Vol. IV, No. 5
November 2003

STATE COVERAGE INITIATIVES

issue brief

Evaluating ROI in State Disease Management Programs

by Thomas W. Wilson

In light of soaring health care premiums and plummeting state revenues, many states are turning to disease management (DM) programs as a means of controlling spending in their Medicaid programs and high-risk pools. DM is believed to help prevent major disease events (e.g., stroke or heart attack) and thus reduce the costs associated with hospitalization and other medical services for such events.

DM does not lead to instantaneous savings. Therefore, choosing to spend money on DM is truly an investment that is hypothesized to lead to a return at a later time. DM programs are designed to encourage prevention and regular monitoring of patients with chronic disease—which, in turn, is thought to improve health and consequently lower health care resource use. If the fees that states must pay to develop a DM program, or to hire a vendor to implement one, are less than the money they save from decreases in the use of inpatient and outpatient care, prescription drugs, and other services, then states have earned a positive return on investment (ROI).

The big challenge for policymakers is to prove that DM both improves health and yields an economic return. "Those who are paying the health care bills are increasingly anxious for DM to show a return on investment,"[1] says David Nash, M.D., associate dean for health policy at Jefferson Medical College.

This issue brief addresses how program managers can determine, in a credible way, whether the investment in DM has led to a positive ROI. It introduces two basic principles—comparability and equivalence—that managers must address when evaluating the economic impact of DM on Medicaid and high-risk pools.

The ROI Problem

Measuring the financial return associated with DM is difficult because changes in health care costs over time cannot be assumed to be solely due to the DM intervention in the population that received DM services. Numerous external factors could have played a role as well. For example, costs could have dropped in the DM population because that group had been exposed recently to a heavily promoted new drug, or because they experienced a change in benefit design, or a number of other external factors.

To address this problem, analysts must ensure that they compare their DM population to an appropriately chosen and measured reference population. Measurements on this population will allow them to answer the following essential question: What would have happened to the DM population's health and health care resource use had they not received that intervention?

Key Principles

To ensure that the reference group to whom the DM population is compared is appropriately chosen and measured, two principles must be followed:[2]

1) The risk of experiencing the economic outcome over time in the absence of the DM program must be *equivalent*

AcademyHealth
Advancing Research, Policy and Practice

AcademyHealth is the national program office for SCI, an initiative of The Robert Wood Johnson Foundation.

between the reference and DM intervention populations.

2) The metrics used to assess that risk and measure the economic outcome must be *comparable* in both the intervention and reference groups.

Equivalent comparison groups are like identical twins who have had similar life experiences. Because the two groups were so alike at the outset, any changes observed between the group subsequently given DM and the one not given the intervention can be assumed to be due to DM. The numerator (e.g., the percentage of people with a given outcome) and denominator (e.g., the total number of people initially at risk for that outcome) must be defined the same way in both groups. When the two groups are both equivalent and comparable, ROI can be calculated in a credible way.

Achieving Comparability

Defining the Population

Which criteria should be used to define the initial denominator for all principal metrics? In other words, what is the best way to identify the population at risk for experiencing a certain event in the future? Should they be characterized by age, diagnosis, certain drugs, enrollment period? As an example, a population of children and young adults at risk for an asthma attack could be defined in the following way: all individuals below the age of 21 who have had a medical claim for asthma and a pharmacy claim for an asthma drug and were continuously enrolled in a Medicaid plan for one calendar year and identified by administrative claims and membership data.

Whatever criteria are chosen to represent the population at risk, they must be used to classify the denominators for both the DM intervention and reference groups. Thus, if a patient is excluded from the DM group based on the chosen criteria, that same patient would have to be excluded from the reference group had he or she been part of that population.

That sounds more straightforward than it is. For instance, using the asthma example mentioned earlier, a significant number of individuals in the DM group may state to a nurse on a telephone call that they do not have asthma. Although it would seem logical to exclude them from a DM intervention aimed at preventing asthma attacks, these individuals should not be excluded from the ROI analysis (however, they can be excluded from the intervention) because they still meet the formal definition for the population at risk (under 21, with medical and pharmacy claims for asthma and asthma-related medication, and a one-year Medicaid enrollment). The likely practical consequence of differentially excluding individuals from the intervention group and not from the reference group is that the ROI calculation will be incorrect.[3]

Table 1: Key Disease Management Metric Types

Type #	Description	Example	Numerator / Denominator
Type I	Intervention	Enrollment statistics in diabetes program	People contacted / Total defined population
Type II	Intermediate or Proximate Outcome	HbA1c screening in a diabetes population	People with HbA1c screening / Total defined population
Type III	Ultimate Outcome	Amount spent on health services in a diabetes population	Total claims dollars spent / Total defined population

Defining the Intervention and Outcome Metrics

Measures used in ROI analyses should include an "intervention" metric (also known as a Type I metric) and an "outcome" metric or metrics (there are two kinds of these: a proximate outcome, or Type II, metric and an ultimate outcome, or Type III, metric).

An intervention metric measures an aspect of the DM intervention and might be calculated as follows: Of all the people in the defined population (i.e., the denominator), how many of them were actually contacted? Let's suppose that, of 1,250 people in the defined population (i.e., the denominator), 1,000 were contacted (i.e., the numerator). This would mean that the "percent identified" would be 1,000 divided by 1,250 (0.8 or 80 percent). These metrics may be used to help understand the cost or investment side of the ROI calculation. Type I metrics are unique in that they are typically not used in the reference group, unless that population has also had some kind of intervention.

To calculate the return part of the ROI calculation, at least one outcome metric is essential and must be measured in both groups. Proximate outcomes assess the number of screening tests or other measures that correlate with a certain health outcome, and ultimate outcomes evaluate the final financial impact of the intervention. An example of a proximate outcome metric is the number of patients who have had lipid screening among a defined population of diabetics, while an ultimate outcome metric might measure the number of inpatient admissions or total claims costs divided by the total number in the defined population.

For any Type II or Type III metric, it is essential that the numerator and denominator for both the intervention and reference populations are defined in precisely the same way. In practical terms, this means that outcome metrics can only be used when the same information is available to calculate the numerator and denominator in both populations. Table 1 summarizes the three major types of metrics.

A fourth kind of metric, called "confounding variable metrics," represent potential factors other than the intervention that could influence outcomes; they could include age, gender, co-morbidities, or drugs, for example. These metrics are important in assessing equivalence, the topic of the next section.

Equivalence in Populations

In the absence of the DM intervention, it is essential that the reference population's risk for a certain outcome is equivalent to the risk in the intervention population. That is not to say that everything must be identical in the two populations; the key is ensuring that the factors other than the intervention that could independently influence the outcome are the same.

This "equivalence principle" is the most important factor to consider when estimating ROI. Equivalence can be achieved by selecting an equivalent reference group and/or adjusting for differences in key predictor variables during the analysis of ROI. If equivalence is achieved, the ROI calculation compares what happened to the DM population's health and health care resource use compared to what was expected to happen in the same population in the absence of DM. Some common confounding factors include inflation, age, and gender.

For example, it is necessary to account for the rising price of goods and services over time if one is conducting an analysis that uses a population measured in the prior year as a reference group. Because medical inflation would have occurred in the absence of the DM intervention, the reference and intervention groups are non-equivalent. Similarly, because age and gender can each independently predict disease onset and progression in many cases, it is important to ensure that the age and gender distribution of the reference and intervention groups are roughly equivalent.

There are other, more pernicious influences that could seriously jeopardize equivalence.[4] Thus, it is useful to carefully examine each confounding variable, as all situations are different. Content experts in the disease of interest or in quantitative analysis should be consulted to help ensure equivalence in both groups.

Figure 1: Similar Random Background Fluctuations in Intervention and Reference Groups

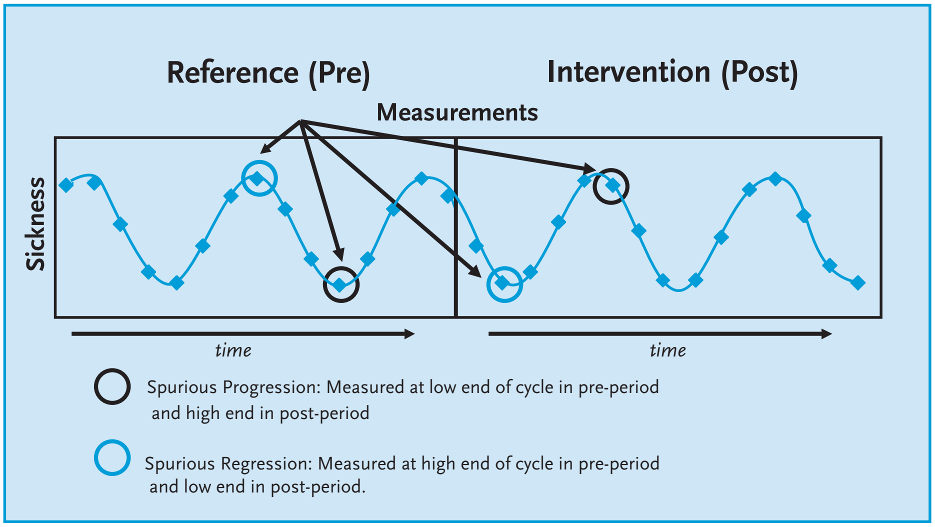
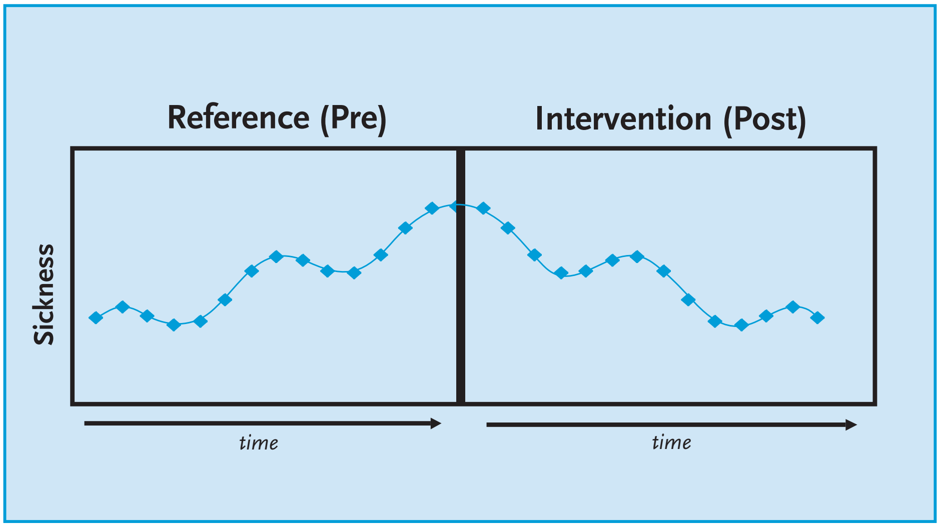

Reference (Pre)
Intervention (Post)
Measurements
Sickness
time
time
○ Spurious Progression: Measured at low end of cycle in pre-period and high end in post-period
○ Spurious Regression: Measured at high end of cycle in pre-period and low end in post-period.

Figure 2: Background Non-Random/Systematic Difference Between Intervention and Reference Groups



As illustrated in Figure 1, a population may experience background fluctuation in a population metric over time that is identical in both groups. Thus, it appears that equivalence is achieved. However, non-equivalence can arise if the selection of patients is biased. This may occur, for example, when individuals are selected (or self-select) into a DM program based on their location in the fluctuation curve. Let's assume that the fluctuating metric is cost in a population with otitis media. Some individuals may be selected when they happen to be at the upper part of the curve. The next time they are measured, many will likely be at a lower part of the curve. Thus, a decline in costs is measured. However, this reduction does not represent a decrease due to the intervention; rather, it is just part of the standard fluctuating cycle. This phenomenon is called regression to the mean.

Figure 2 demonstrates a more serious problem. It shows a systematic background rise in the average risk in the reference group and a systematic fall in the average risk in the intervention group. This could occur if the intervention group had access to a new drug (unrelated to the DM intervention) that improved health and lowered the overall costs, but the reference group did not. This might occur in a study with a concurrent reference group—if, for example, the intervention group was given the drug as part of a promotion. It might also occur in a pre-post design that uses a year when the drug was not as widely available as in a reference period. In this case, the trend projected from the pre-period (the reference group) would overestimate the trend to be expected in the intervention group had they not had the intervention. ROI would also be overestimated.

The Analysis

Selecting the Reference Population

There are two major categories of reference populations that can be used in ROI analyses: 1) historical comparison groups, which are comprised of individuals chosen from a prior time period; and 2) concurrent comparison groups, which include people chosen from the same time period.

Historical controls can be a different group of people than those who receive the DM intervention (e.g., benchmark designs, case-control studies). They can also be the same individuals who subsequently receive the DM intervention (e.g., pre-post studies, which use patients as their own controls) or some combination (e.g., pre-post studies that include the entire population in both the pre and post period, with some overlap). In fact, a review of DM services in the managed care industry concluded that the pre-post design (mostly using patients as their own control) using administrative claims data is the most prominent design used in the managed care industry.[5] Recent work in the Medicaid community indicates that the population-based pre-post design is growing in popularity for DM analyses.[6]

The reasons for the widespread use of this design were not discussed, but one explanation for its popularity may be that it is relatively easy to conduct pre-post studies because they do not require an external reference group. A recent white paper recommended that the population-based pre-post design is "the most practical and appropriate method to measure DM program results at this time."[7] Although it may be practical, the design is subject to numerous biases unless equivalence is achieved between the pre and post periods.[8]

Designs that use the concurrent comparison groups have one advantage over those that use historical references: They can more easily account for confounding variables caused by "secular trends" that were not in existence before (e.g., introduction of a new drug, health policy changes, etc.).

The most common concurrent design used in DM is a participant/non-participant design,[9] which compares a group of study participants given the DM intervention with a defined subset of the general population. It is well recognized that participants may be quite different from non-participants and thus the results could be seriously confounded unless all major non-equivalences are taken into account. Other examples of designs that use concurrent reference populations include cross-sectional studies, prospective cohort studies, and randomized controlled trials (RCT).

The great advantage of the classic, individual-level RCT is that it is the design that is most likely to achieve equivalence because individuals are randomly assigned into the intervention and reference groups.[10] None of these designs are without potential bias, however. Some pharmaceutical products that have passed the rigor of the RCT were later found to produce serious side effects when marketed to the general public.[11]

Addressing Non-Equivalence

There are many ways to address the common problem of non-equivalence between the DM intervention and reference populations.

Some forms of regression to the mean (the problem highlighted in Figure 1) are due to selection bias and should thus be addressed at the selection stage. If a pre-post study design is used, one must measure the entire selected population in the period before and after the intervention, not just those who choose to participate in the DM program. As explained earlier, the measurement of participants only can result in a situation of manufactured non-equivalence. Systematic fluctuation in a population (as seen in Figure 2) can be addressed at the selection stage as well. A reference group may have to be chosen that has the same anticipated fluctuations in risk or costs as the DM intervention.

In most cases, non-equivalence must be focused on at the analytic stage. A rise in prices over time, for example, can be taken into account by using some agreed-upon inflation adjuster. The problem is determining which adjuster to use. The published data from the Bureau of Labor Statistics on medical inflation from the Consumer Price Index is highly recommended.[12] By itself, this adjuster may not be sufficient, but it should still be used as a standard. Alternative methods may include average price increase of claims in the health plan, insurance premium price increases, and actual price inflation in a concurrent (and equivalent) reference group. Aside from inflation, other factors that can influence pricing are changes in practice patterns, cost of new drugs, public policy changes, and changes in business processes (e.g., prior authorization for hospitalization); these are all more difficult to take into account.

Age and gender can be adjusted for in the selection stage by strict age or gender criteria for defining the population or in the analysis stage. In the analytic stage, one fairly straightforward method is to divide or stratify the population into age-sex groups (e.g., males, aged 45 to 65) and compare outcome metrics by subgroup in the reference and intervention populations. Other important confounding variables—including demographic factors, time in natural history of disease cycle, specific clinical factors (e.g., ejection fraction in heart failure patients), and so forth—can also be stratified in this fashion. Sophisticated matching or modeling techniques, some of which can help address large numbers of confounders, are available as well. Many of these methods require the use of an expert. However, it is difficult to adjust for confounders that are not measured.

Calculating ROI

There are two general approaches to calculating ROI: direct and indirect. A direct assessment of ROI uses primary data only. It must include at least one ultimate outcome metric that is available in both the intervention and reference populations. To substantiate the ROI estimate, it is strongly recommended that measures for at least one proximate outcome metric be taken in both groups as well. Assuming the clinical metrics change in the same direction as the financial metric, this will provide some succor that the financial impact was paralleled by a change in a clinical metric.

An indirect ROI assessment is one that uses secondary data, such as that found in a benchmark-type design. This analysis must include at least one proximate outcome metric; the ultimate outcome is inferred. An example of an indirect assessment would be the imputed savings of lowering blood pressure over a five-year period, based on an acceptable formula for calculating savings such as the Framingham Multiple-Risk-Factor Assessment Equation.[13] It should be noted that indirect ROI assessments are easily biased by both non-equivalence and lack of comparability.

The estimation of ROI requires comparing the cost differences between the intervention and reference groups on the ultimate outcome metric (Type III) divided by the DM program cost. For example, if program costs were calculated at $50,000, and the cost difference between the intervention and reference groups was $100,000, ROI would be $100,000 divided by $50,000, or 2:1. In other words, for every dollar invested in DM, two were returned.

ROI is always an estimate, and there are many biases that can influence it. Even tests to ensure statistical significance—which are recommended—address only one type of bias (random sampling error); they cannot control for most factors capable of confounding the results.[14]

Interpretation Stage

For any study, pure objectivity and independence is a myth; therefore; one should still assess the extent to which potential conflicts of interest might compromise the results. Ideally, the evaluator is: 1) qualified, 2) has little or no conflict of interest bias prior to the study, and 3) not subject to any pressure, whether direct or indirect, from the client during the study. To help ensure that these criteria are met, the expert conducting the analysis should sign a conflict of interest statement.[15]

The optimal arrangement for preventing outside influences on research findings is termed an "unrestricted educational grant" in the academic world. The most troubling scenario occurs when an evaluator is instructed by the client about which methods to use and how the results should be disseminated. Even though such an analysis may be scientifically valid, the objectivity of the findings will likely be questioned by the public.

There will never be an ROI study that cannot be improved. Thus, the analysis should include a discussion of the study's strengths and weaknesses. Was an outside expert consulted to discuss metric comparability? How successful were the methods used to address non-equivalence? Were important risk factors measured and accounted for? If equivalence was not achieved, what is the likely impact on the ROI estimate?

Given the uncertainty associated with DM analyses, it is probably not possible to "prove" that DM positively affects ROI by the legal standard of "beyond a reasonable doubt." A more reasonable goal is to aim for a "preponderance of evidence." That means that ROI assessments should not be based on a single study. Rather, evidence should be refreshed constantly with new data. This will assure those who pay the health care bills that the investments they made months or years earlier were intelligent ones.

About the Author

Thomas W. Wilson, Ph.D., M.P.H., Dr.P.H., is an epidemiologist and principal with Wilson Research, LLC in Loveland, Ohio. He has worked in the private sector since the mid-1990s, and is a frequent speaker at health care industry conferences on ROI evaluation and program improvement. He can be reached at twilson@wilsonresearch-llc.com or 513.289.3743.

Endnotes

[1] Nash, D.B. "Managing At-Risk Populations: Disease Management: What Does the Research Evidence Show?" *Drug Benefit Trends*, December 2002.

[2] For a more detailed explanation of these principles, see: Wilson, T., and M. MacDowell, "Framework for Assessing Causality in Disease Management Programs: Principles," *Disease Management*, 2003, Vol . 6, pp. 143–58. Research for this paper was partially funded by an unrestricted educational grant from the Disease Management Association of America.

[3] For practical recommendations to avoid errors of "manufactured non-equivalency," as well as implications for contracting, see: NMHCC Workgroup: Gruen, J. et al. "Crossing the Measurement Chasm: Evaluating Disease Management Measurement Methodologies," presented at the 8th Annual Disease Management Congress, San Diego, California, September 30, 2003. Accessed October 10, 2003, at www.jacksonmediagroup.com/DMC/.

[4] Gold, M.R. et al. *Cost-effectiveness in health and medicine*. New York: Oxford University Press, 1966.

[5] Johnson, A. Disease Management: *The Programs and the Promise*, Milliman USA, 2003. www.milliman.com/health/publications/research_reports/hrr14.pdf. Accessed August 1, 2003.

[6] The RFP to introduce DM to the New Hampshire Medicaid program uses this pre-post method, with adjustments (New Hampshire Department of Health and Human Services Office of Health Planning and Medicaid. Request for Proposal. Disease Management Program. January 1, 2004 – December 31, 2006. www.dhhs.state.nh.us/DHHS/OOF/LIBRARY/default.htm). The results of the Pfizer DM initiative in the Florida Medicaid program also used as one of their methodologies a pre-post study of participants ("Florida DM Medicaid Saves 15.9 Million in First Year," *DM News*, July 11, 2003. Vol. 8, No. 18, p. 1).

[7] American Healthways, Johns Hopkins University. "Standard Outcomes Metrics and Evaluation Methodology of Disease Management Programs." 2nd Annual Disease Management Outcomes Summit, November 7–10, 2002, Palm Springs, California. American Healthways Inc., Nashville, Tenn.: 2003.

[8] Linden, A. et al. "An assessment of the total population approach for evaluating disease management," *Disease Management*, Vol. 6, No. 2, Summer 2003, pp. 93–101 for article on bias of pre-post study. These "threats to validity" include member-based (e.g., selection bias), health plan-/program-based (e.g., unit cost increases), physician-/provider-based (e.g., Hawthorne effect), data-based (e.g., sensitivity/specificity), measurement-based (e.g., reliability), and general issues (e.g., secular trends).

[9] Johnson, A. Disease Management: *The Programs and the Promise*, Milliman USA, 2003. www.milliman.com/health/publications/research_reports/hrr14.pdf. Accessed August 1, 2003.

[10] There are other RCT designs where the individual is not the unit of randomization. "Field-based" RCTs randomize groups by time or place. They are not as likely to automatically achieve equivalence as is the classic individual-level RCTs.

[11] Food and Drug Administration. Center for Drug Evaluation and Research (CDER) 2002 Report to the Nation: *Improving Public Health Through Human Drugs*. Rockville, Maryland, 20857. www.fda.gov/cder/reports/rtn/2002/rtn2002.pdf. Accessed August 1, 2003.

[12] U.S. Government. Bureau of Labor Statistics. www.bls.gov. Accessed August 1, 2003.

[13] D'Agostino, R.B. et al. "Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation," *Journal of the American Medical Association*, 2001, Vol. 286, pp. 180–7.

[14] Bradford-Hill, A. "The Environment and Disease: Association or Causation?" *Proceedings of the Royal Society of Medicine*, 1965, Vol. 58, pp. 295–300.

[15] International Committee of Medical Journal Editors (ICMJE). Uniform Requirements for Manuscripts Submitted to Biomedical Journals. *Annals of Internal Medicine*, 1997, Vol. 126, pp. 36–47. www.icmje.org/. Accessed September 1, 2003.

AcademyHealth
Advancing Research, Policy and Practice

1801 K Street, NW, Suite 701-L
Washington, DC 20006
tel: 202.292.6700
fax: 202.292.6800
e-mail: SCI@academyhealth.org
web: www.statecoverage.net