Paper 5: Evaluating Disease Management Savings Outcomes[1]

Ian Duncan[2], FSA, FIA, FCIA, MAAA

Recently, because of the controversy over its results, many authors have discussed general methodological principles for ensuring validity in disease management ("DM") savings outcomes measurement. Examples may be found in Wilson and McDowell (2003), Wilson, et al (2004), Fetterolf, Wennberg and DeVries (2003), Linden (2004), and Fitzner (2004 (1) and 2004 (2)). Several of these references, while discussing measurement principles, also provide lists of different study types and designs. This paper will address the theory of measurement design and provide a practical evaluation of the most common designs for the practitioner.

Evaluating a Savings Calculation

The actuary may not always have the chance to design a measurement study, and will more frequently be called in to evaluate a vendor's or colleague's results. Three questions should be considered when evaluating results:

1. Has the measurement been performed according to a valid methodology?
2. How has that methodology been applied in practice? In other words, what assumptions, adjustments and calculation processes have been used to prepare the results?
3. Are the results arithmetically correct? Have data processing, arithmetic or calculation errors been made in the preparation of results?

This paper addresses the first issue, namely assessing the validity of the methodology. Other papers in this series provide insight into the second issue of practical application. Audits of actual calculations are, however, beyond our scope. With regard to the third point, calculations may be audited or a parallel test may be performed in which the results of the analysis are reproduced in order to confirm that results have been correctly prepared. We assume readers will be able to perform audits to validate the calculations, or, if necessary, a parallel test (although the latter is often highly resource-intensive).

Before we discuss evaluation of savings calculations, we define two terms used in this paper: Causality and Methodology.

---

[1] Paper 5 in a series of papers under the general title: "Evaluating the Results of Care Management Interventions: Comparative Analysis of Different Outcomes Measures" sponsored by the Society of Actuaries Health Section. Available at www.soa.org

[2] Lotter Actuarial Partners, New York, NY and Solucia Inc., Hartford, CT.

Causality

Causality is an important concept in both scientific and commercial studies of disease management outcomes. Just because savings are associated with a program does not necessarily mean that these savings are the result of the program. Attributing causality to an intervention program is a difficult problem, and one that has not been much studied in the field of DM outcomes. Research to date has been focused on attempts to obtain an accurate estimate of savings, no matter the source. Because of the difficulties inherent in proving causality, commercial purchasers of DM programs are usually satisfied with a weaker standard of proof: "demonstration" of savings, rather than proof of causality. Appendix 1 contains a more detailed discussion of causality for readers who are interested in more information about this issue.

What is a "Methodology"?

Methodology is a term frequently used but rarely defined in outcomes studies. The definition of a methodology is: "a body of methods, rules and postulates employed by a discipline**:** a particular procedure or set of procedures" (*Merriam-Webster Unabridged Dictionary Online*). What distinguishes a methodology from a calculation technique, however, is the fact that a methodology stands alone and can be implemented alone. A technique (such as an adjustment for age, or for trend) does not stand on its own but is rather an input to a methodology. Below, we compare the characteristics, including validity, of 10 different methodologies for calculating savings results.

Principles of Measurement Design: What Constitutes a Valid Methodology?

Evaluation of a methodology is a different problem than the evaluation of the results of an analysis. The former is a question of conformance to evaluation principles, while in the latter case we evaluate whether or not the author's hypothesis is rejected.

Whether designing an analysis from scratch, or evaluating a published study, the same principles determine whether a methodology is likely to be judged acceptable. The principles below are repeated from Paper 2 in this series[3]:

- *Reference Population:* Any outcome's measurement requires a reference population against which to evaluate the statistic of interest.

- *Equivalence:* To ensure validity in outcomes measurement, the reference population should be equivalent to the intervention population.

[3] Dove, H and Duncan, I: "Actuarial Issues in care Management Evaluations" Paper 2 in a series of papers under the general title: "Evaluating the Results of Care Management Interventions: Comparative Analysis of Different Outcomes Measures" sponsored by the Society of Actuaries Health Section. Available at www.soa.org

- *Consistent Statistics:* The comparison needs to measure the same statistic, the same way, in the reference and intervention populations[4].

- *Avoid irrelevant and potentially confounding data.* At its most extreme, this may imply measurement only of what the DM organization is paid to manage. The average DM program is usually only responsible for a limited subset of conditions, claims and patients, not the entire health plan population or necessarily all claims of the chronic population.

- *Control the exposure.* Assign each member to appropriate measurement categories for each month of exposure.

- *Reconcile the results.* The data going into an evaluation should be controlled (that is, reconciled to a valid or published source). The outcomes, too, should be reconciled to a valid source and should be plausible. An example often cited, implausible savings outcomes, consists of studies that show all or almost all of the cost incurred in an asthmatic population being reduced by a DM program.

In addition to the requirement for scientific rigor that is necessary for an academic paper, commercial purchasers of DM are likely to have additional requirements.

- The methodology must be one that a purchaser (or its consultant) is familiar with, or at least can grasp readily, and that should be perceived in the marketplace as sound;

- The methodology must be documented in sufficient detail for another practitioner to replicate the analysis, and, if required, allow the client to be able to replicate the savings estimates themselves (or at least major components of the calculation);

- The results of the application of the methodology must be consistent with the client's savings expectations and plausible overall;

- The application should lead to stable results over time and between clients, with differences between different studies and clients that can be explained; and

- The methodology must be practical, that is, it must be possible to implement it cost-effectively, without significant commitment of resources relative to the potential benefit being measured.

[4] For example, if the measured statistic is claims per member per month, this measurement should be performed the same way in both the baseline and intervention period, including identification of members, number of months of run-out, etc.

Study Designs for DM: A Summary

Many of the methodological differences in published studies that calculate savings are the result of the application of different methods of addressing population equivalence. As we survey methodologies, we find it useful to group those with similar characteristics together. So, for example, methods in the control group category have in common that they set up a control group; they differ in the way equivalence is achieved between the intervention and reference populations. Our groupings are differentiated by whether or not they incorporate an experimental control or reference group, or use primarily statistical methods for their conclusions. We believe that it is useful to identify similarities and differences this way; other evaluations of methodologies, for example those in DMAA's *Program Evaluation Guide,* simply list methodologies, leaving the reader to determine how each methodology differs from the others. But, as noted above, there is no single, agreed classification in the industry. Our view is that most major methodologies encountered in the literature or in practical commercial analyses may be mapped into the following three classes:

- Control group methods
- Population methods without control groups
- Statistical methods

Control Group Methods

Control group methods, that is, methods that attempt to match the intervention subjects with other subjects that are not part of the intervention group generally rate higher than other methods in terms of validity, scientific rigor and ability to replicate. The "matching" that takes place in these methods can be random (that is, subjects are selected randomly from the same population) or nonrandom. (We describe several nonrandom control groups below.) These methods also have high market acceptance, because it is simple to understand how the methods achieve equivalence. Except for random fluctuations, two large enough samples drawn from the same population will exhibit the same risk factors.
A control group may be:

- *Randomized* (comparing equivalent samples drawn randomly from the same population). It is important that randomization be performed prior to any interventions, if the results are to be generalized to the population from which the groups are drawn. Equivalence between the intervention and control groups is also not assured and should be demonstrated. This methodology is encountered more in academic than commercial studies, although the new Medicare Chronic Care Improvement Program requires randomized evaluation in large-scale implementations, so it may become more prevalent commercially.

  Table 1 shows a simple application of a randomized control design:

Table 1: Application of a Randomized Control Design

Claims Per Member Per Month

	Reference Population	Intervention Population
Baseline Period Cost	$90	$90
Analysis Period Cost	$105	$102

In this example, the two populations pre-intervention (baseline period) are equivalent, as demonstrated by the equal claims per member per month. In the analysis period, the intervention population experience is lower than that of the reference population, which indicates a positive effect of the program. (See Appendix 1, however, for a discussion of causality and interpretation of this result.)

- *Geographic* (comparing equivalent populations in two different locations). Unlike randomized controls, in which the control group is subject to the same forces as the intervention group, the risk profile and market forces present in different geographies may cause differences that obscure ("confound") the true difference in the intervention and reference populations. In many cases, these differences may be anticipated and corrected for in the analysis.[5] This adjustment is easier to make when there is no dynamic effect on the reference population over time. Consider the following example shown in Table 2.

Table 2: Application of Actuarial Adjustment to a Reference Population

Claims Per Member Per Month

	Reference Population	Intervention Population
Baseline Period Cost	$90	$100
Analysis Period Cost	$105	$102

In this example, geographic differences drive differences in the intervention population costs per member per month and the reference population costs in the baseline period. Initially there appears to be no savings: costs increased by $2 pmpm between the baseline and intervention periods. However, this result is due to the confounding effect of healthcare cost trend. Comparison with the reference population experience shows that trend at the rate of 16.7 percent was present in

[5] See, for example, "Actuarial Adjustment" in "Dictionary of DM Terminology" edited by Ian Duncan (DMAA, 2004).

the reference population. The obvious adjustment to the intervention population data is to multiply the intervention period cost by $100/$90, or $105 *1.11 = $116.67. Then the estimated savings from the intervention would be: $116.67 - $102.00 or $14.67.

This savings estimate may, however, be subject to forces that impact the two populations in the intervention period differently (for example, benefit design changes or changes in provider contracts), something that should be further explored before the results are accepted.

- *Temporal* (also known as the Adjusted Historical Control Design). This design compares equivalent samples drawn from the same population but at different points in time, specifically before and after the intervention program. This is the most common approach used in the disease management industry. To account for the fact that, over time, utilization and cost of care within a population is empirically observed to rise, a medical trend adjustment is used to adjust the historical experience to the same time period as the intervention data.

- *The Product Control Methodology* compares samples drawn from the same population at the same point in time, but differentiates between members who have different products, such as HMO vs. PPO, or indemnity vs. ASO. Clearly the product differences introduce the potential for the confounding effect of product selection, different medical management, included benefits or providers (often a factor with ASO groups) and reimbursement, and this approach should be treated with caution. The mathematics of this methodology are similar to those of the geographic methodology (see above).

- *"Patient as Their Own Control" (Pre-post Cohort Methodology).* This method differs from the "temporal" method described above, in which the intervention and comparison populations are resampled in each period to ensure equivalence. Applying the same rules of identification to create an equivalent population in a different time period is somewhat analogous to the "with replacement" and "without replacement" problems with which actuaries are familiar from introductory statistics courses. In the "Patient as Their Own Control" method, the comparison group is the population as initially defined, but measured post-intervention. In this design, there is no equivalent reference population. One conclusion from our discussion of regression to the mean (see Paper 2 in this series) is that regression to the mean is potentially present in the post-intervention population. If the extent of the regression in the baseline population were known, an adjustment could be applied to the intervention population to correct for the inherent regression. However, in most cases, the extent of regression is not known. This phenomenon is illustrated in Table 3:

Table 3: Application of a Pre- Post Study (Patient as Their Own Control)

Claims Per Member Per Month

	Intervention Population
Baseline Period Cost	$100
Analysis Period Cost	$95

In this example, the initial impression is that the program has reduced the cost within the population by $5 ($100 - $95). However, the following illustrates the population in the absence of an intervention program. The reduction observed in cost from the baseline to the analysis period is due to the effect of regression to the mean (see Paper 2 in this series).

Claims Per Member Per Month

	Intervention Population
Baseline Period Cost	$100
Analysis Period Cost	$98

In this case, regression to the mean inherent in the intervention population accounts for a reduction of $2 in cost, which needs to be subtracted from the observed reduction of $5. In most cases, unless a randomized control study is performed, it is not possible to estimate the value of the regression effect within the intervention population.

- *Participant vs. Nonparticipant Studies.* In this method, the experience of those who voluntarily elect to participate in a program is compared with the experience of those who choose not to participate. The participants represent a group with potentially different risk factors than those of the nonparticipants (we already know that they differ with respect to the important factor of willingness to take control of their own health by engaging in a program). Some authors appear to believe that it is possible to adjust the reference population (nonparticipants) to bring them into equivalence with the intervention population; however, it does not appear to us that the effect of selection can be estimated. In any event, the existence of selection bias is known in the industry and this methodology is not assigned a high credibility, with or without adjustment.

- *Other types* of control group methods are cited in the literature, for example a "staggered roll-out[6]." However, these methods can be deconstructed to fit one or more of the above five categories.

Within control group methods, randomized control methods and adjusted historical methods both exhibit a high degree of validity. Other types of control methodologies also achieve high market acceptance because of their intuitive appeal (even though the technical aspects of achieving equivalence in nonrandom controls may be daunting). Methods using other controls (geographic or product) are less practical to implement and may require a highly sophisticated system of risk classification and risk adjustment to ensure equivalence between the intervention group and reference group. The "patient as their own control" or pre-post cohort methodology, as discussed elsewhere in these papers, while it is well-known and understood, suffers from potential bias due to regression to the mean. Results produced using this method cannot be considered valid (something that is increasingly being recognized in the market). A similar conclusion is drawn for participant vs. nonparticipant studies, which, while they may be simple to understand and implement, suffer from a fundamental flaw in that they compare a self-selected population with its complement (and therefore fail to demonstrate equivalence).

Non-Control Group Methods

- Among non-control group methods, *Services Avoided* methods are commonly used, particularly for case management applications (see Paper 3 in this series). In this methodology, the intended resource utilization of the member prior to the intervention is estimated because the member calls a health plan to pre-authorize a particular service. Savings are then calculated by estimating the cost of the requested service and comparing this estimate of utilization with actual utilization after the intervention. In the specific example of case management, an estimate of the likely resource utilization of the member is compared with actual approved utilization (including any alternative services arranged or approved by the case manager), and the difference is counted as savings due to the case management program. Some applications track the utilization of members who report a change in intent (for example, the intent to have surgery) for as long as six to 12 months post-intervention to ensure that the change in intent was not later reversed. Because of its widespread use, the methodology scores high on familiarity but lacks a reference population. The method also includes a high degree of subjectivity, both in selection of candidates and in estimating what utilization would have been, absent intervention. For these reasons, the validity of savings calculated by this method is questionable.

- *Clinical Improvement* methods have achieved reasonable market acceptability. In a clinical improvement method, the change in an objective clinical measure is first

[6] This methodology is used in a paper on Diabetes outcomes: Villagra, V. and Ahmed, T. "Effectiveness of a Disease Management Program for Patients with Diabetes." *Health Affairs* July/August 2004; (23) 4: 255-266.

observed (for example, the rate of use of a particular medication by members who have a diagnosis for which the medication is indicated). The peer-reviewed clinical literature is searched for studies that indicate how health is improved (and resource utilization is decreased) in the population with the particular diagnosis, as a result of adherence to treatment. A dollar value is assigned to the reduction in resource utilization, which is then applied to the observed change in the clinical measure due to the program. This method appeals to some evaluators because it involves objective causal factors. Unlike some other methods that measure changes in claims, this method can point to actual improvement in a clinical factor that can cause reduction in claims. Despite its appeal, the methodology rates relatively low because the results are achieved through a subjective process and often lack a reference group—in this case, the subjective element is the estimation of financial savings by inference from published clinical studies (somewhat akin to a benchmark method). To our knowledge, no study has ever been published that compares estimates of savings in a population using a clinical improvement approach with estimates in the same population using a randomized or other control group approach. We should also remember that our review of the literature does not yet show a strong, demonstrated, link from clinical to financial effects[7].

Statistical Methods

We term these methods "statistical methods" because they involve the use of purely statistical techniques (for example, regression or benchmarks), rather than the construction of an explicit reference population. The term "statistical methods", however, should not be confused with the statistical tests that underlie hypothesis testing. Statistical tests of hypotheses should be applied to any study to determine whether the results are significant.

- *Time-Series Methods.* The objective of these methods is to fit a curve or series of curves to the data over time, and then to demonstrate a divergence from the best-fit line once the intervention is applied. This method is a generalization of the trend-adjusted historical control methodology, which focuses on just one historical period. The difficulty of fitting a curve to healthcare data over time appears to us to be almost insuperable, because of the need to capture in the model the effect of a multitude of factors, both endogenous and exogenous. Because of the difficulty in demonstrating a high correlation between the actual data and the fitted line, demonstrating divergence from that line and assigning causality to the DM program is complicated.
- *Regression Discontinuity*[8]. This design may be thought of as a special case of the Time Series method (above). At its core, this method looks for a statistically

[7] See Dove, H and Duncan, I, "Selective Literature Review of Care Management Interventions". Paper 3 in a series of papers under the general title: "Evaluating the Results of Care Management Interventions: Comparative Analysis of Different Outcomes Measures" sponsored by the Society of Actuaries Health Section. Available at www.soa.org.

significant difference between two similar subsets of the population. A regression line is first fitted to data that relates pre- and post-intervention experience. A dummy variable is included in the regression to capture the difference between the intercept of the intervention population's regression line at the "cut-off point" and that of the reference population.

To understand the regression discontinuity method, consider Figure 1. In this example, we plot the relationship between an individual's risk score in a baseline period (Year 1) and a cost per member per month in the follow-up period (Year 2). Each point in the scatter represents the pair of observations for a single member. The method requires an objective method for segregating those members eligible for, and those not eligible for, the intervention. The risk score is a useful variable because frequently intervention programs are targeted at members whose risk score exceeds some predetermined minimum. (Risk score is the preferred measure for the Year 1 variable over cost because, while the group targeted for intervention often includes high-cost members, this is seldom the sole criterion for targeting.) An upward-sloping regression line implies that members with a high Year 1 risk score tend to have high Year 2 costs, as well. The closer this relationship, the closer the data points will be to the line. On the other hand, a line that slopes upward at less than 45° indicates regression to the mean (high-cost Year 1 members tend to have lower Year 2 costs; low-cost Year 1 members tend to have higher Year 2 costs). While this may not have any direct impact on measurement of outcomes, there is a deeper significance to the slope of this line when it is not 45°, because of the implications of this slope for predictive modeling and selecting intervention subjects.

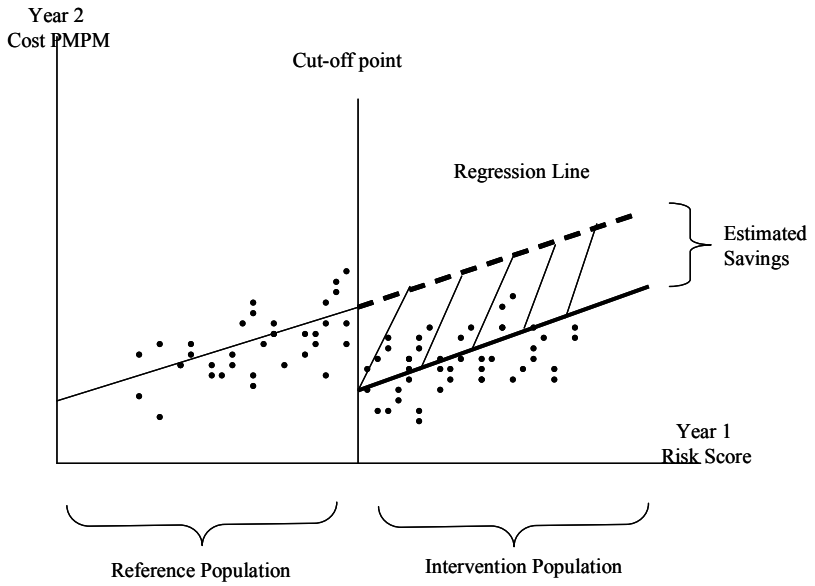
A regression is fitted to the Year 1 vs. Year 2 data. An example of this regression is: $Y_i = ß_0 + ß_1 X_i + ß_2 Z_i + \varepsilon_i$

Where:
Y_i = Dependent variable (Year 2 cost for the i-th person)
$ß_0$ = Regression intercept
X_i = Independent variable (Year 1 cost for the i-th person)
$ß_1$ = Regression coefficient for variable 1
Z = Dummy variable with value zero (if observation is in the reference population) or 1 (if in the intervention population)
$ß_2$ = Regression coefficient for dummy variable Z
ε_i = Random error term for the i-th person

[8] This section is adapted from publications of William M. K. Trochim ("Research Design for Program Evaluation" by W. M. K. Trochim. Sage Publications, 1984) and Linden A, J.L. Adams and N. Roberts: "Evaluating Disease Management programme effectiveness: an introduction to the regression discontinuity design". Journal of Evaluation in Clinical Practice, 0 (0) 2005. I want to thank Dr. Ariel Linden for his helpful discussion on this section.

In some applications the independent variable, X, is transformed so that the cutoff point intersects the X-axis at a value of Zero. However, this does not appear to be essential to the successful application or understanding of the method.

Figure 1 Relationship Between Costs in Two Periods



A significant value for the dummy variable regression coefficient ($ß_2$) implies that there is a statistical difference between the intercept of the basic line (reference population line) and the intervention population line at the cutoff point. The value of this coefficient gives an estimate of the effect of the program at that point. The focus on the cutoff point may seem excessive, but an important feature of this method is that the effect is calculated at a point (the cutoff point) at which the reference and intervention populations are most similar. This overcomes a potential objection that there is not a reasonable "goodness-of-fit" throughout the entire population, particularly at the extremes of the distribution. Second, while this method may have been used to demonstrate a significant effect, clients who are purchasers of care management interventions require an estimate of the savings due to the program, and we are not aware of an actual savings calculation using this method. Savings could, however, be calculated by projecting the expected cost of the intervention population using the regression analysis and subtracting the intervention population's actual expenses (as in the hatched area above). We have described this method at greater length than some others because there is a considerable interest in its potential application in commercial calculations, and we expect to see the method used more in the future.

- *Benchmark Methods:* Certain key statistics in the population under management are compared with the value(s) of the same statistics in another (benchmark) population. Benchmark studies compare outcomes for a managed population with

an independent statistic—either a national, regional or other external benchmark, or a metric available from an external source. It is difficult to demonstrate adequate equivalence between the intervention population and benchmark population in these studies. The principle of equivalence requires consistency between the populations on a very large number of risk factors. While it may be possible to apply the principle of equivalence in theory, it is unlikely that a published study or benchmark source will provide sufficient detail needed to apply adjustments in such a way that equivalence can be assured. Actuaries, who are used to making adjustments to population data and inferring from one set of data conclusions in another, (for example in rating and underwriting applications) will be familiar with the issues that exist in using "external" data sources.

Statistical methodologies are restricted to the scientific community; we have not seen them used in widespread commercial application. The more commonly used control group and non-control group methods are simpler to understand and the calculations are transparent. Statistical methods involve techniques with which most business users are not familiar and may regard with a degree of suspicion (for their "black box" aspect). For example, the regression-Discontinuity Method is often discussed favorably in the literature, but we have yet to find significant analyses using this method. Time series methods have one important advantage in that they draw attention to long-term utilization and cost trends in the population, which provides the evaluator with valuable information about what was happening in the population before the intervention program began. We question the practical usefulness of this method in a health plan environment where so many variables change over time, making it virtually impossible to control for confounding. Some authors favor benchmark methods (and they have some appeal to actuaries, who are used to making the type of adjustments required to compare different populations). However, the sheer number of variables and risk factors (and lack of information about their values) that could potentially affect a benchmark analysis will make this another difficult methodology to apply in practice. Statistical methodologies may yet prove useful, but none are developed to the point of being practical for implementation in a commercial environment.

Some authors suggest propensity scoring as a methodology, equivalent to others considered above. We recognize the importance of propensity scoring, a method of identifying or creating populations of the same degree of risk as the intervention population. (Propensity scoring is potentially of importance to actuaries because of its similarity to risk adjustment. See Appendix 2.) However, because we do not consider propensity scoring a methodology, but rather a technique for adjusting other populations or creating matched populations, we do not include it in our comparison. For more information on this technique, see (Linden A, Adams J, Roberts N. "Using propensity scores to construct comparable control groups for disease management program evaluation". *Disease Management and Health Outcomes*. 2005;13(2):107-127).

Comparative Assessment of Methodologies

The table on pages 14 and 15 summarizes our conclusions about how different measurement methodologies meet our key criteria (above) that determine whether a methodology is valid; inherent validity (lack of obvious bias) and scientific rigor; familiarity (how commonly used is the methodology in the industry?); market acceptance (how is the method perceived in the marketplace?); ease of replication and auditability; application (how is the methodology applied in practice?); and other important issues in the application of each methodology. These criteria for assessing methodologies are our own and reflect our experience as consulting actuaries in this area. Other actuaries, or practitioners from other disciplines, may have different criteria by which to judge methodologies. The point, however, is that methodologies are not equally valid, and results that are prepared according to a higher-scoring methodology should be given more weight.

TABLE 1: COMPARISON OF CERTAIN COMMONLY USED DM SAVINGS CALCULATION METHODOLOGIES

	Method Type	Method	Application	Validity/ Scientific Rigor	Famil-iarity	Replicability/ Auditability	Evaluation of Methodology	Other issues
1	Control Group Methods	Randomized control	Requires randomized, control group not subject to intervention. Metric in the Intervention group is compared with the same metric in the control group, and the difference is assigned to the effect of the intervention.	High	High	Difficult to replicate and audit; need another randomized group.	"Gold Standard" method, although requires demonstration of equivalence. Need for incurred claims results in delays in evaluations.	Practical to implement and avoids adjustment issues, although requires sufficient number of members. Viewed by health plans as difficult to implement and potentially unethical. Randomization must occur at the population level if results are to be applied to the population.
2		Temporal (Historical) control	Requires population drawn according to identical rules from two periods. Metric from the Intervention period is compared with the same metric from the Baseline period, adjusted with trend. Requires adjustment of the comparison population to be equivalent to the Intervention population.	High	High	Replicable and auditable	Becoming the most widespread methodology in the industry. Need for incurred claims results in delays in evaluations.	Implicit assumption that regression to the mean is uniformly distributed in the Baseline and Intervention periods, and that a robust trend estimate is available. Differs from the Pre-post cohort (Patient as own control) method because a new cohort is used for comparison, including all members that meet the identification criteria in the period.
3		Geographic or product line controls	Requires population drawn according to identical rules from two different groups (e.g. geographies). Metric from the Intervention period is compared with the same metric from the control, adjusted for all appropriate risk-factor differences.	High/Medium	High/ Moderate	Replicable and auditable	Not widely used.	Sometimes difficult to adjust for the many risk factors that affect a population and its utilization (see Paper 2).
4		"Patient as their own control" (Pre-post cohort)	Patients are identified pre-intervention and then followed post-intervention. Pre-intervention metric is compared with post-intervention metric.	Low	High	Replicable and auditable	Widely used, but regression to the mean issues are causing purchasers to re-evaluate (see Paper 2).	Theoretically possible to correct for the effect of regression, but no method has yet been developed to do so. Differs from the Temporal (historical control) method because the same cohort is used for comparison, and newly identified members are not added.
5		Participant vs. Non-participant	Patients are invited to enroll in a program. Those who choose to enroll are subject to treatment; those who choose not to enroll form the control group.	Low	High	Replicable and auditable	Widely used, but selection bias causes this methodology to be highly suspect.	Theoretically possible to correct for the effect of selection bias, the effect of a member's "willingness to change" is unmeasurable.

TABLE 1 (contd): COMPARISON OF CERTAIN COMMONLY USED DM SAVINGS CALCULATION METHODOLOGIES

	Method Type	Method	Application	Validity/ Scientific Rigor	Famil-iarity	Replicability/ Auditability	Evaluation of Methodology	Other issues
6	Non-Control Group Methods	Services Avoided (also called pre-Intent/post-Intent)	Record intent of different patients, track for a period of time to determine actual outcome, and assign a dollar value to the avoided event (adjusted for alternative treatment, if any).	Moderate	High	May be difficult to replicate; auditable.	Frequently used for small, highly-specialized programs (such as case management).	Two issues: participant bias (participants who are more likely to change their minds seek information and support) and evaluation and recording of intent is subjective.
7		Clinical improvement methods	Measure clinical improvement and estimate financial savings using a model based on the difference in cost of well-managed and other patients.	Moderate	Moderate	Difficult to replicate; difficult to assemble comparable clinical trial data.	Useful for small volume studies and when a result is required more quickly than data-based evaluations.	Requires review of the significant literature on clinical improvement, and a method for projecting financial from clinical improvement. To our knowledge there is no comparative study of results of clinical improvement and other methods.
8	Statistical Methods	Regression-discontinuity	A regression line is fitted on the relationship between Year 1 Risk Score and Year 2 PMPM costs in a population; a dummy variable is included to indicate membership in the intervention group. The difference at the "cut-off point" between the non-intervention and intervention population regression lines indicates that the intervention has had an effect.	Unknown	Low	Replicable and auditable	Highly-regarded as a theoretical method in the scientific literature, but we are not aware of a specific practical DM application.	To be determined.
9		Time-series	Extension of the Adjusted historical control methodology to multiple periods.	Low	Low	Replicable and auditable	Not widely used in commercial evaluations.	The effect of changes in risk-factors (often reflected in variations in Trend) is compounded over a period of years, making it very difficult to control this calculation.
10		Benchmark	Metric in the intervention group is compared with the same metric in another population. The difference is assigned to the effect of the interventnion and savings are estimated accordingly.	Low	Low	Replicable; difficult to assemble valid comparison data	Occasionally encountered in commercial applications.	Comparison populations are unlikely to be described in sufficient detail to determine their degree of comparability (or the extent to which adjustment is required).

Conclusion

As the DMAA statement referenced at the start of this paper implies, a non-control group methodology is unlikely to be a satisfactory method for calculating DM savings results, except under unusual circumstances. The DMAA conclusion is that the preferred method for any evaluation is a randomized control study (in our experience, easier to implement than is commonly believed).

As our evaluation above shows, a nonrandomized control group study can also be valid (provided equivalence is maintained and can be satisfactorily demonstrated). This nonrandomized control group could be temporal, geographic or product-based, but not based on self-selected members (such as nonparticipants). The achievement and demonstration of equivalence is an area in which actuaries may contribute. Actuaries are qualified, through their experience in rating and underwriting of health risks, to perform the process of drawing conclusions and making projections in one population from data or experience of another population. In the next paper in the series, two issues will be addressed: how to set up an analysis to maintain control of the data and the achievement and demonstration of equivalence.

In addition to the issues of equivalence that are important in choosing a valid methodology, the application of the methodology needs to be carefully controlled, or the results of the measurement will be invalidated for other reasons. Paper 6 also explores the issue of the controls that should be in place in applying an important methodology, the adjusted historical control methodology.

APPENDIX I: CAUSALITY

Any discussion of outcomes measurement needs to consider the role of causality and the standard of proof required in an *actuarial* or financial analysis or savings calculation.

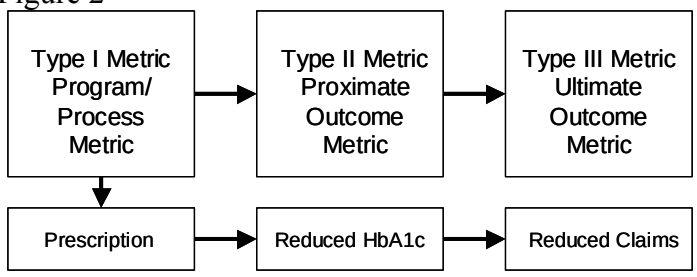
Papers cited in the references approach outcomes measurement from the perspective of multiple users, both scientific and business. The scientific emphasis tends to promote a higher degree of proof than is likely to be encountered in a business setting. This higher standard in the scientific community results in a focus on a need for proof that the DM intervention "caused" the specific outcome, or "causality". Business users, while demanding considerable rigor in other aspects of an evaluation, (such as validation and reconciliation of source data) may be satisfied with a "demonstration" standard, where "association" between cause and effect may be sufficient, rather than the stricter test of "causation". Demonstration may be satisfied with an analysis that shows association between the intervention program and a favorable outcome, together with adequate demonstration that the results are not biased or confounded by factors that could impact the result.

Proof of causation requires that the mechanism whereby the outcome is achieved be unambiguously demonstrated. For example, if the result to be proved is savings, then a study that proves causation would have to establish "missing" components in the target population (such as compliance with best-practice medical care), then show how the intervention improved compliance with care in the population, and finally, the resulting financial outcomes.

The DMAA committee that drafted "Principles for Assessing Disease Management Outcomes" recommends that "absent independently funded research trials using a randomized control design, … the preferred evaluation method for assessing causality of DM in achieving outcomes is comparison to an equivalent control group. This assessment should include a baseline measure in the intervention and control populations with a remeasure of the outcomes of interest following the intervention in the DM population versus the control group."

Wilson and MacDowell refer to a causal pathway that can be used where "proof" is required, in a pure scientific sense. They describe a pathway as follows in Figure 2, in which we provide an example of different types of metrics:[9]

Figure 2



[9] From Wilson and MacDowell; reproduced by permission.

A "Type I" metric is the most basic measure. These basic metrics measure the components of a process, such as units of input, rather than the results of a process, such as units of output. For example, a Type I metric in a DM program could be a volume measure such as the overall number of members of a health plan, the number of chronic members, the number of enrolled chronic members, the number of nurses assigned to manage patients, the number of calls attempted to chronic members over a period of time, or the number of prescriptions for a particular medication consumed by the population.

A "Type II" metric is an intermediate measure, and it represents the process outcome of the input measured in the Type I metric. As such, it may measure the result of an input that contributes to the final output measure of interest. So if our ultimate measure of interest is reduced claims, an intermediate measure could be the rate at which patients refill their prescriptions, or the proportion of a diabetic population whose hemoglobin A1c score is less than 7.0 (the maximum level set by the American Diabetes Association as indicating blood sugar "control" in a diabetic).

Finally, a "Type III" metric is the outcome metric of interest that the program has been designed to achieve, for example per member per month incurred claims in the chronic population.

Clearly, establishment of causality implies that it should be possible to trace the ultimate result back to its source (the inputs and improved outcomes that contribute to an improved financial measure). Conversely (and importantly in disease management financial outcomes measurement) it should be possible to demonstrate the process and clinical causal factors that lead to the clinical improvements that lead to the measured financial outcomes.

The ultimate outcome metric of interest to actuaries and other financial buyers is financial savings to the payer. Assuming that actual savings results are available, the steps to understand causality include:

- The mechanism that produced the savings results, i.e., the "causal pathway". For example, first-level proximate metrics would include enrollment of chronic members. Higher-level (Type II) metrics would include an increase in testing and further stratification of high-risk members, increased medication adherence, increased compliance with physician-ordered treatments or improved test-scores on clinical tests. Finally, the ultimate metric will be that metric of interest to the program sponsor—for example, reduced hospital admissions.

- The influence of treatment variations. For example, is there evidence that varying outreach efforts results in different enrollment experience (implying that enrollment is in fact controllable by the program, rather than a constant)?

- Consistency of the proximate outcome metrics with the ultimate outcome metrics (e.g., do we observe higher compliance or testing rates at the proximate level, and is this experience sufficient to lead to the observed ultimate outcome metric)? An observation of improved financial results, accompanied by deterioration in proximate measures such as test scores, would tend to suggest that the improved financial results are aberrant. Conversely, as we noted in Paper 4, "Understanding the Economics of Disease Management" there remains an unanswered question in disease management about how we reconcile improved clinical measures (proximate measures) with financial measures that do not show the same degree of improvement.

Although "demonstration" is a weaker standard than a scientific test of causality, if it is performed rigorously, it should still be sufficient for most financial buyers. A demonstration standard simply measures the comparative values in the intervention and reference populations and assigns causality to the intervention.

APPENDIX 2: PROPENSITY SCORES AND RISK ADJUSTMENT

Propensity score methods are not a methodology on a par with other methodologies discussed above, but represent a way of identifying or scoring different members for evaluation using another methodology. Because of the interest that propensity score methods have raised in the industry, and their similarity to the risk adjustment methodologies with which actuaries are familiar, we cover these methods briefly in this section.

In order to grasp the basics of propensity scoring, consider a randomized control method. In this method, every member is equally likely to be selected for assignment to the intervention or reference population, irrespective of the member's characteristics. In a randomized control method, the outcomes from the intervention and reference groups are compared, and if savings are the statistic of interest, this is estimated from the difference in mean costs of the two groups. In a propensity scoring method, an intervention group is first identified (or may self-identify, as, for example, program participants). A matching reference population is then identified from a population of nonparticipants, for example, those who elect not to enroll, or members of a similar population not offered the intervention. Savings are estimated as the difference between the mean costs of the intervention and reference groups. The problem in such a method, of course, is identifying members with exactly the same characteristics as the intervention population. Consideration of the potential risk variables (age, sex, condition, co-morbidities, residence, plan of benefits, etc.) quickly indicates that the likelihood of finding an intervention group member that matches exactly on every one of the risk characteristics is small. The propensity score method was developed to reduce the effect of the multiple risk factors to a scalar variable, or single score, using multiple regression. Thus, for example, a reference group member may have different age, sex, co-morbidities, etc. to those of an intervention group member, yet have the same score, allowing the two members to be "matched". The similarity of the score that this method produces to the score produced by a "risk adjustment" methodology is obvious.

Some authors believe that this method holds considerable promise for DM evaluations because it allows participants to be "matched" with nonparticipants in a pre-post analysis, yet apparently overcome the two major shortcomings of this method: selection bias and regression to the mean.

We believe that this method has a place within other methods, but we are less optimistic about its potential than other authors. While it may be true that a propensity score may be calculated for observable (and measurable) variables, the method cannot accommodate nonobservable or nonmeasurable variables. Important examples of the latter are the member's "readiness to change," which influences the member's likelihood to enroll in a program. As we have shown elsewhere,[10] members who enroll in programs do not experience the same claims experience as those who do not enroll in programs. Examples of otherwise nonmeasurable

[10] See discussion about selection bias in Dove and Duncan, *op cit.*

variables that affect risk include the quality of care provided by the treating physician, the type of care and support that the member receives in the home environment, etc.

This method continues to be of considerable interest to the industry and we are likely to see continued research into its application in the future.

References

Fetterolf, D, D. Wennberg, and A. DeVries. "Estimating Return on Investment in Disease Management Programs Using the Pre-Post Analysis." *Disease Management* 7(1) Winter 2004.

Fitzner, Karen, et al: *Disease Management Program Evaluation Guide.* DMAA, 2004.

Fitzner, Karen, et al: "Principles for Assessing Disease Management Outcomes." *Disease Management* 7(3) October 2004.

Linden, A., J.L. Adams, and N. Roberts "An Assessment of the Total Population Approach for Evaluating Disease Management Program Effectiveness." *Disease Management* 6(2) Summer 2003.

Linden, A., J.L. Adams, and N. Roberts "Evaluating Disease Management Programme Effectiveness: An Introduction to the Regression Discontinuity Design." *Journal of Evaluation in Clinical Practice* 0(0) July 2005.

Linden A, Adams J, Roberts N. "Using propensity scores to construct comparable control groups for disease management program evaluation." *Disease Management and Health Outcomes* 13(2) 2005.

Trochin, William M. K., *Research Design for Program Evaluation*. Beverly Hills, CA: Sage Publications 1984.

Wilson, T.W., J. Gruen, W. Thar, D. Fetterolf, M. Patel, R. Popiel, A. Lewis, and D.B. Nash, "Assessing Return on Investment of Defined-Population Disease Management Interventions." *Joint Commission Journal on Quality and Safety*. 30(11) November 2004.

Wilson, T.W. and M. MacDowell, "Framework for assessing causality in Disease Management Programs." *Disease Management.* Fall 2003.