EVALUATION METHODS IN DISEASE MANAGEMENT: DETERMINING PROGRAM EFFECTIVENESS

Ariel Linden, DrPH, MS
John L Adams, PhD
Nancy Roberts, MPH

Commissioned by the Disease Management Association of America

EVALUATION METHODS IN DISEASE MANAGEMENT: DETERMINING PROGRAM EFFECTIVENESS

Ariel Linden, Dr.P.H., M.S., [1] John L. Adams, Ph.D., [2] Nancy Roberts, M.P.H. [3]

[1] President, Linden Consulting Group, Portland, OR. ariellinden@yahoo.com
[2] Senior Statistician, RAND Corporation, Santa Monica, CA. adams@rand.org
[3] Regional Director of Integrated Performance/Six Sigma Champion, Providence Health System, Portland, OR. nancy.roberts@providence.org

ABSTRACT

The purpose of this paper is to provide an overview of various evaluation techniques that should be considered for assessing disease management (DM) program effectiveness. Each design has a particular suitability, depending on the type of DM program and the outcome variables measured. In all cases however, threats to validity must be controlled for or the accuracy of the results may be called into question. Of particular concern, is the current use of utilization measures as the focus of the intervention and cost as the outcome measure. To mitigate the effects of this potential measurement bias, utilization variables should be used as the outcome measure and a transformation made to financial performance. This document should provide DM program planners the tools necessary for choosing the evaluation method that is most appropriate for assessing effectiveness in improving health outcomes and reducing morbidity of the population.

INTRODUCTION

The disease management industry is currently at a crossroads. Health plan and employer group partners are no longer willing to blindly assume that positive health outcomes and cost savings are guaranteed without evidence, and disease management (DM) programs are scrambling to demonstrate that their programs indeed work.[1-3]

The purpose of this paper is to provide an overview of various evaluation techniques that should be considered for assessing DM program effectiveness. However, no matter which design is chosen, careful attention must be paid to the identification and control of potential biases that may invalidate the results. Therefore, this paper will also address the issue of internal validity and provide methods of controlling for bias. This manuscript covers a wide breadth of content at a summary level. When necessary, readers will be referred to more comprehensive references for the given topic under discussion. Ultimately, this document should provide DM program planners with the tools necessary for choosing the evaluation method that is most appropriate for assessing DM program effectiveness.

EXPECTED IMPACT OF DM PROGRAMS

According to the Disease Management Association of America (DMAA), a full-service disease management program must include all of the following: population identification processes, evidence-based practice guidelines, collaborative practice models to include physician and support-service

providers, patient self-management education, process and outcomes measurement, evaluation and management, and routine reporting/feedback loops (which may include communication with patient, physician, health plan and ancillary providers, and practice profiling). The primary purpose of these programs is to improve the quality, consistency and comprehensiveness of care for those with chronic illness, thereby improving health outcomes, reducing morbidity of the population, and reducing use of high cost interventions when lower cost alternatives with equal efficacy are available.

Evidence of improvement in chronic disease management will first be seen by changes in patient and provider behavior directly impacted by DM program interventions. For example, as a direct result of a congestive heart failure (CHF) program intervention, more patients will be monitoring their weight daily and physicians will be more likely to prescribe ACE inhibitors for their patients with heart failure. If DM program interventions successfully influence patient and provider behavior, these changes will result in improvements in physiologic measures, such as blood pressure, pulmonary function or glucose control. These clinical measures are indirectly impacted by the DM interventions. The third level of evidence of DM program effectiveness flows from the prior two. If clinical measures are improved patients are less likely to experience acute exacerbations of their illness requiring an emergency visit or hospitalization. Given evidence of the effectiveness of the direct DM interventions and their indirect impact on clinical measures, improvements in health services utilization measures are a reasonably expected impact of DM programs.

THREATS TO VALIDITY OF DM PROGRAM OUTCOMES

Any design chosen to evaluate DM program effectiveness is subject to exposure to factors that may influence the results, outside of the intervention itself. These factors, termed biases, may cause the results to look better or worse than what was actually achieved by the intervention. Consequentially, these potential sources of bias must be identified and controlled for when developing an evaluation strategy.

Table 1 provides a comprehensive matrix of potential sources of bias that may impact the results attained from a disease management program intervention.[4] The first column indicates the source of the bias, the second column specifies the type of bias (using the nomenclature normally appearing in the research literature), the third column explains how these factors may be manifested, and the fourth column shows how these variables may affect the results. While a thorough discourse on controlling for biases is beyond the scope of this paper (the reader is referred to references 4 through 7 for a comprehensive discussion on the topic), several of the most prevalent and controllable biases will be presented here.

Selection Bias

In simple terms, selection bias means that program participants are not representative of the population of all possible participants. In other words, there are some fundamental differences between those members enrolled in the DM program as compared to those members suitable for the program but

Table 1. Potential Threats to Validity of Disease Management Program Results[4]

Source	*Type*	*How Evidenced?*	*Possible Result?*
Member	Selection Bias	Motivated members more likely to enroll and achieve desired behavior	+
		Sicker members may enroll due to the "fear-factor"	-
		Voluntary enrollment model may enroll motivated members	+
		Engagement model will force members to enroll, including the unmotivated	-
	Loss to Attrition	Members may disenroll from the program or health plan, or die	?
	Maturation	Progressive disease, patients will get sicker	-
	Benefit Design	Member cost sharing may influence use of health services	?
Health Plan/Program	Unit Cost Increases	Costs will appear higher if unadjusted for changes in pricing of services	-
		Reimbursement method and coding changes may alter unit cost	?
	Regression to the Mean	High cost members in the 1st year will cost less in the 2nd year	+
		Low cost members in the 1st year will cost more in the 2nd year	-
	Case-mix	Turnover of health plan membership may change population health status	?
	Treatment interference	Members may be exposed to more than one competing intervention	+
	Access to Services	Utilization may be impacted by access to and availability of providers	?
	New Technology	New technologies may increase some costs and/or reduce others	?
Physician/Provider	Hawthorne Effect	Changing practice patterns due to being observed	+
	Reimbursement method	Practice patterns may be influenced by risk and reimbursement models	?
Data	Sensitivity/Specificity	Imprecise disease identification algorithms may miss suitable members	?
	Missing Information	Algorithms may require specific data sources which may be unavailable	-
	Seasonality	Outcomes may be effected if unadjusted for seasonal influences	?
Measurement	Reliability	Can the same results be produced repeatedly?	?
	Validity	Does the outcome measure make sense?	?
General	Secular Trends	External factors may mask the impact of the DM program intervention	?

Note: In the table, a plus indicates that the re-measurement period may show a better result than the baseline, a minus indicates a worse result than baseline, and a question mark indicates that the result may be difficult to determine.

who are not enrolled. Typically, people who choose to enroll in a DM program are motivated to achieve a desired improvement in health status. Similarly, individuals who have recently undergone an acute adverse health event may be more inclined to participate in a DM program than those who have a chronic illness with no acute exacerbations. If an evaluation of a program's impact does not control for selection bias, then the results may appear better or worse than expected because this cohort of enrolled members may include healthier members, sicker members, motivated members or unmotivated members in unequal proportions to what is representative of the population from which they were drawn.

There are several ways to control for selection bias, the most preferred method being the addition of a control group that is randomly drawn from the same population as the cohort in the DM program but who does not receive any of the program's influence. If a positive outcome is achieved in the program's cohort and not in the control group, the results can be considered valid (assuming that all other threats to validity were controlled for). As a practical matter, DM programs currently do not use control groups under the belief that; (a) it would be costly and difficult to track behavioral change and outcomes for a group not under their purview, (b) the organization may be hesitant to offer services to one sub-set of the population while withholding that same "value-added" benefit to others, and (c) given the belief that these interventions are clinically effective, the DM program may argue the need to treat all members with the disease, since each member receiving the intervention has the potential of adding to the medical-cost savings and positive clinical outcomes promised by the program.

There are two rather simple methods that can be employed in a natural setting for creating control groups drawn from readily available administrative data. *Predictive modeling*, which is a recently developed method for identifying individuals appropriate for a DM program[8] is also quite suitable as a method for determining a control group for evaluation purposes. Using this technique, patients are assigned a risk score based on their likelihood of using health services in the near future. Members enrolled in the DM program can be matched to those members not in the program based on their risk score. The program evaluation will entail a comparison of outcomes of the enrolled members to the control group matched on those scores.

Another method that can be used for developing a control group using administrative data is called *Propensity Scoring*.[9-11] Using this simple technique, all individuals eligible for the program at the outset are given a score (derived from logistic regression) based on their likelihood of being in the program. Thus, scores range from 0 to 1 with a score of 1 indicating perfect likelihood of being enrolled and 0 indicating a perfect likelihood of not being enrolled. Several independent variables can be used to drive this regression, with the most common being age and sex. In fact, a member's predictive modeling score could also be added as a variable to increase the model's sensitivity. Ultimately, enrolled members are matched one-for-one with controls based on their propensity score, and outcomes are compared between the two groups. Using this method, baseline

characteristics between both groups should be statistically similar.

Both methods for creating control groups are dependent on the presence of a pool of patients suitable for the program yet who did not receive any of the program interventions. However, for DM programs using "presumptive" enrollment strategies, this may be problematic. Under this enrollment method, all eligible members are presumed to agree to program participation unless they actively opt-out. This results in higher enrollment penetration, therefore a smaller pool of non-participants from which to develop a control group. In this situation, level or intensity of program participation may also need to be a factor in the matching process.

Regression to the Mean

At the aggregate health plan or DM program level, regression to the mean poses a serious threat to the validity of the results.[4] Also referred to as statistical regression, this concept suggests that, without the effect of the intervention, members with high costs and utilization in the baseline year will tend to cost less and use fewer services in the following year (a move toward the mean). Conversely, members using few services in the baseline year will use more services and accrue higher costs in the subsequent year. Some DM companies contend that, due to the nature of progressive chronic disease, increased costs can, and should be expected year-over-year. The results shown in Figure 1 should dispel that notion.[4] As illustrated in this single health plan example (using a continuously enrolled cohort over the course of two years during which no chronic disease interventions were in place), regardless of type of chronic illness, the regression to the mean occurs. 7% to 11% of members in the highest cost quintile in year 1 will move to the lowest cost quintile in year 2. Conversely, 11 to 17% of members in the lowest cost quintile in year 1 will move to the highest cost quintile in year 2. While these data show that there is movement toward the mean, the actual total costs may vary from year to year, and from quintile to quintile. Therefore, it would be incorrect to conclude that total average costs remain stable across measurement periods.[4]

Regression to the mean can be controlled for using several methods. If a pre- and post-test evaluation design is used, the addition of a randomly chosen control group (or matched using the propensity scoring method) should nullify the effect of regression. In other words, two similar groups chosen at baseline should produce similar outcomes at the end of the study because both groups are expected to regress similarly. However, if one group receives an intervention and achieves better outcomes, we can attribute that success to the intervention. A novel method to consider for controlling regression when using a pre-post design is lengthening each period of measurement to two years or more. This will mitigate the effects of regression by allowing natural movement of individuals back and forth across the range of measurement and reach equilibrium around the mean. Perhaps the best method for controlling regression is time-series analysis. As will be discussed later in this paper, the time-series design is the principal method for controlling for regression because, over time, the natural movement of observations will achieve equilibrium.

Measurement Error

One of the biggest concerns facing health plans when contracting with DM vendors is the reconciliation process. Due to the sometimes, large discrepancies found in the results when data analyses are performed by both the vendor and the plan, this period has been coined "the reconciliation blues".[12] This is a textbook example for demonstrating the importance of having reliable outcome measures. In other words, if the iterative process for arriving at the results is clearly defined and followed, then the outcome measure should be identical whether the vendor analyzes the data repeatedly, or whether the health plan and the vendor run the analysis separately and compare results in the end.

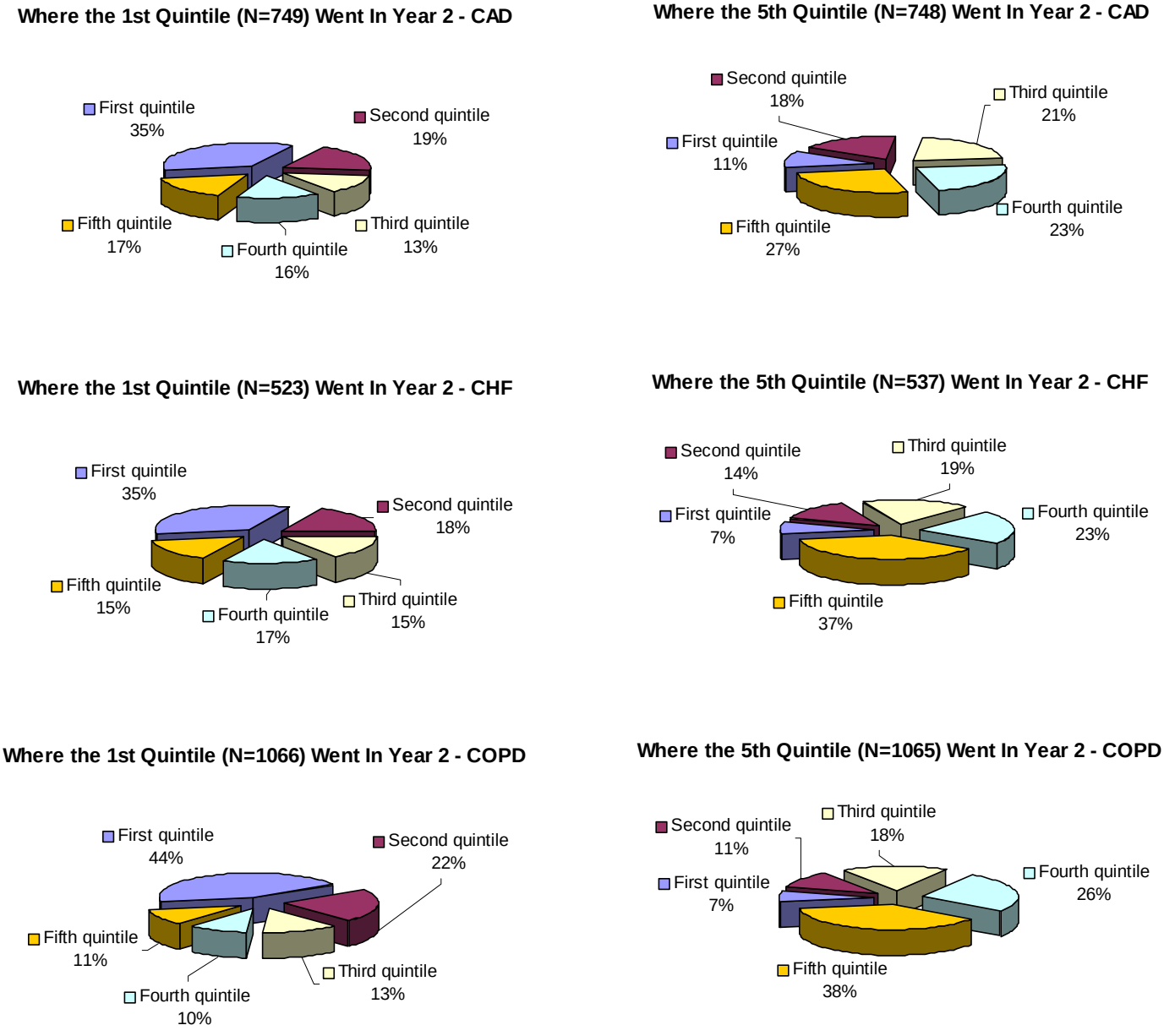

Figure 1. Actual data illustrating the regression to the mean phenomenon in Coronary Artery Disease (CAD), Congestive Heart Failure (CHF), and Chronic Obstructive Pulmonary Disease (COPD). Quintiles are ranked from 1 to 5, with 1 being the lowest cost group and 5 being the highest. All patients were continuously enrolled during the 2-year period

Poorly chosen outcome measures may invalidate the entire program evaluation. In simple terms, validity means that the outcome measure is a true portrayal of what it is supposed to represent. Sometimes the measure appears to be valid, but the underlying process to achieve the results invalidates the outcome. For example, assume that the chosen outcome measure is cost, and that total healthcare costs are included in the calculation (including disease-specific and non-disease specific costs). It is entirely possible that disease-specific costs increased (as a result of increased hospitalizations and ED visits), while costs decreased enough in the non-disease category to outstrip those disease-specific increases. As a result, the total costs for the measurement period would be lower than in the baseline. Clearly, a successful DM program would be expected to reduce utilization in these categories not increase them, yet the inclusion of non-disease costs into the equation could bias the resulting outcome measure.

In summary, bias can influence the results of a DM program whether the intervention was effective or not. Therefore, it is imperative to carefully choose an evaluation design that controls for biases or reduces the types of biases that cannot be identified or controlled for.

EVALUATION DESIGNS FOR ASSESSING DM PROGRAM IMPACT

In this section, three evaluation designs will be examined for assessing the impact of DM program interventions. The first method, the *Total Population Approach*, is currently the most widely used model in the DM industry. The second model, *Survival Analysis*, offers a better solution to the current method because it takes into account the effect of enrollment and disenrollment patterns during the program's implementation. The third method, *Time Series Analysis*, is another method to be considered for evaluating DM program effectiveness because it (a) controls for the effect of many biases, (b) provides a timeline for changes in outcome measure patterns, and (c) is more suitable for evaluating changes at the population level, which DM programs are intended to do.

The Total Population Approach

Figure 2 illustrates the generalized Total Population Approach (TPA) model typically used for evaluating impact of DM programs on medical utilization and cost.[4] The baseline measurement year usually denotes the twelve-month period ending with the month prior to the program launch. Each subsequent measurement period equals the baseline period in duration. To improve measurement accuracy, analysis of each measurement year does not occur until after a complete claims run-out period (usually 3 to 4 months is sufficient to collect close to 100% of medical and pharmacy claims).

This model is a pretest-posttest design, which is a relatively weak research and evaluation technique.[4-7] The most basic limitation of this design is that there is no control group for which comparisons of outcomes can be made. Consequently, the results may be influenced by many of the biases discussed above and presented in Table 1.

There are two principal methods of improving the model design to enhance the validity of the results: (1) As discussed earlier, a control group

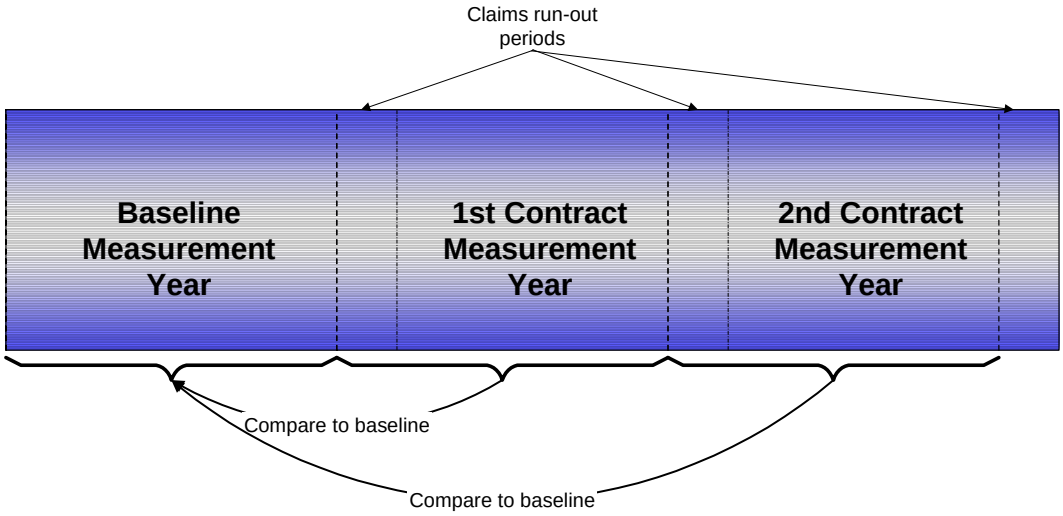

Figure 2. Conceptual model of the "total population approach" to DM program evaluation. Each contract measurement year is compared to the baseline, after a claims run-out is completed (usually with in 3-4 months after year end[4]

receiving no DM intervention can be selected for use as a comparison group. The easiest method is to match enrolled and non-enrolled members based on propensity scores. Because this can be done with administrative data, all enrolled members can be found a suitable matched control. A more costly and cumbersome alternative would be to select a random sample from the eligible, but not enrolled, population for use as a control group. However, this method may not be practical or financially feasible for most DM programs to implement, and may still be subject to selection bias if not performed correctly. (2) To decrease the measurement error, utilization or quality indicators should be used as outcome measures instead of cost, as is the current practice in DM evaluations. There are many reasons why costs can change from period to period, several which are difficult to predict or quantify accurately (e.g., the effect of member benefit design changes or the introduction of new technologies).[4] Conversely, utilization measures (e.g., hospitalizations, emergency department visits, etc.) are subject to less variability over time and the influence of bias. As will be discussed later in this paper, utilization can be transformed to estimates of cost for evaluating financial performance.

Survival Analysis

The advantage of survival analysis over the total population approach method is that it offers insight into the effect of disease process progression over time while providing the ability to measure the impact of secondary prevention techniques on these processes. More specifically, as shown in Figure 3, survival analysis can be used to determine how long it takes for the DM interventions to improve patient physiologic markers (i.e., HbA1c levels in diabetics, LDL levels in patients at risk for coronary artery disease, pulmonary function in asthmatics, etc.), and how long after that reductions in utilization and cost become evident.[13] Survival analysis

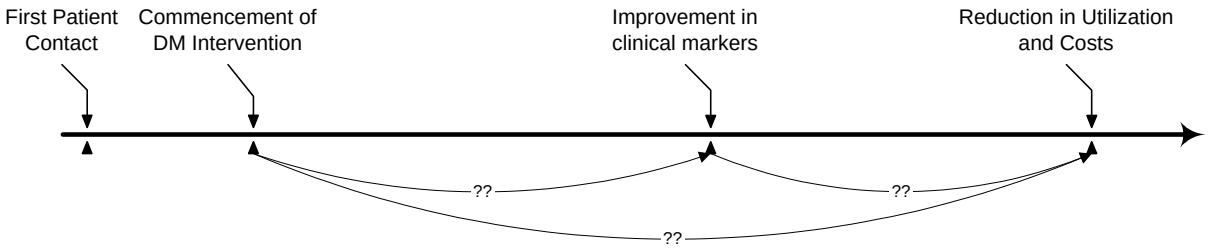

Figure 3. The intended process of a typical disease management program. The timelines still left undetermined are between the commencement of the disease management (DM) intervention and improvement in patient-level physiological markers and the translation of that improvement into reduced utilization.[13]

does not refer just to mortality outcomes. Any well-defined event can be analyzed using these methods. The more general term for this technique is “time to event analysis.”

What makes survival analysis unique among the many statistical methods is that data from patients who do not realize the “event” by the end of the study are included in the development of the model. These patient’s survival times are called *censored*, indicating that that the study period ended before the event occurred, or that the patient may have been lost to follow-up at some point during the study. In either case, the censored survival times are used along with the survival times of patients who ultimately experienced the event in order to construct the survival analysis model.[13]

An important feature of survival studies is that patients are enrolled at various points during the observation period and then followed until the end of the study. As such, patients enrolled near the end of the study will be followed for a shorter period of time than those patients enrolled early on, and thus will have a lower likelihood of experiencing the event. Nonetheless, it would be incorrect to assume that because these patient data are censored they have a better or worse prognosis than those patients who experienced the event after being observed for a longer period during the study. For example, a patient who “survived” the study for 1 year before its termination may not have a shorter survival time than a patient who was enrolled in the study for 2 years before ultimately experiencing the event.

The general principles discussed thus far support survival analysis as being one of the most powerful analytic tools available for evaluating DM program effectiveness. For example, DM programs manage a population of patients that move freely in and out of the program during the contract period (analogous to the end of a study). This is due to new enrollment or disenrollment in the health plan, recent identification of suitability based on disease state, members opting in or out of the program, or death. Using survival analysis, each of these patient’s censored data can be used to develop the model for predicting time to event. Additionally, the outcome event for a DM program can be either a utilization measure (e.g., time to first hospitalization from enrollment) or a clinical indicator (e.g. time to receipt of a lipid panel from first nursing intervention). Moreover, survival analysis allows for a comparison between cohorts and the inclusion of explanatory variables to assist in

determining which specific patient or program characteristics are related to better outcomes.[13]

Figure 4 illustrates how survival curves are typically displayed. In this example,[14] newly enrolled HMO Medicare members completed a health risk assessment called the P_{RA} test[15] where the results classified them as being either at low or high risk of hospitalization. A survival analysis was then performed to determine what the probability of hospitalization was for the low and high-risk cohorts over the course of the following 25-month period.

Each curve is illustrated as a step function, with each step corresponding to times at which a hospitalization was observed. The times of the censored data are indicated by X markers. Upon visual inspection we see that the probability of hospitalization in the high-risk group is about twice that of the low-risk group, reaching approximately 50% at month 25 of the study.

In order to make an inference about the population from which these cohorts of patients were drawn, several different statistical tests are available.

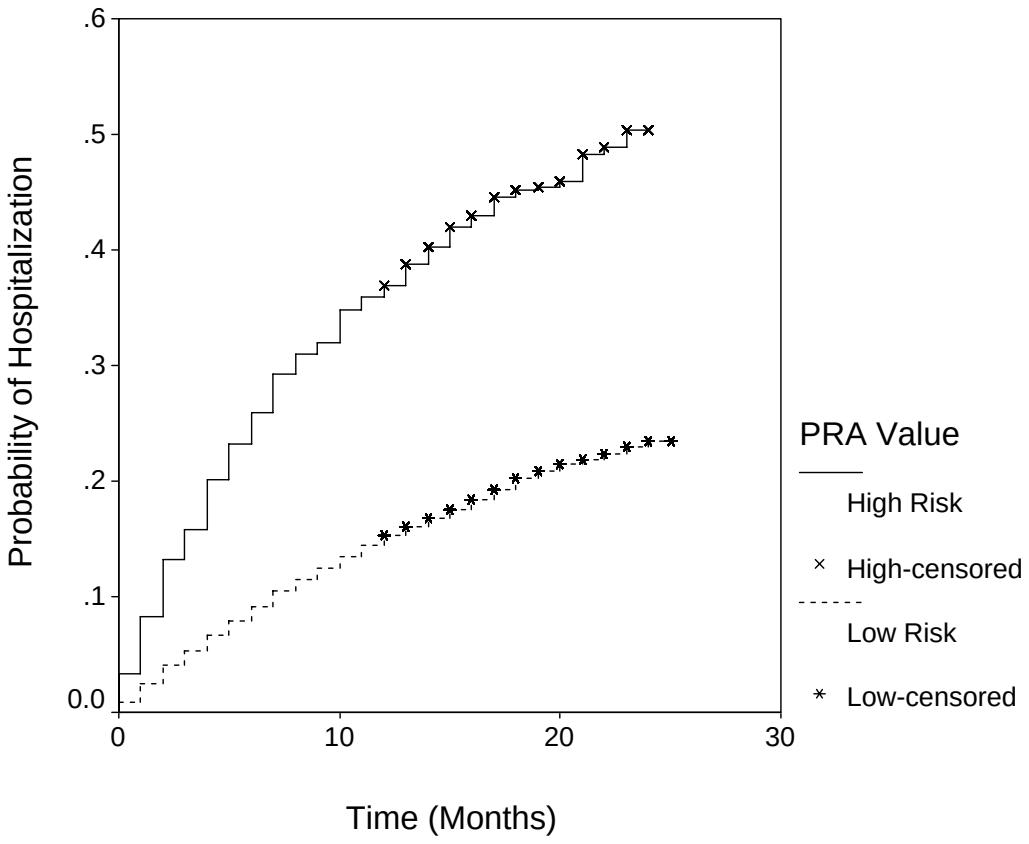

Figure 4. Cumulative survival probability curves, plotted for both low and high-risk subjects. As illustrated, there is an approximate two-fold likelihood of hospitalization in the high-risk cohort over the low-risk group[14]

Each of these tests compare the number of hospitalizations that was observed to number that was expected, which is calculated from the number of surviving members and the number of members who had the hospitalization at each period in the study.

To optimize the potential of survival analysis as an evaluation technique of DM program effectiveness the same rules of controlling for bias must be applied. Most notably, a control group should be used for comparing results. Here again, propensity scoring may prove to be the most favorable method.

In summary, survival analysis offers several advantages over the total population approach as an evaluation tool. One advantage is that the effects of the DM intervention can be quantified in terms of the timeline from the introduction of the intervention to a change in a clinical or physiologic marker, and from that point until the outcome is achieved. Another advantage is that there is a "built-in" control for loss to attrition. This model is designed to account for changes in enrollment status as typically experienced in DM programs. For a more comprehensive discussion of survival analysis, the reader is referred to references 13, 16-20.

Time Series Analysis

As discussed above, measuring changes that take place over time is an integral component of the DM program evaluation. The early months of a DM program are geared toward enrollment and initial patient assessments. Most DM programs have a phased enrollment process, bringing a percentage of eligible members into the program each month. Full program enrollment is usually not achieved until 3-6 months after program launch. If the intervention is effective at the patient level, it will not be evident until several months or even a year into the program. At that juncture, patients and providers should have been given the tools necessary to better manage the disease, as well as to ensure that clinical or physiological measures have achieved levels indicating control. In aggregate, over time, patient level improvements will manifest as population wide changes in medical utilization variables. At some point in time, intervention effectiveness may flatten or be reduced. Awareness of these temporal influences assist the DM program evaluator to identify, describe, explain, and predict the effects of processes that bring about change as a result of the program intervention.[21]

A time series can be simply defined as a variable that undergoes a repeated periodic observation, or measurement. The variable can be either at the patient level (i.e., EKG readings, respiratory rates, blood pressure, etc.), or at some aggregate level (i.e., hospitalization rates, mammography rates, etc.). The periodic measurement may be as short as a fraction of a second or as long a century. In general, time series analyses are used to characterize a pattern of behavior occurring in the natural environment over the measurement period, analyze fluctuations of the variable along the continuum, infer the impact of an intervention introduced during the measurement period, and forecast future direction of the time series variable.[21-24]

An important feature of time series is that of serial dependence. Any variable measured over time is potentially influenced by previous observations (autocorrelation). To take advantage of these relationships time series models use previous observations as the basis for predicting future behavior. This is the

essential difference between time series analysis and traditional statistical tests for measuring change, such as regression analysis, which rely on variation in independent variables to explain changes in the outcome. [21]

Some time series methods allow the DM program evaluator to predict future behavior of the observed variable without attempting to measure independent relationships that influence it. This is an extremely important point, since there are countless factors that may govern the behavior of the time series variable that cannot be identified or accurately measured. This last point indicates why time series analysis is a preferred design over the currently used total population approach for assessing impact of DM programs. [21]

Figure 5 illustrates three of the more prevalent time series patterns that may be identified in healthcare data: (a) *Trends*. This is the long-term movement of a data series that may slope either upward or downward. (b) *Seasonality*. This pattern emerges as spikes at regular intervals in a time series (usually monthly, quarterly or annually) and (c) *Stationarity*. A stationary series reflects data that have a constant variance around a constant mean. Therefore, a stationary series would not have a linear trend or seasonal component, but instead, would appear as relatively horizontal along the x-axis.[21]

While a visual inspection of a time series plot may clearly identify a linear trend or a seasonal effect, most healthcare data contain enough variability to lead to unreliable results using this method. However, this task can also be performed empirically using a statistical tool called the autocorrelation function (ACF), which will produce more accurate results. The underlying principle supporting the ACF is quite straightforward. As discussed earlier, what differentiates time series analysis from other statistical methods is that time series models are built on the premise of relationships between observations (as opposed to independent explanatory variables). An ACF simply indicates how each observation is correlated to prior observations (hence the term autocorrelation). A review of the ACF allows patterns to be detected. For example, the ACF in Figure 5 that corresponds with the time-series showing a trend, indicates a significant autocorrelation with the two most recent observations. Intuitively, this should make sense, since a trend is not a random occurrence but a movement in a specific direction. Therefore, we would expect to see a correlation between subsequent observations. Similarly, the ACF corresponding to the seasonal time-series indicates a significant autocorrelation with every twelfth observation (these data represent monthly values, therefore the seasonal effect appears one month every year). When the time series does not indicate the existence of a trend or seasonality, the ACF shows no significant autocorrelation. Again, this should make sense intuitively, since no trend indicates that time-series observations appear to be distributed randomly. Understanding the pattern in the data helps the analyst choose the appropriate model.

There are several different time-series models available to the DM program evaluator. The one that is chosen depends both on the pattern observed in the data and the statistical training of the evaluator. When the data appear stationary, *Simple Exponential Smoothing* (SES) is the preferred method.[21] *Double Exponential Smoothing* (DES) [25] is appropriate if there is a noticeable trend in the data but no obvious seasonality.[21]

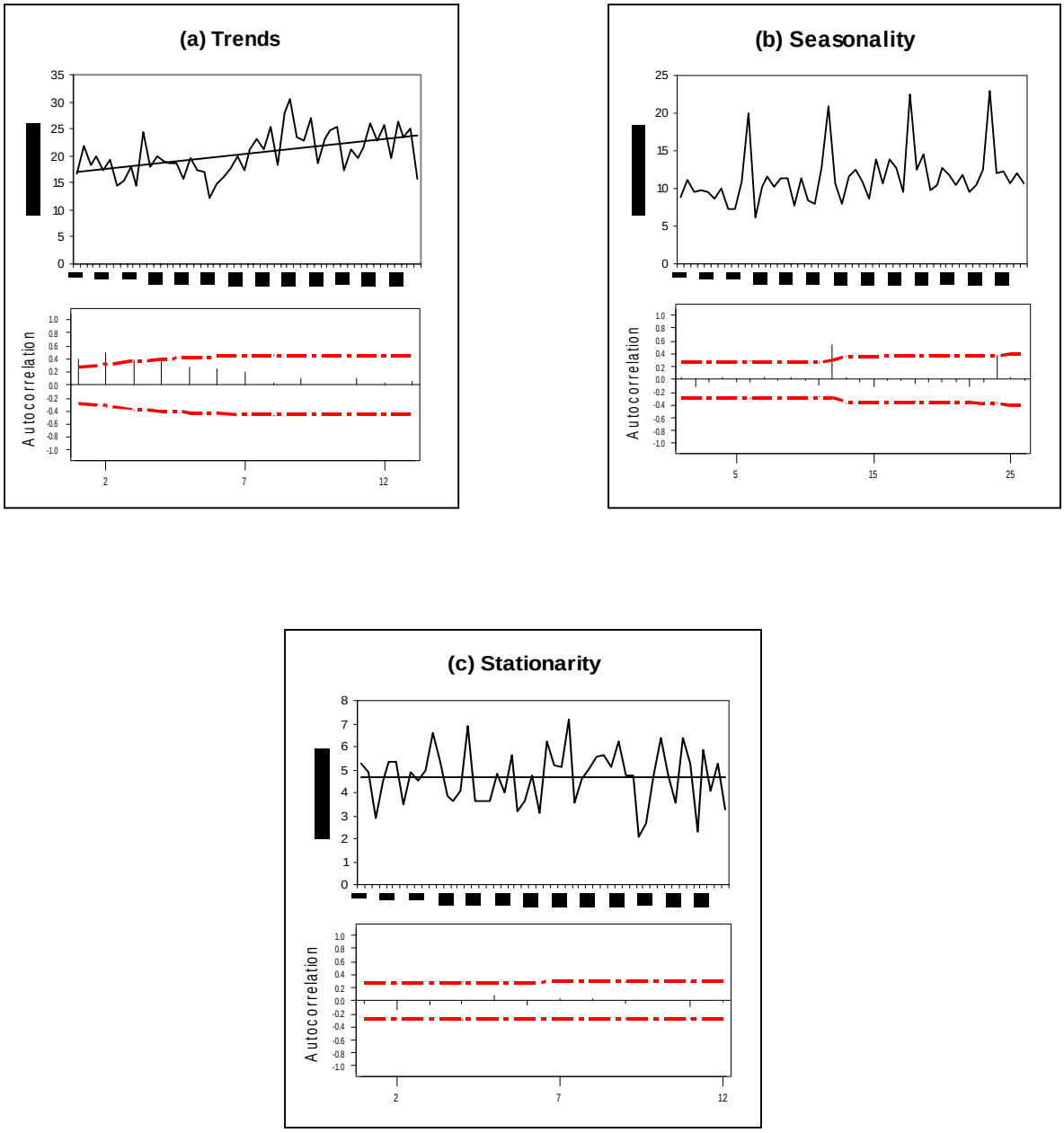

Figure 5. Three hypothetical time-series patterns that may be identified in healthcare data: (a) a long-term slope in the data represents a trend, either up or downward. This is represented in the autocorrelation function (ACF) by the current observation showing significant autocorrelation with the two most recent values. (b) Seasonality is identified by sharp spikes appearing at evenly distributed periods. The ACF shows a significant autocorrelation with the 12th and 24th observation. (c) A stationary series is identified as having a constant variance around a constant mean. The ACF illustrates this factor by indicating no significant autocorrelation between the current observation and any past ones.

The *Holt-Winters design*[26] includes variables to account for randomness, trend, and seasonality component of the time series. *The Autoregressive Integrated Moving Average* (ARIMA) design[27] is currently the most widely used time series methods found in the health services literature.[28-35] This, despite being the most complicated and theoretically based time-series models available. It is intended to describe, mathematically, the changes in a series over time.

Irrespective of which time-series design is chosen, the fundamental rules of application are identical; (1) choose an outcome variable that will reduce measurement bias (e.g. utilization measure in lieu of cost), (2) in developing the best-fit model, at least 50 observations should be analyzed. For a DM program, this would require at least 4 years of past data leading up to the month prior to the commencement of the intervention. This will allow the model to accommodate any patterns in the data that may impact the fitting parameters, (3) the forecast period should extend to no longer than the first 12 months of the program to ensure an accurate forecast horizon, and (4) a comparison of actual versus forecasted values should be performed so that results have a tangible meaning to the program evaluator.[21]

Figure 6 illustrates how to use a time series design. As shown in this hypothetical data set; (1) observations from the historical data set were used to develop the time series model (2), forecasts were then produced for the test period or baseline (which is the 12 months immediately prior to the commencement of the DM program) (3), actual values were compared to predicted (in these data, the best fitting model was determined to be the SES based on that model producing the lowest percentage error compared to the other models), (4) the model was then recalculated adding the actual observations from the baseline period to the historical set, (5) forecasts were produced for the first measurement year, (6) at the end of that program year actual utilization was compared to the forecasted values. In this example, the admission rates appeared to be 3.6% higher than what was predicted for the period, indicating that the DM program was not successful in reducing admissions in this period.

In summary, time series analysis is a suitable model for evaluating DM program effectiveness and its uses should be further explored. It controls for most biases simply because it monitors behavior of observations over an extended period of time (at least 50 data points are used for establishing a historical period). This eliminates the concern of regression to the mean and selection bias. Additionally, outcome measures are aggregated to the population level (e.g. hospitalization rates per thousand, etc), thereby allowing for ease of observation in the natural environment without needing a control group for comparison. A forecast is made for the following 12-month period, and at the end of that measurement period, the actual data is compared to the forecasts. Using this process the analyst can measure whether the program achieved the desired result. For a more comprehensive discussion of time-series analysis, the reader is referred to references 21-24.

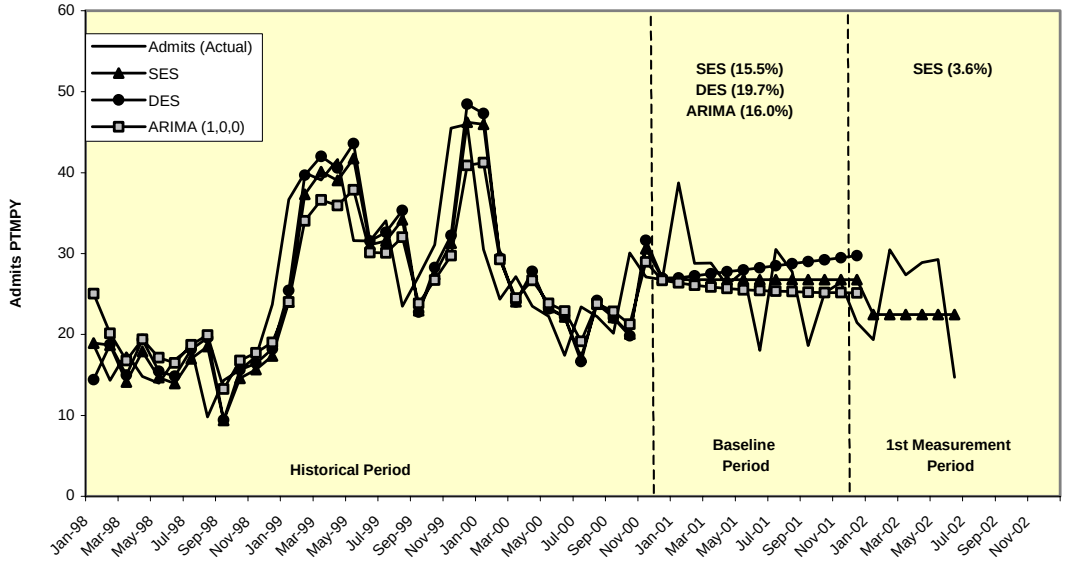

Figure 6. An illustration of how time series analysis can be used for evaluating DM program effectiveness. The observations are disease specific hospital admissions (per thousand members per year). [21]As illustrated, three different models were fitted to the historical data and were used to forecast into the baseline 12-month period. The best-fitting model (that which has the least percentage difference between the actual and forecasted value) was chosen as the model that will be used for future forecasts. In this example, simple exponential smoothing (SES) had the best fit. A comparison of actual to forecasted values during the 1st measurement period indicated a 3.6% higher than expected admission rate (PTMPY), indicating that the program was not effective during the first 6-months of the 1st measurement period.

DETERMINING FINANCIAL PERFORMANCE USING DISEASE MANAGEMENT PROGRAM EVALUATION MODELS

For most DM vendors and their clients, the final determination of program "effectiveness" is based on a determination of whether or not medical cost for the given population was less than what would be expected had there been no DM intervention. As discussed earlier in the paper (and presented in Table 1), using cost as an outcome variable can expose the results of the evaluation to measurement bias. Cost of medical services is comprised of three elements (1) the type/mix of services used, (2) the amount of services used (i.e. utilization) and (3) the price paid (i.e. unit cost) for services used. Typically, disease management interventions are aimed at impacting the amount of services used (e.g. the number of ED visits or hospital admissions). Less commonly DM interventions impact the type/mix of services (e.g. chemotherapy provided on an out-patient versus in-patient setting or encouraging medical therapy over surgical for low back pain). Rarely do DM programs have any impact or influence over unit cost of services. Factors such as reimbursement models, payment rates, benefit design and covered services all impact unit cost and are all outside the influence of DM interventions.

While it is possible to adjust for many factors that may impact cost (such

as annual provider payment rate increases and changes in member cost sharing), there are many more causes that may be obscured and therefore cannot be controlled for. For example, the effect of new technologies, or a change in the use of existing ones, may be impossible to explicate when only reviewing aggregate cost. Despite these limitations, reconciliations between DM vendors and their health plan partners continue to be based on cost.

This section will provide suggestions on how to apply the results derived from those DM evaluation models described above, to determine financial performance of the program. The two methods by which to gauge financial performance is (a) directly, using cost as the outcome measure, and (b) indirectly, translating a utilization or clinical outcome into a financial outcome.

Direct Measurement of Cost

In order to conclude that a DM program did in fact impact costs, the evaluation must negate or control for the measurement bias associated with this variable. In the Total Population Approach, using a properly matched *concurrent* control group is the only method available to ensure that the realization of lower costs is indeed a result of the intervention. Historic controls may be subject to very different costs, some of which may not be identifiable for adjustment. Therefore, for financial results to be valid using a pre-post test model, measurement bias must be controlled and a tightly-matched control group must be selected.

Time series analysis can also estimate financial performance. After each measurement period is completed, actual normalized aggregate costs can be compared to forecasted costs with the difference indicating whether a savings effect was achieved. Again, bias may be introduced if any part of the cost variable was measured differently from period to period, or if no adjustment was made to account for inflationary factors, changes in reimbursement or benefit package, etc.

Survival analysis may be tailored to estimate financial performance if the outcome variable is designed to estimate the probability of hitting a given cost threshold. For example, a DM program evaluation may assess the median number of months it took to reach a $25,000 threshold in total costs per enrolled member compared to a control group. Program success would be indicated by a longer period of time until the intervention group reached that threshold. The difference in those costs over that time period can be easily computed.

Indirect Measurement of Cost

The basic premise of disease management is that by influencing the adoption of evidence-based medicine by providers and providing patient health education and case-management support, patients will be less likely to suffer an acute exacerbation of their illness requiring an emergency department (ED) visit or hospitalization. This concept is supportive of the notion that utilization measures should be chosen as the outcome variable as opposed to cost.

More specifically, the utilization measures selected should flow directly from the intended impact of the intervention. For example, a common intervention in DM programs involves education and coaching of patients to monitor daily vital signs and symptoms (e.g. weight, blood pressure, difficulty breathing, etc) and to report these daily

measures to the DM program. The DM program nurses monitor these self-reported measures for concerning trends or triggers. The intended impact of this intervention is to identify problems early enough for outpatient intervention (e.g. altering medication dosage) thereby avoiding an ED visit or hospitalization. Therefore, It is both logical and plausible that an effective DM program should result in lower rates of ED visits and hospital admissions than without such intervention.

Additionally, health plans should require the DM program administrators to specify where and how they intend to impact population medical costs. Intervention and enrollment strategies should then be examined to determine if it is logical and plausible that the program could result in utilization improvements consistent with the claimed cost impact. It is also important to identify areas where the DM program interventions will intentionally increase utilization. For example, improving the management of many chronic diseases involves increasing the appropriate use of drug therapies. For populations with a drug benefit, this increased utilization offsets the savings from reduced utilization in other areas. To summarize, if the measurement variable targeted for intervention is different from the outcome measure, measurement bias is introduced. To circumvent this issue, utilization measures should be compared vis-à-vis the chosen evaluation model and then financial performance be extrapolated from the results.

In the Total Population Approach, using utilization outcome variables allows the evaluator to match DM program enrollees to either a *concurrent* or *historical* control group, since utilization is less susceptible to bias than cost over time. Once the evaluation has been completed between the groups, present day costs can be assigned to the utilization rate to derive at an estimate of financial performance. In survival analysis, a marginal cost effect for the probability of hospitalization (or other utilization measure) can be estimated and compared across enrolled and control groups. In time-series analysis utilization measures can be translated into estimated cost for each observation period and compared to the actual utilization value and its estimated cost.

A simpler, more novel approach for contractual purposes is to determine whether a pre-established target is met. Each incremental percentage reduction in utilization is worth a given reimbursement amount. For example, each percentage point decrease in hospitalizations may be worth X dollars bonus to the DM vendor above a set fee limit. Conversely, an increase in hospitalizations may lead to reimbursement of fees to the health plan or some other payment penalty.

In summary, there are various ways to measure financial performance of a DM program using the evaluation tools described in this paper. Measuring costs directly as an outcome variable can be performed when the related biases can be identified and controlled for. However, the preferred method is to use utilization outcome measures as a proxy for financial performance. This is the most reliable way to maintain the integrity of the outcome measurement by reducing the impact of bias.

CONCLUSIONS

This paper provides an overview of several evaluation methods that can be used for assessing DM program effectiveness. Before an evaluation is launched, the evaluator must ensure that

the design chosen best fits the data and mitigates the effects of bias that may raise concerns about the validity of the results. One bias of particular concern in DM programs is measurement bias, which results from the intervention focusing on a reduction in acute health service utilization yet use cost as the outcome measure for assessing program effectiveness. An alternate and more appropriate method would use utilization as the outcome measure and then transform the results to financial performance based on current indices.

REFERENCES

1. Carroll J. Health plans demand proof that DM saves them money. *Managed Care.* 2000;9(11):25-30.
2. Diamond F. DM's motivation factor can skew study results. *Managed Care.* 1999;8(6):45-50.
3. Johnson A. Measuring DM's net effect is harder than you might think. *Managed Care* 2003;(6):28-32.
4. Linden A, Adams J, Roberts N. An assessment of the total population approach for evaluating disease management program effectiveness. *Disease Management* 2003;6(2): 93-102.
5. Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.
6. Cook TD, Campbell DT. Quasi-experimentation: design and analysis issues for field settings. Chicago: Rand McNally College Publishing Company, 1979.
7. Shadish SR, Cook TD, Campbell DT. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston, MA:Houghton Mifflin, 2002.
8. Cousins MS, Shickle LM, Bander JA. An introduction to to predictive modeling for disease management risk stratification. *Disease Management*. 2002;5(3):157-167.
9. Dehejia RH, Wahba S. Propensity score-matching methods for non-experimental causal studies. *Rev Econ and Stats* 2002;84:151-161.
10. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41-55.
11. Rosenbaum P, Rubin D. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Stat* 1985;39:33-38.
12. Cellini GL, Disease management – the reconciliation blues. Available at: http://www.healthleaders.com/news/feature1.php?contentid=39481&CE_Session=e53af34ff8342f37dc0fa4cd29b04690. Accessed December 3, 2002.
13. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: An introduction to survival analysis. *Disease Management* (under review).
14. Linden A, Schweitzer SO. Applying survival analysis to health risk assessment data to predict time to first hospitalization. *AHSRHP Annual Meeting*. 2001;18:26.
15. Boult C, Dowd B, McCaffrey D, Boult L, et al. Screening elders for risk of hospital admission. *J Am Geriatr Soc* 1993;41:811-7.
16. Kaplan E, Meier P. Nonparametric estimation from incomplete observations. *J Amer Statist Assoc* 1958;53:457-481.
17. Cox DR, Regression models and life tables (with discussion). *J R Statist Soc B* 1972;34:187-220.
18. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br J Cancer* 1977;35:1-39.
19. Kalbfleisch JD, Prentice RL. *The statistical analysis of failure time data*. New York: Wiley; 1980.
20. Andersen PK. Testing goodness-of-fit of Cox regression and life model. *Biometrics* 1982;38:67-77.
21. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: An introduction to time series analysis. *Disease Management* (in print).
22. Chatfield C. The analysis of time series, 5th edition. London: Chapman and Hall, 1996.
23. Makridakis SG, Wheelwright SC, Hyndman RJ. Forecasting: methods and applications, 3rd edition. New York, NY: John Wiley and Sons, 1998.
24. McCleary R, Hay R. Applied time series analysis for the social sciences. Beverly Hills, CA: Sage, 1980.

25. Holt CC. Forecasting seasonal and trends by exponentially weighted moving averages. Office of Naval Research, Research Memorandum No. 52, 1957.
26. Winters PR. Forecasting sales by exponentially weighted moving averages. *Management Science*. 1960;6:324-342.
27. Box GEP, Jenkins GM. Time series analysis: forecasting and control. San Francisco, CA: Holden Day, 1976.
28. Choi K, Thacker SB. An evaluation of influenza mortality surveillance, 1962-1979. I. Time series forecasts of expected pneumonia and influenza deaths. *American Journal of Epidemiology*. 1981;13:215-226.
29. Gibson E, Fleming N, Fleming D, et al. Sudden death syndrome rates subsequent to the American Academy of Pediatrics supine sleep position. 1998;36(6):938-942.
30. Haines LM, Munoz WP, Van Gelderen CJ. ARIMA modeling of birth data. *Journal of Applied Statistics*. 1989;16:55-67.
31. Linden A, Schweitzer SO. Using time series ARIMA modeling for forecasting bed-days in a Medicare HMO. *AHSRHP Annual Meeting*. 2001;18:25.
32. Martinez-Schnell B, Zaidi A. Time series analysis of injuries. *Statistics in Medicine*. 1989;8:1497-1508.
33. Tsouros AD, Young RJ. Applications of time-series analysis: a case study on the impact of computer tomography. *Statistics in Medicine*. 1986;5:593-606.
34. van Walraven C, Goel V, Chan B. Effect of population-based interventions on laboratory utilization: a time series analysis. *JAMA*. 1998;280(23):2028-2033.
35. Zechnich AD, Greenlick M, Haxby D, Mullooly J. Elimination of over-the-counter medication coverage in the Oregon Medicaid population: the impact on program costs and drug use. *Medical Care*. 1998;36(8):1283-1294.

Address reprint requests to:
Ariel Linden, Dr.P.H., M.S
President Linden Consulting Group
6208 NE Chestnut Street
Hillsboro, OR 97124

E-mail: ariellinden@yahoo.com