Administrative Services Only (ASO) groups of the same health plan that did not elect to purchase the CCM program; these members were not offered interventions.

Descriptive statistics

Member-level administrative claims and eligibility data for the baseline (2005) and first program year (2008) were utilized in this study. The temporal gap in data was because of contractual issues in which the health plan transitioned vendors, resulting in 2005 as the true nonintervened period and 2008 as the first full year of 1 vendor-led program. Evaluated variables, derived from claims and eligibility data, are shown in Table 1 along with the corresponding mean baseline values for the pre-matched treatment and comparison groups.

Quantitative statistics

Two quantitative measures were utilized to evaluate the extent of imbalance and heterogeneity between the study group members both prematching and postmatching—the L1 metric and the Wald test. Generally, imbalance is defined

Table 1. Comparison of Baseline (2005) Variables Prior to Matching*

Explanatory Variable	*Description*	*Treatment (n=12,202)*	*Comparison (n=7,914)*	% $\Delta^{\dagger}$
Age	Age of the member, restricted to 18 to 64.9 years	51.38 (8.6018)	50.47 (8.3921)	−1.8%
Sex	If member is male, then 1; else 0	0.61 (0.4882)	0.58 (0.4941)	−5.1%
Disease program				
CAD	If member was identified through administrative claims as having CAD, then 1; else 0	0.24 (0.4255)	0.21 (0.4088)	−10.7%
CHF	If CHF, then 1; else 0	0.03 (0.1683)	0.03 (0.1779)	12.2%
COPD	If COPD, then 1; else 0	0.07 (0.2514)	0.08 (0.2715)	18.2%
Diabetes	If Diabetes, then 1; else 0	0.67 (0.4718)	0.68 (0.4684)	1.4%
Base Member Months	Total number of eligible member months during baseline	11.78 (1.0027)	11.92 (0.5810)	1.2%
2004 Allowed PMPM Medical Costs ($)	Average allowed monthly administrative claims expenditures in the pre-baseline period	570.87 (1,071.4987)	489.34 (1,013.8602)	−14.3%
Allowed PMPM Medical Costs ($)	Average allowed monthly administrative claims expenditures	629.11 (828.8645)	561.75 (788.7325)	−10.7%
Exceed Average Medical Costs	If member's annual allowed medical costs exceed their group-specific average medical cost, then 1; else 0	0.35 (0.4757)	0.31 (0.4644)	−9.1%
Primary Plan State	If member resides in the 2-state core network area of the health plan, then 1; else 0	0.93 (0.2497)	0.60 (0.4907)	−36.1%
Plan Type	Ordinal variable denominated 1 to 8, with 1 representing the most frequently reported type of health insurance coverage (eg, HMO) and 8 the least frequently reported	1.70 (0.8301)	1.38 (1.3044)	−18.9%
Duration Since Chronic Disease Indication Dummy	If member is identified in administrative claims to have been diagnosed with a chronic condition for less than 24 months prior to program year, then 1; else 0	0.37 (0.4830)	0.44 (0.4959)	17.7%
Inpatient Stays	Total number of unique inpatient stays	0.52 (1.8275)	0.60 (2.0025)	16.7%
Emergency Department Visits	Total number of unique emergency department visits	0.22 (0.7782)	0.26 (0.7292)	19.0%
Outpatient Visits	Total number of unique outpatient visits	12.79 (11.6860)	12.56 (11.8912)	−1.8%
Physician Visits	Total number of unique physician visits	19.75 (13.9824)	13.04 (14.1139)	−34.0%
2004 ICD9 Count	Count of the number of first position ICD9 diagnoses (not unique) recorded in administrative claims in the pre-baseline period	54.74 (50.8496)	43.27 (49.2509)	−21.0%
ICD9 Count	Count of the number of first position ICD9 diagnoses (not unique) recorded in administrative claims	70.57 (86.2643)	70.12 (86.2194)	1.3%
CPT4 Rank	Indexed value based on the weighted average ranking of cost and frequency of recorded procedures in administrative claims	15.88 (13.4794)	15.49 (14.0804)	−2.5%
Prescription Medication (NDC) Count	Count of the number of prescriptions recorded in administrative claims	35.22 (27.6712)	22.59 (26.1818)	−35.8%
Clinical Risk Group (CRG) Score	Proprietary, customer computed measure of morbidity	227.90 (126.4608)	235.96 (108.1854)	−6.7%
Weighted Average Disease Severity	Indexed value based on the member's chronic condition(s) and the associated weighted average cost of the condition(s)	3.89 (1.3313)	3.72 (1.3158)	−4.2%

*All values based on the average of 250 simple random samples taken with replacement, with the treatment group sampled (simple random with replacement) at a rate equal to the original comparison group (analyzed comparison group sampled up to 1.5 times the original size). Members were between the ages of 18 and 64.9 years, evidenced to have a chronic condition of coronary artery disease, congestive heart failure, chronic obstructive pulmonary disease, and/or diabetes (based on administrative claims data), and excluded if medical costs in baseline or program year exceeded the 99th percentile, or if their change in costs over these 2 periods exceeded the 99th percentile.

†Independent, relative percent difference in prematching values (by Explanatory Variable) between comparison and treatment group members. As an example, for the explanatory variable Age, the delta is computed as: (50.47-51.38)/ 51.38 = −1.8%.

CAD, coronary artery disease; CHF, congestive heart failure; COPD, chronic obstructive pulmonary disease; CPT, Current Procedural Terminology; HMO, health maintenance organization; ICD9, *International Classification of Diseases, Ninth Revision*; NDC, National Drug Code; PMPM, per member per month.

in the context of nonparametric evaluation of the comparability of 2 groups of members, whereas heterogeneity is more relevant to statistical tests of group comparability. For both measures, the purpose is to assess quantitatively the extent to which treatment and comparison members are unequally distributed, or conversely, share a common attribute, for a given set of evaluated factors such as those listed in Table 1. Relatively high levels of imbalance and heterogeneity suggest the study groups are not equivalent and are expected to have differential impacts to the variance around the mean of the outcome (dependent) variable.

On the other hand, low levels of imbalance and heterogeneity indicate the 2 groups are of similar composition and should have similar variability around the outcome, allowing for aggregation of the data from each group into 1 data set for analysis within a statistical model. The objective of this analysis would then be to compute the difference in the mean value of the outcome (here, trend in medical expenditures) related to either having received interventions or not, while controlling for all other relevant observable factors. This calculated difference is the monetary measure of the effectiveness of the program, or causal effect.

The L1 metric, which is a nonparametric measure originally developed within Coarsened Exact Matching (CEM) (to be described),[13] quantifies imbalance by comparing relative frequencies of treatment and comparison group members assigned across each of the bins (or strata) created within CEM. For example, consider a simple case of 10 treatment and 20 comparison members stratified based on 2 factors—age less than 40 or greater than or equal to 40 (2 strata denoted "<40," "$=40$") and sex (2 strata denoted "m" and "f"). The distribution of these members is observed to be [$n_{<40\_m}=\{3,7\}$, $n_{<40\_f}=\{4,10\}$, $n_{>=40\_m}=\{1,3\}$, $n_{>=40\_f}=\{2,0\}$], where the first value in each set refers to the number of treatment members and the latter to the comparison members. The L1 is computed as one-half the sum of the absolute difference between the relative proportions across each of the 4 strata; in the case of the first strata capturing men under 40 the contribution to L1 is: $[(3/10) - (7/20) = 0.05]$; the remaining differences in proportions are $n_{<40\_f}=0.10$, $n_{>=40\_m}=0.05$, and $n_{>=40\_f}=0.20$. The resulting prematch L1 for this stratification rule (age demarcated at 40 and sex) is 0.20. In the postmatch case, the L1 would be computed for only those strata in which 1 or more members are exactly matched to at least 1 other member from either the treatment or comparison group. In the example, the stratum denoting women greater than or equal to 40 would be excluded from subsequent analysis and the resulting post-CEM match L1 would be 0.10. Values of L1 close to zero indicate a higher quality match (an equal number of treatment and comparison members in each strata), whereas an L1 value of 1 indicates perfect dissimilarity or disproportionality between the groups (no overlap between groups in the strata assignment). Ideally, the goal is to achieve optimum balance across a given evaluated set of stratification criteria such that L1 is minimized while retaining a relatively high proportion of the original set of members.

There are several statistical-based counterparts to L1 that can be utilized to ascertain balance and homogeneity. In this study, 2 statistical tests are most relevant, the Chow and Wald tests, because the primary reason for specific member assignment to a given study group is known (ASO purchasing decision).[14–16] The Chow and Wald tests are similar in regard to providing a statistical test of whether significant differences are present between 2 sets of data in the magnitude, direction, and variability of influence of a given set of covariates on the same dependent variable. More explicitly, and relevant to the current study, the Chow and Wald tests can be used to determine whether observed data for treatment members can be combined with data for the comparison members such that a dichotomous (explanatory) variable of group identity can be used to measure the influence of the intervention on the outcome of medical costs over time (ie, the causal effect). If the null hypothesis of homogeneity is rejected because of failure to achieve the critical value (F statistic in the case of the Chow test and chi-square statistic in the case of Wald test), then the dichotomous variable of group identity may yield a biased estimate of program effectiveness.

However, the Chow and Wald tests differ concerning robustness in the presence of violation of the statistical modeling assumption of constant error variance (homoscedasticity). If the assumption of homoscedasticity is met, then either the Chow or Wald test can be applied, with selection dependent upon the econometric estimation technique applied. In the case where this assumption is not met, the Wald test is more robust.[14–16]

Matching techniques

To minimize potential imbalance between the treatment and comparison groups, CEM and Propensity Score Matching (PSM) techniques were utilized. CEM is a nonparametric matching method applicable in a quasi-experimental design for the purpose of comparing an outcome between 2 groups over time.[13,17,18] Compared with PSM and in the context of a comparison group framework, CEM has been found to yield estimates of the causal effect with the lowest variance and bias for any sample size (G. King, R. Neilson, J. Pope, C. Coberley, A. Wells, unpublished data, 2011). The increased efficiency and lower bias properties of CEM are attributed to stratification and exact matching of the 2 study groups based on variables that explain variance in the outcome of interest, difference-in-difference computations, and strata-based weighting within a nonparametric framework. Specifically, CEM distinctly assigns each member into one of a defined set of strata in which the members are exactly matched on a set of "coarsened" variables (ie, variables divided into 2 or more meaningful ranges or categories). The matched members are then assigned a weight specific to their stratum and representative of the proportion of all members present in said stratum. Effectively, CEM is a quasi-experimental methodology that facilitates more comparable evaluation of study groups by creating proportionality among the factors contributing to the outcome of interest through blocking members into distinct strata.

The more widely applied PSM method[19–25] consists of estimating a logistic model to derive a propensity score measuring the likelihood a given member is in the treatment group compared to the comparison group based on a common set of explanatory variables. A principal assumption of PSM pertains to independence of irrelevant alternatives (IIA). For this study, the IIA assumption is assumed to be met because an executive-level decision was made, as opposed to member-level decision (where the member is the unit of measurement), to not purchase the CCM program.

Specific to the application of PSM in this study, the Greedy algorithm was utilized within the SAS 9.2 statistical software (SAS Institute Inc., Cary, NC) to minimize the weighted sum of the absolute differences between the comparison and treatment group propensity scores.[26#] More than 30 explanatory variables were evaluated for inclusion in this PSM application, with the final set of variables chosen based upon use of a forward selection algorithm.

Sampling design

Due to the considerable difference in number of treatment ($n = 12{,}202$) and comparison ($n = 7914$) members in addition to the restriction of the Greedy PSM algorithm that the number of comparison members exceed treatment members, a 2-step sampling plan was developed and applied in this study. The plan consisted of first selecting a subset of treatment members equal in number to the comparison group using simple random sampling with replacement. The comparison group was then augmented by 50% through the inclusion of a subset of the original comparison group sampled using simple random sampling with replacement. This combined comparison group (original 7914 members with an additional 3957 members) was then matched to the sampled treatment group ($n = 7914$) in PSM and CEM. In both sampling plans, 250 iterations were applied.

Econometric estimation techniques

The estimate of causal effect or, alternatively, monetary value of program effectiveness, was derived from multivariate statistical modeling using members matched through application of CEM and PSM. The dependent (outcome) variable in these models was differenced-allowed medical expenditures computed at the member level; the difference was computed as the average monthly allowed medical expenditures in program year less baseline year. Given the differenced-dependent variable, a negative coefficient for the causal effect from the regression analysis indicated the comparison group trend in medical expenditures exceeded the treatment group trend and thus reflects gross per member per month savings as a result of the intervention.

The multivariate statistical modeling methods ordinary least squares (OLS) and generalized estimating equations (GEE; normal distribution) were utilized. OLS is a common estimation technique, whereas GEE is a more robust, in terms of estimated coefficient variance and model explanatory power, yet complicated multivariate statistical method. For purposes of this analysis, both GEE and OLS were evaluated to demonstrate the different econometric methods that may be applied in matching studies; however, only GEE results are reported because of the presence of heteroscedasticity in the OLS estimates.

Case Study

Comparability of unmatched study groups

Descriptive and statistical methods were used to assess the level of balance between comparison and treatment group members. Descriptively, imbalance was observed to be most pronounced (greater than 15% relative percent difference in mean baseline values between treatment and comparison members) among members with residence outside of the 2 states defining the health plan's primary network, prescription utilization, physician visits, count of pre-baseline diagnoses, type of health coverage, emergency department visits, COPD prevalence, proportion of members with less than 24 months of chronic disease exposure prior to program start, and number of inpatient stays (Table 1). Conversely, the groups were most balanced in regard to composition of the sexes, a disease severity measure, number of procedures, outpatient visits, age, number of eligible member months, count of diagnoses, and diabetes prevalence. For the remaining descriptive measures of the population, the average absolute relative percent difference in mean baseline values between the 2 groups was 11%, with a standard deviation of 3%.

Reinforcing the descriptive results observed in Table 1, the prematch L1 value for the study groups was 0.42 (Table 2). This value is indicative of a relatively high level of imbalance between the groups and supports the use of matching. However, the Wald test showed a statistically insignificant difference between the study groups in the estimated effect (within a multivariate statistical model) of evaluated explanatory variables on the dependent variable. In other words, the Wald test results indicated that heterogeneity was not observed between the treatment and comparison group members for the specified GEE econometric model. This was an interesting result following the imbalance observed in the L1 and highlights the need for multiple quantitative measures of group comparability prior to assessing program effectiveness.

Effectiveness of two matching methodologies

Based on the qualitative and quantitative imbalance results observed in Tables 1 and 2, CEM and PSM were employed within the aforementioned sampling design. Table 2 lists the postmatch L1 and Wald test results and Table 3 shows a comparison of explanatory variable mean values following application of CEM and PSM. To demonstrate the methodological differences between CEM and PSM, Table 2 reports CEM results compared with PSM when the logit model of PSM was specified using the explanatory factors used to create the CEM strata. Additionally, results are shown when CEM was applied to members selected through the use of the optimal PSM specification (based on forward selection) as well as different specifications of the multivariate model using the propensity score.

The objective of including multiple CEM and PSM specifications within the comparison of matching methods was to enable analysis of the methods from inter (across methods) and intra (within method) perspectives, with the overarching goal to address design concerns surrounding matching criteria specification and matching method selection. In the intermatching method analyses, the objective was to evaluate the extent to which the 2 methods differed in regard to matching metrics (L1 and Wald test) and causal effects when the analyzed factors remained the same. With these analyses, the differential member reduction combined with CEM stratification (blocking) weights principally distinguished the 2 matching methods. Similarly, the objective of the intramatching method analyses was to quantify the difference in matching metrics and causal effects when the method was held fixed, yet alternative specifications of the matching criteria were applied.

Results of the intermatching and intramatching method analyses showed that CEM reported a lower postmatching

Table 2. Comparison of Matching Metrics and Causal Effects Estimated from Coarsened Exact Matching (CEM) and Propensity Score Matching (PSM)*

Matching Method†,‡	n	L1 Metric**		Wald Test		LL***	Estimated Causal Effect (PCMPM)†
		Pre	Post	Pre	Post		
CEM – A	15,027	0.42	0.00	0.81	0.84	−125.2	−$26.46
PSM – A	13,120	0.42	0.23	0.81	0.64	−106.7	−$19.85
PSM – B	13,151	0.42	0.23	0.81	0.71	−106.9	−$19.68
CEM – B	13,024	0.42	0.00	0.81	0.80	−106.5	−$24.67
PSM – C	13,151	0.42	0.23	0.81	0.76	−106.9	−$19.93
PSM – D	13,151	0.42	0.23	0.81	3.02	−130.9	−$34.25

*All values based on the average of 250 simple random samples taken with replacement, with the treatment group sampled (simple random with replacement) at a rate equal to the original comparison group (analyzed comparison group sampled up to 1.5 times the original size). Members were between the ages of 18 and 64.9 years, evidenced to have a chronic condition of coronary artery disease, congestive heart failure, chronic obstructive pulmonary disease, and/or diabetes (based on administrative claims data), and excluded if medical costs in baseline or program year exceeded the 99th percentile, or if their change in costs over these 2 periods exceeded the 99th percentile.

†Letters denote specific CEM or PSM stratification criteria and are described below (note that all variables not denoted as "04" are derived from the baseline period of 2005).

A = DURATION_SINCE_CHRONIC_DISEASE_INDICATION_DUMMY, AGE40_DUMMY, GENDER, IP_DUMMY, ED99_DUMMY, PLAN_STATE_DUMMY, PLAN_TYPE

B = BASE_MM, AGE, GENDER, IP_STAYS, IP_STAYS04, ED_VISITS, OP_VISITS, PHY_VISITS, PLAN_STATE_DUMMY, PLAN_TYPE, NDC_COUNT, WTAVGSEV, CPT_RANK, ICD_COUNT, ICD_COUNT04, SQRT_CLAIMS04, BASE_PMPM, EXCEED_AVG_COST, CRG, ASTHMA_IND, CAD_IND, CHF_IND, DM_IND *(note that propensity score not included beyond use in creating the match)*

C = **B** & propensity score as an explanatory variable

D = **B** & propensity score as the weighting instrument

‡Corresponding Percent Concordant statistic (*c*-statistic) values for the propensity score generating model within the PSM-based Matching Methods, by letter, are as follows (note that all values are based on the aforementioned random sampling plan):

A: *c*–statistic = 0.76

B: *c*–statistic = 0.79

C: *c*–statistic = 0.79 (*identical propensity score generating model as* **B**)

D: *c*–statistic = 0.79 (*identical propensity score generating model as* **B**)

**Pre-L1 metric computed based on the stratification criteria listed as opposed to searching across the relevant parameter space to define the matching variables, using Scott's Binning Algorithm to determine the cut points (eg, age separated at cut points 25, 35, 45, and 55) and then choosing the median L1 value from these specifications.

***Log likelihood values reported in thousands and based on use of generalized estimating equations (GEE) regression model (normal distribution, identify link).

†Estimated Causal Effect is the difference in per chronic member per month (PCMPM) medical costs between the 2 study years and treatment and comparison group members (referred to as the difference-in-difference value). Negative values imply the comparison group trend in medical expenditures exceeded the treatment group trend and thus reflect gross savings. For CEM and PSM, the Estimated Causal Effect is the coefficient for GROUP_TYPE from the GEE regression model (normal distribution, identify link). Note that for each result listed above, the set of evaluated explanatory variables was comprised of the following: GROUP_TYPE, BASE_MM, AGE, GENDER, IP_STAYS, IP_STAYS04, ED_VISITS, OP_VISITS, PHY_VISITS, PLAN_STATE_DUMMY, PLAN_TYPE, NDC_COUNT, WTAVGSEV, CPT_RANK, ICD_COUNT, ICD_COUNT04, SQRT_CLAIMS04, BASE_PMPM, EXCEED_AVG_COST, CRG, MONTHS_AFTER_TRIGGER_DUMMY, ASTHMA_IND, CAD_IND, CHF_IND, DM_IND [*COPD_IND was the reference variable*].

L1 value, failed to reject the Wald test, and demonstrated a higher causal effect. On average, the significant estimated causal effects from CEM exceeded that of PSM by 29% (average savings in terms of the trend in medical expenditures per treatment group member per month equal to -$25.57, compared to -$19.82, from CEM and PSM, respectively). In the intermatching analysis, significantly more members were retained following the CEM matching process. The intramatching analyses indicated that PSM was relatively stable to alternative specifications of the matching criteria whereas CEM evidenced a higher causal effect when the PSM criteria were not imposed. These results indicate that the improved balance between treatment and comparison groups achieved within CEM, regardless of the specification criteria, enabled the effect of the program to be more pronounced.

Findings from the intermethod comparisons showed that the member-level propensity score should not be used as a weighting instrument. The propensity score applied in a weighted GEE model yielded the highest, yet most variable, estimate of program effectiveness. This result indicates the propensity score as a weighting instrument actually may have induced bias in the causal effect estimated with the matched members. On the other hand, use of the propensity score as an additional explanatory variable did not alter conclusions drawn from the aforementioned comparative matching analyses.

In order to gain a better understanding of the difference between CEM and PSM, an ad hoc exploratory analysis was conducted of the treatment and comparison group members included in the CEM match yet not in the PSM match and vice versa (Table 4). In the former analysis, identical treatment group members were retained following CEM and PSM, whereas in the latter analysis CEM removed an additional set ($n = 49$) of treatment members. From a CCM purchaser perspective as well as the researcher perspective, the goal is to estimate the effectiveness of the program using an analytical design in which balance is maximized with minimal removal of treatment group members while retaining only those comparison group members who are most similar to the intervened population. In the analysis of comparison members included in CEM (but excluded from PSM)

Table 3. Comparison of Baseline (2005) Variables Following Application of Coarsened Exact Matching (CEM) and Propensity Score Matching (PSM)*

	Post CEM†		*Post PSM*		
Explanatory Variables	*Treatment* (n = *7833*)	*Comparison* (n = *7221*)	*Treatment* (n = *7833*)	*Comparison* (n = *5295*)	% *Δ*‡
Age	51.41 (0.0579)	51.22 (0.0408)	51.37 (0.0572)	50.84 (0.0276)	−0.4% CEM ‖ −1% PSM
Sex	0.61 (0.0033)	0.61 (0.0033)	0.61 (0.0034)	0.60 (0.0018)	0% CEM ‖ −0.9% PSM
Disease program					
CAD	0.24 (0.0029)	0.23 (0.0008)	0.24 (0.0029)	0.23 (0.0021)	−5% CEM ‖ −1.6% PSM
CHF	0.03 (0.0011)	0.03 (0.0002)	0.03 (0.0011)	0.03 (0.0005)	16.8% CEM ‖ 11% PSM
COPD	0.07 (0.0017)	0.08 (0.0004)	0.07 (0.0017)	0.08 (0.0007)	14.4% CEM ‖ 13.6% PSM
Diabetes	0.67 (0.0032)	0.66 (0.0011)	0.67 (0.0032)	0.66 (0.0021)	−0.4% CEM ‖ −1.3% PSM
Base Member Months	11.78 (0.0068)	11.93 (0.0007)	11.78 (0.0067)	11.89 (0.0011)	1.3% CEM ‖ 1% PSM
2004 Allowed PMPM Medical Costs ($)	564.44 (6.9170)	532.63 (1.2736)	567.76 (7.1124)	523.48 (3.3224)	−5.6% CEM ‖ −7.8% PSM
Allowed PMPM Medical Costs ($)	622.24 (5.2277)	575.17 (3.5667)	626.62 (5.2750)	594.26 (2.6700)	−7.6% CEM ‖ −5.2% PSM
Exceed Average Medical Costs	0.34 (0.0030)	0.34 (0.0017)	0.34 (0.0030)	0.33 (0.0017)	−1.6% CEM ‖ −3% PSM
Primary Plan State	0.93 (0.0016)	0.93 (0.0016)	0.93 (0.0016)	0.79 (0.0039)	0% CEM ‖ −15.8% PSM
Plan Type	1.82 (0.0076)	1.83 (0.0076)	1.84 (0.0077)	1.66 (0.0162)	0.3% CEM ‖ −9.9% PSM
Duration Since Chronic Disease Indication Dummy	0.37 (0.0033)	0.37 (0.0033)	0.37 (0.0033)	0.41 (0.0019)	0% CEM ‖ 10% PSM
Inpatient Stays	0.50 (0.0119)	0.50 (0.0094)	0.51 (0.0120)	0.57 (0.0066)	−0.6% CEM ‖ 11.2% PSM
Emergency Department Visits	0.20 (0.0048)	0.21 (0.0022)	0.21 (0.0051)	0.24 (0.0029)	4.9% CEM ‖ 12.4% PSM
Outpatient Visits	12.71 (0.0781)	13.24 (0.0362)	12.76 (0.0789)	13.14 (0.0441)	4.1% CEM ‖ 3% PSM
Physician Visits	19.71 (0.0965)	15.99 (0.0438)	19.73 (0.0957)	15.27 (0.0571)	−18.9% CEM ‖ −22.6% PSM
2004 ICD9 Count	54.35 (0.3434)	52.55 (0.1017)	54.43 (0.3466)	49.57 (0.1710)	−3.3% CEM ‖ −8.9% PSM
ICD9 Count	55.88 (0.2703)	56.37 (0.1358)	56.07 (0.2719)	57.55 (0.1764)	0.9% CEM ‖ 2.6% PSM
CPT4 Rank	15.83 (0.0882)	16.27 (0.0354)	15.86 (0.0888)	16.02 (0.0464)	2.8% CEM ‖ 1% PSM
Prescription Medication (NDC) Count	35.15 (0.1917)	27.81 (0.0787)	35.18 (0.1902)	26.54 (0.1003)	−20.9% CEM ‖ −24.5% PSM
Clinical Risk Group (CRG) Score	233.81 (0.2242)	231.85 (0.0790)	233.83 (0.2227)	229.02 (0.2912)	−0.8% CEM ‖ −2.1% PSM
Weighted Average Disease Severity	3.88 (0.0089)	3.85 (0.0035)	3.88 (0.0089)	3.81 (0.0054)	−0.8% CEM ‖ −1.9% PSM

*Results based on CEM stratification {DURATION_SINCE_CHRONIC_DISEASE_INDICATION_DUMMY, AGE40_DUMMY, GENDER, IP_DUMMY, ED99_DUMMY, PLAN_STATE_DUMMY, PLAN_TYPE}; all values based on the average of 250 simple random samples taken with replacement, with the treatment group sampled (simple random with replacement) at a rate equal to the original comparison group (analyzed comparison group sampled up to 1.5 times the original size). Members were between the ages of 18 and 64.9 years, evidenced to have a chronic condition of coronary artery disease, congestive heart failure, chronic obstructive pulmonary disease, and/or diabetes (based on administrative claims data), and excluded if medical costs in baseline or program year exceeded the 99th percentile, or if their change in costs over these 2 periods exceeded the 99th percentile. Standard deviation in parentheses.

†CEM results are based on application of CEM-derived weights.

‡Independent, relative percent difference in postmatch values (by Explanatory Variable) for comparison and treatment group members by matching method. As an example, for the explanatory variable Age, the delta is computed as: (51.22 − 51.41)/51.41 = −0.4% [CEM] and (50.84 − 51.37)/51.37 = −1 % [PSM].

CAD, coronary artery disease; CHF, congestive heart failure; COPD, chronic obstructive pulmonary disease; CPT, Current Procedural Terminology; ICD9, *International Classification of Diseases, Ninth Revision*; NDC, National Drug Code; PMPM, per member per month.

Table 4. Comparison of Baseline (2005) Variables for Members Included or Excluded between Coarsened Exact Matching (CEM) and Propensity Score Matching (PSM)*

	Post-CEM Match (Included in CEM, Excluded in PSM)		*Post-PSM Match (Included In PSM, Excluded in CEM)*		
Explanatory Variables	*Treatment*† (n=.)	*Comparison* (n=*1994*)	*Treatment* (n=*49*)	*Comparison* (n=*37*)	% Δ‡
Age	AMR	50.23 (0.2047)	44.55 (0.9669)	44.97 (1.3747)	−10%
Sex	AMR	0.50 (0.0351)	0.41 (0.0389)	0.38 (0.0794)	−24%
Disease program					
CAD	AMR	0.13 (0.0155)	0.22 (0.0364)	0.23 (0.0545)	85%
CHF	AMR	0.04 (0.0018)	0.05 (0.0194)	0.02 (0.0229)	−39%
COPD	AMR	0.10 (0.0037)	0.07 (0.0219)	0.14 (0.0334)	43%
Diabetes	AMR	0.74 (0.0160)	0.66 (0.0405)	0.60 (0.0612)	−18%
Base Member Months	AMR	11.87 (0.0080)	11.36 (0.1351)	11.79 (0.0529)	−1%
2004 Allowed PMPM Medical Costs ($)	AMR	374.56 (11.4938)	1,094.25 (235.0966)	441.45 (90.8875)	18%
Allowed PMPM Medical Costs ($)	AMR	465.35 (26.1465)	1,322.08 (112.2179)	1,099.00 (128.3874)	136%
Exceed Average Medical Costs	AMR	0.26 (0.0131)	0.68 (0.0396)	0.72 (0.0584)	178%
Primary Plan State	AMR	0.27 (0.0211)	0.75 (0.0386)	0.41 (0.0769)	52%
Plan Type	AMR	2.40 (0.0821)	4.78 (0.1664)	3.85 (0.3932)	60%
Duration Since Chronic Disease Indication Dummy	AMR	0.54 (0.0373)	0.34 (0.0383)	0.61 (0.0856)	13%
Inpatient Stays	AMR	0.54 (0.0677)	1.96 (0.2651)	2.02 (0.4067)	273%
Emergency Department Visits	AMR	0.22 (0.0090)	1.52 (0.2077)	1.84 (0.4763)	719%
Outpatient Visits	AMR	11.13 (0.2870)	19.33 (1.5772)	20.07 (1.3314)	80%
Physician Visits	AMR	9.65 (0.5128)	23.03 (1.8850)	19.23 (1.5928)	99%
2004 ICD9 Count	AMR	29.59 (0.7897)	67.47 (7.5389)	45.37 (5.6799)	53%
ICD9 Count	AMR	53.35 (1.3111)	86.63 (4.7918)	98.51 (6.5595)	85%
CPT4 Rank	AMR	14.09 (0.2699)	20.76 (1.7155)	21.02 (2.0439)	49%
Prescription Medication (NDC) Count	AMR	16.73 (0.8627)	40.06 (3.3940)	30.57 (2.7375)	83%
Clinical Risk Group (CRG) Score	AMR	212.42 (1.1264)	238.20 (2.3102)	230.91 (6.1218)	9%
Weighted Average Disease Severity	AMR	3.49 (0.0336)	4.01 (0.1199)	4.23 (0.1981)	21%

*Results based on CEM stratification {DURATION_SINCE_CHRONIC_DISEASE_INDICATION_DUMMY, AGE40_DUMMY, GENDER, IP_DUMMY, ED99_DUMMY, PLAN_STATE_DUMMY, PLAN_TYPE}; all values based on the average of 250 simple random samples taken with replacement, with the treatment group sampled (simple random with replacement) at a rate equal to the original comparison group (analyzed comparison group sampled up to 1.5 times the original size). Members were between the ages of 18 and 64.9 years, evidenced to have a chronic condition of coronary artery disease, congestive heart failure, chronic obstructive pulmonary disease, and/or diabetes (based on administrative claims data), and excluded if medical costs in baseline or program year exceeded the 99th percentile, or if their change in costs over these 2 periods exceeded the 99th percentile. Standard deviation in parentheses.

†AMR denotes "All Members Retained," implying CEM and PSM retained identical treatment group members.

‡Independent, relative percent difference in postmatch values (by Explanatory Variable) between CEM and PSM for comparison group members. As an example, for the explanatory variable Age, the delta is computed as: (44.97 − 50.23)/50.23 = −10%.

CAD, coronary artery disease; CHF, congestive heart failure; COPD, chronic obstructive pulmonary disease; CPT, Current Procedural Terminology; ICD9, *International Classification of Diseases, Ninth Revision*; NDC, National Drug Code; PMPM, per member per month.

compared to members included in PSM (but excluded from CEM), the results showed that the composition of the 2 sets of members was vastly different. On an absolute value basis, the members were most different in regard to emergency department visits, inpatient stays, likelihood to exceed group-specific average costs, medical expenditures, and physician visits. Overall, the results demonstrate that PSM was more likely to include comparison group members who were higher utilizers, of greater morbidity, and more costly in terms of medical costs.

A final analysis was conducted to assess the plausibility of the causal effects from a utilization perspective (Table 5). The analysis showed that on average, members in the treatment group reported a reduction in the number of inpatient stays over the 2-year study period, compared to an increase in the number of stays in the comparison group. For emergency room visits, though, the comparison group showed an increase while the treatment group increased slightly more. For the remaining 2 utilization metrics, both the treatment and comparison groups evidenced increases over the 2 study years. For outpatient claims, the comparison group trend was approximately 1.5 times greater than the treatment group; for physician claims, the treatment trend exceeded the comparison by 1.4 times. These results suggest that the causal effect reported in Table 2 was predominately related to reductions in inpatient followed by outpatient utilization, with the increase in physician claims a likely outcome of members being encouraged to increase proactive care via their physicians.

Conclusion

As the industry moves toward more rigorous methods of program evaluation that employ statistical research designs based on comparison groups, it is important that robust methods for accurate and objective creation of comparison groups be advanced. Although significant developments have been made in standardizing CCM study designs and methodologies over the years, standardization among other rapidly growing programs of interest (eg, wellness and total population health) is less well developed.[27] However, the industry foresees the adoption and implementation of similar standardized practices across all program types, where applicable, in the near future.[1] This study aimed to generate an open discussion among researchers regarding challenges and considerations that arise during CCM study design and analysis, while also informing readers of viable alternative methodologies.

A primary objective of this study was to provide a scientific framework for CCM program evaluation within a quasi-experimental, multivariate statistical setting, using matched comparison groups. Evaluation of 2 matching methods based on descriptive measures, the Wald test and L1 metric, for the analysis of CCM data derived over a 2-year period from a large health plan was reported. Baseline differences between comparison and treatment members across a common set of attributes were observed; however, application of an advanced matching method (CEM) and more rigorous statistical procedures enabled construction of a comparable data set for assessment of the causal effect of the program.

The methodologies presented in this article should not be considered definitive solutions to the problem of forming balanced comparison groups adequate for estimating cost savings; instead, they represent alternative methodological options researchers can consider to validate the integrity of their comparison groups within quasi-experimental designs. It is important to note that each methodology will have its own unique strengths and weaknesses that may vary based on the application and or sample population. Given our findings of the differences in balance and causal effects generated by use of different matching methods, future studies of CCM outcomes should focus on comparing not just CEM and PSM but other less commonly utilized

Table 5. Comparison of Utilization Variables During Baseline (2005) and Program (2008) Year Following Application of Coarsened Exact Matching (CEM)*

	Baseline Year†		*Program Year*		
Explanatory Variables	*Treatment (n = 7833)*	*Comparison (n = 7221)*	*Treatment (n = 7833)*	*Comparison (n = 7221)*	*% Δ*‡
Inpatient Stays	0.50 (0.0119)	0.50 (0.0094)	0.43 (0.0107)	0.51 (0.0029)	2.0% CP || −14.0% TX
Emergency Department Visits	0.20 (0.0048)	0.21 (0.0022)	0.21 (0.0042)	0.22 (0.0012)	4.8% CP || 5.0% TX
Outpatient Visits	12.71 (0.0781)	13.24 (0.0362)	13.08 (0.0819)	13.83 (0.0239)	4.5% CP || 2.9% TX
Physician Visits	19.71 (0.0965)	15.99 (0.0438)	22.08 (0.1047)	17.36 (0.0385)	8.6% CP || 12.0% TX

*Results based on CEM stratification {DURATION_SINCE_CHRONIC_DISEASE_INDICATION_DUMMY, AGE40_DUMMY, GENDER, IP_DUMMY, ED99_DUMMY, PLAN_STATE_DUMMY, PLAN_TYPE}; all values based on the average of 250 simple random samples taken with replacement, with the treatment group sampled (simple random with replacement) at a rate equal to the original comparison group (analyzed comparison group sampled up to 1.5 times the original size). Members were between the ages of 18 and 64.9 years, evidenced to have a chronic condition of coronary artery disease, congestive heart failure, chronic obstructive pulmonary disease, and/or diabetes (based on administrative claims data), and excluded if medical costs in baseline or program year exceeded the 99th percentile, or if their change in costs over these 2 periods exceeded the 99th percentile. Standard deviation in parentheses.

†Results are based on application of CEM-derived weights.

‡Independent, relative percent difference in postmatch values (by Explanatory Variable) for comparison (denoted CP) and treatment (denoted TX) group members by year. As an example, for the explanatory variable Inpatient Stays, the delta is computed as: (0.43 − 0.50)/0.50 = −14.0%.

methods. In exploring the applicability and effectiveness of a matching method, effort should be directed toward comparing alternative specifications of the factors upon which the match is based. For example, age, sex, and baseline health risk severity are necessary but not sufficient matching factors; however, a definitive, sufficient set of matching factors has yet to be defined, leaving CCM purchasers to consider various specifications from CCM providers without knowledge of the most robust option. In addition, future efforts should consider different regression models and explanatory variables applied in these models to determine program savings estimates. Beyond CCM program evaluation, comparison of matching methods should be applied to wellness program evaluation and, subsequently, programs that seek to understand, manage, and improve health risks across an entire population.

An important study limitation to note is that although the results presented here demonstrate that, in addition to matching, the inclusion of a comprehensive set of explanatory variables improves study group comparability, only observable differences are considered. Another limitation of matching techniques in general is that when members are matched based on similar variables into blocks or strata, as in CEM, reductions in the sample population usually occur, especially when the pool of comparison group members is small relative to the treatment. Thus, researchers should consider the trade-off of fewer analyzed members from the original population and potential variance inflation related to sparse strata with the benefits of improved balance and homogeneity in the matched cohorts.

In conclusion, the need for new and improved methodological approaches is critical to continuously advance the design, implementation, and execution of scientifically rigorous studies in the CCM field. While eager to share our research experiences and methodology applications, we are aware that such applications should be cross-validated among different CCM providers, programs, and time periods to standardize and ensure robustness. The hope is that our findings will serve as a catalyst for more substantive discussion so that, collectively, we can continue to advance the field.

Acknowledgment

The Center for Health Research would like to acknowledge William Greene, Ph.D., Toyota Motor Corporation Professor of Economics, at the New York University Leonard N. Stern School of Business for offering his expertise and insight regarding Chow test analysis and data interpretation. We also would like to acknowledge Gary King, Ph.D., Albert J. Weatherhead III University Professor and Director of the Institute for Quantitative Social Science at Harvard University, Department of Government.

Disclosure Statement

Drs. Wells, Hamar, Gandy, Coberley, Rula, and Pope, and Mr. Sidney are employed by Healthways, Inc., a provider of population health management programs. Dr. Bradley and Ms. Harrison were employed by Healthways, Inc. at the time of manuscript development. Neither Dr. Bradley nor Ms. Harrison disclosed a potential conflict of interest.

References

1. Disease Management Association of America. *Outcomes Guidelines Report.* Vol 4. Washington, DC: DMAA; 2009.
2. Disease Management Association of America. *Outcomes Guidelines Report.* Vol 3. Washington, DC: DMAA; 2008.
3. King G, Nielsen R, Coberley C, Pope JE, Wells A. Avoiding randomization failure in program evaluation, with application to the Medicare Health Support program. Popul Health Manag 2011;14:S11–S22.
4. Atherly A, Thorpe KE. Analysis of the treatment effect of Healthways' Medicare Health Support phase 1 pilot on Medicare costs. Popul Health Manag 2011;14:S23–S28.
5. Stel VS, Zoccali C, Dekker FW, Jager KJ. The randomized controlled trial. Nephron Clin Pract 2009;113:c337–c342.
6. Sirey JA, Bruce ML, Kales HC. Improving antidepressant adherence and depression outcomes in primary care: The treatment initiation and participation (TIP) program. Am J Geriatr Psychiatry 2010;18:554–562.
7. Lorig K, Ritter PL, Laurent DD, et al. On-line diabetes self-management program: A randomized study. Diabetes Care 2010;33:1275–1281.
8. Ellis JL, Bayliss EA, Totsch J, Steiner JF. C-B2-01: Disease and care management for multimorbid patients in an integrated system: How much is too much? Clin Med Res 2010;8:44.
9. Sawamura K, Ito H, Koyama A, Tajima M, Higuchi T. The effect of an educational leaflet on depressive patients' attitudes toward treatment. Psychiatry Res 2010;177:184–187.
10. Patel DN, Lambert EV, da Silva R, et al. The association between medical costs and participation in the vitality health promotion program among 948,974 members of a South African health insurance company. Am J Health Promot 2010;24:199–204.
11. Mudge A, Denaro C, Scott I, Bennett C, Hickey A, Jones MA. The paradox of readmission: Effect of a quality improvement program in hospitalized patients with heart failure. J Hosp Med 2010;5:148–153.
12. Ramalho de Oliveira D, Brummel AR, Miller DB. Medication therapy management: 10 years of experience in a large integrated health care system. J Manag Care Pharm 2010;16: 185–195.
13. Iacus S, King G, Porro G. cem: Software for coarsened exact matching. J Stat Software 2009;30:1-=26.
14. Greene W. *Econometric Analysis*. 5th ed. Upper Saddle River: Prentice Hall; 2003.
15. Maddala G. *Introduction to Econometrics*. 3rd ed. New York: John Wiley & Sons; 2001.
16. Gujarati D. *Basic Econometrics*. 4th ed. Boston, MA: McGraw-Hill; 2003.
17. Iacus S, King G, Porro G. Causal inference without balance checking: Coarsened exact matching. Pol Analysis 2012; 20:1-=24.
18. Iacus S, King G, Porro G. Multivariate matching methods that are monotonic imbalance bounding. J Am Stat Assn 2011;106:345-=361.
19. Hirano K, Imbens G. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. Health Serv Outcomes Res Methodol 2001;2:259–278.
20. Lipkovic I, Adams D, Mallinckrodt C. Evaluating dose response from flexible dose clinical trials. BMC Psychiatry 2008;8:1–9.
21. Allen-Ramey FC, Duong PT, Goodman DC, et al. Treatment effectiveness of inhaled corticosteroids and leukotriene modifiers for patients with asthma: An analysis from managed care data. Allergy Asthma Proc 2003;24:43–51.

22. Perkins SM, Tu W, Underhill MG, Zhou XH, Murray MD. The use of propensity scores in pharmacoepidemiologic research. Pharmacoepidemiol Drug Saf 2000;9:93–101.
23. Shepardson LB, Youngner SJ, Speroff T, Rosenthal GE. Increased risk of death in patients with do-not-resuscitate orders. Med Care 1999;37:727–737.
24. D'Agostino RB Jr. Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265–2281.
25. Brandt S, Gale S, Tager IB. Estimated effect of asthma case management using propensity score methods. Am J Manag Care 2010;16:257–264.
26. Bergstralh E, Kosanke J. *Computerized Matching of Controls*. Rochester, MN: Mayo Foundation; 1995. Technical Report 56.
27. Mattke S, Serxner SA, Zakowski SL, Jain AK, Gold DB. Impact of 2 employer-sponsored population health management programs on medical care cost and utilization. Am J Manag Care 2009;15:113–120.

Address correspondence to:
Aaron R. Wells, Ph.D.
Center for Health Research
Healthways, Inc.
701 Cool Springs Blvd.
Franklin, TN 37067

E-mail: aaron.wells@healthways.com