



---

# **FRAMEWORK FOR ASSESSING CAUSALITY IN DISEASE MANAGEMENT PROGRAMS**

---

Martin MacDowell, DrPH, MBA  
Thomas Wilson, PhD, MPH, DrPH

Commissioned by the Disease Management Association  
of America Quality and Research Committee

**DMAA**

Disease Management Association of America



# FRAMEWORK FOR ASSESSING CAUSALITY IN DISEASE MANAGEMENT PROGRAMS

Martin MacDowell, DrPH, MBA <sup>a</sup>

Thomas Wilson, PhD, MPH, DrPH <sup>b</sup>

<sup>a</sup>University of Illinois at Rockford, Rockford, Illinois

<sup>b</sup>Wilson Research, LLC, Loveland, Ohio.

## **Acknowledgements**

The authors would like to acknowledge the members of the Quality and Research Committee of the Disease Management Association of America for helpful comments, especially their Chair, Rose Maljanian, RN, MBA. Also Kathleen Brody, BSN, PHN of Kaiser Permanente, Center for Health Research, Portland Oregon for her clinical review of the paper and assistance in developing the congestive heart failure case study. Finally, to the fall 2002 Xavier University students in the descriptive epidemiology class for improving the authors' abilities to better communicate the practical concepts of epidemiology to non-epidemiologists.

Research for this paper was partially funded by an unrestricted educational grant from the Disease Management Association of America.



Commissioned by the Disease Management Association  
of America Quality and Research Committee



# Executive Summary

The paper, part of a series of papers commissioned by the DMAA, is designed to help answer the question about value of a disease management intervention. Its purpose is to provide recommendations regarding strategies for assessing causality of disease management (DM) programs. The recommendations are as follows:

- 1) The DM program must explicitly state the intervention pathway. This would include three metrics used to measure the actual intervention (“cause”) as well as the proximate and ultimate outcomes that are hypothesized to result from that intervention (“effect”).
- 2) The DM program metrics must be compared to metrics from an independent source, broadly defined as a “reference population.” At a minimum, this referent metric should be based on benchmark values from commonly accepted practice guidelines.
- 3) Statements that the DM program “caused” specific outcomes (e.g. economic, health, or satisfaction) must be based on investigations that have adopted study designs more rigorous than that of a post-only study design. The evaluators of the DM program should acknowledge other factors that may have impacted the outcome (so called “confounding factors) and, when possible, “control” for them. In addition, write-ups should acknowledge the potential strengths and weaknesses of their approach.
- 4) It is unlikely that one study will provide definitive proof of the value of a specific DM program, thus, we recommend that studies be on-going, at multiple points in time, and at multiple sites. Moreover, if feasible, more than one kind of study design should be used.
- 5) Finally, all studies purporting to show the value of DM interventions should be capable of passing expert review such as through a submission to a peer-reviewed journal or an outcomes validation program.

The paper goes into some detail about methods available to conduct a credible analysis of DM programs. It begins with a discussion of importance of identifying the criteria by which the patient population is selected. This includes, among other issues, the need to explicitly state the differences, if any, between those individuals who subsequently participate compared to those who do not, both at baseline and follow-up periods.

A second area of discussion revolves around the need to clearly state the path the program takes from the DM intervention, to the intermediate outcome, to the ultimate outcome. How exactly does a phone call, for example, lead to a doctor's visit, and subsequently, to lower costs? The paper discusses three different categories of metrics: Type I, II, and III. Type I metrics refer to a measurement of the actual intervention initiated by the DM program. Type II metrics are intermediate or proximate factors that are directly caused by the DM program (e.g., higher level of adherence to medication orders) and, in turn, lead to a greater probability of the improvement in an "ultimate" or Type III metric (e.g., fewer inpatient admissions, lower costs). These categories of metrics should make up the intervention pathway and the intervention pathway should be explicitly stated.

The remainder of the paper deals with methods of dealing with the very significant issues of taking into account factors that could be related to the outcome in the absence of the DM program. These include a discussion of mathematical adjustments to "control for" include "confounding factors" such as economic inflation. Additional areas of discussion include the need to use the appropriate "test" to examine the "statistical significance" of age differences, for example, between participants and non-participants in a DM program.

Most importantly, since all "confounding variables" cannot be measured, let alone anticipated, sophisticated study designs that include a reference population are discussed. This reference population is used to better understand the trends that DM program participants would likely have had in the absence of a DM program. These designs differ in the methods for selecting a reference population that is "equivalent" to the DM population in all aspects, ideally, except the DM program. Seventeen different designs are introduced, including two "Post Only" designs, three "Benchmark" Designs, three "Quasi-Experimental" designs, one "Ecological" design, two "cross-sectional" designs, one "Case-Control" design" and five "Follow-up" designs. Four of the five follow-up designs are variations of randomized trials (either field level or individual level). The authors state that all except the two "post only" designs can be very useful to better understand the effectiveness of a DM program.

The establishment of any degree of methodological sophistication will greatly enhance the reputation of the important and growing efforts in disease management. This paper is far from definitive; it is only an introduction to some of the issues involved in assessing causality and value of a DM program. We strongly urge the readers to consult experts in epidemiology and statistics as well as textbooks and articles on the methods discussed here.

## (A) INTRODUCTION

This paper is an introduction to the principles and methods of determining the impact – health (clinical and humanistic), economic, and satisfaction – of disease management programs. It is designed to be a practical and useful introduction to assist in conducting a credible evaluation of disease management programs and assessing the integrity of evaluation done by others. “Outcomes” are not enough; hypotheses that outcomes are “caused” by the DM program must be articulated and tested.

**“OUTCOMES” ARE NOT ENOUGH; HYPOTHESES THAT OUTCOMES ARE “CAUSED” BY THE DM PROGRAM MUST BE ARTICULATED AND TESTED.**

Disease management (DM) programs are facing scrutiny from increasingly well informed and analytically prepared customer groups including managed care organizations, self-insured employers or government organizations. All want a credible answer to the ultimate question: “How well did the DM program work?” Several issues need to be addressed to enable us to answer this question. First are the metrics themselves; there have been and continue to be numerous efforts to standardize clinical process and outcome metrics. The pioneering work of NCQA’s HEDIS metrics applied across numerous conditions is a case in point, measuring the right things in the right ways are essential prerequisites for assessing the effectiveness of any health or disease management program.

The second issue is related to the methods used to determine the exact “effect” the DM operations have on clinical and financial outcome metrics. If we really knew that a disease management program had an effect of some magnitude on hospitalization rates and some known effect on emergency department (ED) visits, then, by knowing the cost of a typical asthma hospitalization or ED visit one could easily calculate the exact or specific assessment of the financial impact of the program. Thus, a rather precise determination of financial outcome, for example,

return on investment (ROI), could be determined.

But, unless we study these programs with rigor, we really do not know exactly what “effect” a DM intervention truly has on a population of enrollees. Certainly, many patients improve, but to what extent did the DM intervention “cause” that improvement? After all, other external factors could have been partially responsible. These uncertainties about cause-effect are well known in the medical research field and extraordinary thought and efforts have been put into designing methods and tools to increase the probability that decisions regarding resource allocation – clinical and financial – are as accurate and ethical as possible.

The credible assessment of the effect of DM on clinical, financial, and other outcomes is probably the most important issue facing DM today. This document seeks to acknowledge the reality of doing program evaluation, i.e. causality assessment, in a real world setting. We acknowledge that in many cases DM programs are multidimensional and can bring perceived value to a managed care organization, employer, providers, and patients even in the absence of a study that shows cause and effect. This paper, however, is only focused on one dimension, that is, on some of the methods available to measure and assess a direct relationship between a DM program and some desired outcome.

The central organizing theme is that the evaluation of disease management (DM) impact — the determination of causality – is impossible without the explicit use of a population that can be compared to the DM population. As will be discussed in the paper, this “reference population” can appear in numerous forms. However, in whatever form it appears its purpose is the same— to provide an answer to this key question:

*What would the outcome have been in the DM population in the absence of DM?*

The metrics derived from the answers to this question are then evaluated against comparable metrics that were measured in the DM population. Simply put, the difference between the two groups

is a measure of impact of the DM program. Again, it is not simply a measure of outcomes, but an estimate of the extent to which DM was, at least, partially responsible for causing those outcomes.

In this sense, DM program value is based upon a key assumption that DM population and the reference population(s) are *equivalent*, i.e., that bias or confounding between the two is limited.

Thus, there are three key principles that are necessary to consider when assessing causality.

- 1) A reference population (or a benchmark) must be used in any study involving disease management.
- 2) The disease management population and the reference population must be equivalent to each other.
- 3) The metrics methods used in the DM population and the reference population must be *comparable*.

**THUS, THERE ARE THREE KEY PRINCIPLES THAT ARE NECESSARY TO CONSIDER WHEN ASSESSING CAUSALITY.**

- 1) A REFERENCE POPULATION (OR A BENCHMARK) MUST BE USED IN ANY STUDY INVOLVING DISEASE MANAGEMENT.**
- 2) THE DISEASE MANAGEMENT POPULATION AND THE REFERENCE POPULATION MUST BE EQUIVALENT TO EACH OTHER.**
- 3) THE METRICS METHODS USED IN THE DM POPULATION AND THE REFERENCE POPULATION MUST BE *COMPARABLE*.**

If the populations are *equivalent* and metrics methods are *comparable*, then the comparison between the metrics in the DM population and the reference population may be valid and credible. Thus, the estimate of the difference between the

metrics can be an estimate of program impact. Such a study would have what is called, high “internal validity.” Therefore, the results of such a study may then be used as a surrogate to impute the value of other programs, if these other programs are equivalent to the initial program. This latter concept is termed “generalizability” or “external validity.”

These points about validity are graphically depicted in Figure 1. As can be seen, issues of equivalence and comparability are mirror images of each other in regards to internal and external validity. It should be noted that a study without high internal validity can never have high external validity. However, a study with high internal validity, for example, a double-blind, crossover randomized controlled trial, may or may not have high external validity. The latter can occur, for example, if the sample size was too low, the study population was not representative of the target population (this is common in some randomized control trials due to their strict inclusion/exclusion criteria) and a whole host of other factors. One solution to the problem of low external validity in studies with high internal validity is their replication at multiple times and in multiple places.

## **NOTE ON RANDOMIZED CONTROL TRIALS**

From the outset, we want to address the “gold standard” reference population; i.e. that found in the classic double-blind, crossover randomized clinical trial (RCT). A classic RCT is the rigorous study of willing individuals “blindly” randomized to either a placebo or an intervention group. The goal of this rigorous randomization and blinding is to create equivalence between the reference population and the intervention population at the beginning of the study and throughout the follow-up period. Even factors that are not known risk factors for certain outcomes are, theoretically, divided evenly between the intervention and reference groups. Thus, this kind of RCT — and there others — is the most certain method to achieve the goal of equivalence. In this way, an

RCT is not an end in itself; it is a means to an end, the end being equivalence.

**THUS, THIS KIND OF RCT — AND THERE OTHERS — IS THE MOST CERTAIN METHOD TO ACHIEVE THE GOAL OF EQUIVALENCE. IN THIS WAY, AN RCT IS NOT AN END IN ITSELF; IT IS A MEANS TO AN END, THE END BEING EQUIVALENCE.**

The classic double-blind, crossover randomized trial in DM, or any community-based situation for that matter, requires the extremely difficult task of isolating the DM component(s) in the group of individuals randomized to the intervention group from influencing the group of individuals randomized to the control or reference group in the follow-up period. The lack of isolation may be a threat to equivalency and therefore may bias the results. This can occur, for example, when the reference group adopts some or all elements of the DM intervention. The assumption of equivalence should be addressed at both the beginning of an evaluation and at the end. Thus, even if equivalence is achieved at the beginning of a study between the intervention and reference populations there is no guarantee that equivalence will occur throughout a follow-up period..

That is not to say that “randomization” cannot be used in DM, in fact, as will be seen below the technique of randomization of individuals without blinding<sup>1</sup>, random allocation of providers<sup>2</sup> or hospitals<sup>3</sup> over time can be done successfully using the general characteristics of the follow-up designs. We recommend that randomization be used when possible and ethical. However, RCTs are neither necessary, nor sufficient for a credible assessment of causality.

**RCTS ARE NEITHER NECESSARY, NOR SUFFICIENT FOR A CREDIBLE ASSESSMENT OF CAUSALITY.**

There are also very powerful designs that do not use any form of random allocation that can be very successful in establishing causality. The seminal work, for example, showing a causal relationship between smoking and lung cancer is a prime example of determining causality in an observational setting. Indeed, many factors that are thought to be independent risk factors for disease, such as neighborhood or family-level social support, are virtually impossible to subject to a randomized control trial and must be studied using observational designs. These designs are “weaker” in terms of internal validity from the double blind, crossover RCT, but much more realistic in a field setting. Bias is in issue in all kinds of studies in observational studies and must be dealt with in either the selection or analysis stages.

**THERE ARE ALSO VERY POWERFUL DESIGNS THAT DO NOT USE ANY FORM OF RANDOM ALLOCATION THAT CAN BE VERY SUCCESSFUL IN ESTABLISHING CAUSALITY. THE SEMINAL WORK, FOR EXAMPLE, SHOWING A CAUSAL RELATIONSHIP BETWEEN SMOKING AND LUNG CANCER IS A PRIME EXAMPLE OF DETERMINING CAUSALITY IN AN OBSERVATIONAL SETTING. INDEED, MANY FACTORS THAT ARE THOUGHT TO BE INDEPENDENT RISK FACTORS FOR DISEASE, SUCH AS NEIGHBORHOOD OR FAMILY-LEVEL SOCIAL SUPPORT, ARE VIRTUALLY IMPOSSIBLE TO SUBJECT TO A RANDOMIZED CONTROL TRIAL AND MUST BE STUDIED USING OBSERVATIONAL DESIGNS. THESE DESIGNS ARE “WEAKER”**

The science of epidemiology is built around these kinds of designs as they apply to the health care setting. Thus, the discipline of both authors – epidemiology — is very well suited to help evaluate the impact of DM in real world settings.

## **(B) DM PROGRAM COMPONENTS**

One rule of epidemiology is to be precise in the definition of what is being studied. We start with the definition of Disease Management from the Disease Management Association of America (DMAA):

“Disease Management is a system of coordinated healthcare interventions and communications for populations with conditions in which patient self-care efforts are significant. Disease management: supports the physician or practitioner/patient relationship and plan of care, emphasizes prevention of exacerbations and complications utilizing evidence-based practice guidelines and patient empowerment strategies, and evaluates clinical, humanistic, and economic outcomes on a going basis with the goal of improving overall health. Disease Management components include:

- [1] Population Identification processes.
- [2] Evidence-based practice guidelines.
- [3] Collaborative practice models to include physician and support-service providers.
- [4] Patient self-management education (may include primary prevention, behavior modification programs, and compliance/surveillance).
- [5] Process and outcomes measurement, evaluation, and management.
- [6] Routine reporting/feedback loop (may include communication with patient, physician, health plan and ancillary providers, and practice profiling).

Full Service Disease Management Programs must include all six components. Programs consisting of fewer components are Disease Management Support Services.”<sup>4</sup>

It is important to recognize that measurement

and evaluation (Component #5) is part and parcel of this DM definition. The entire industry must be commended for following the recommendations of clinical experts by “building measurement and data collection” into practice.<sup>5</sup> Moreover, we appreciate the precision of this definition and have translated it into a “causal pathway” schematic that can help us organize our thinking to scientifically test the ideas of DM program impact. As can be seen in Figure 2, components #2 and #6 are the prime intervention agents, components #3 and #4 are the recipients of the intervention; components #1 and especially #5 are the primary topics of this paper.

Programs that have one, a few, or all of these components will be considered part of the operational definition of DM for this paper. Indeed, the fact that only one intervention component has been implemented on patients but not providers, for example, or that one sub-component is done in one population and not another, allows for very innovative and credible evaluation of DM impact. In addition, components not explicitly mentioned in the definition could be evaluated. For example, it has been suggested in the DM literature<sup>1</sup> (and is well known in other circles<sup>6</sup>) that the impact of DM could be due, in part, to the psycho-social support the telephone callers are providing to the participants. Studies could be developed to evaluate this or other independent risk factors for DM outcomes using one of the study designs discussed later in this paper.

## **(C) POPULATION/PATIENT SELECTION**

It is important to note that populations are made up of individuals and the criteria by which individuals are selected into the population are critical to the success of a credible evaluation. One must be explicit as to the criteria that was used to select patients for the DM program as the same criteria can be used to select (or find) patient populations to be used for a reference population(s).

These criteria can be based on factors that are defined by the DM program leaders or by others. An example of the latter is referrals from community physicians. Of course, they can also be based on factors that are defined by those in control of the DM program. These criteria could then be made available to community physicians by the DM program. These would then be used to help determine who would be invited to a DM program. Better (at least from an evaluation perspective) is some standardized algorithm applied equally to all potential participants. This standard algorithm can be based, for example, on results of a medical-based survey (e.g. Health Risk Assessment data) or on software rules applied to health insurance claims data.

The latter can be very simple or complex. One simple example is the identification of an individual based upon their reaching some high cost threshold. Such a model would be based upon the assumption that today's high cost individuals are tomorrow's high cost individuals, an assumption that may not be robust.<sup>7</sup> Another example is the identification of individuals and subsequent invitation to a DM program if they have a certain diagnosis, medication use, inpatient admission use, etc. So called "predictive modeling" algorithms are the state-of-the-art in this kind of patient identification. These use historical or concurrent data on diagnostic and other criteria (e.g. age, gender) to generate statistically derived "weights" that are related to high cost cases in the future. These weights are then applied to individuals in new populations to estimate an individual's risk of becoming a high cost case or a case with an exceptionally high disease burden in the future.<sup>8,9,10</sup>

Figure 3 is a graphic depiction of equivalence.<sup>11</sup> It is essential to compare significant "risk factors" in the DM population to the same risk factors in the selected reference population. The more explicit the criteria for population definition, the easier it is to create (or find) an equivalent population for which to compare the metrics. For example, the internal validity of predictive modeling algorithms is based upon several assumptions, one of them being that the new population is equivalent to the

population from which the weights were derived. Another important assumption is that the statistical model being "fit" was appropriate for the data. This assumption should be subjected to rigorous statistical tests.

A significant bias could result from patients who meet the criteria but refuse to participate in the DM program, or decide to opt-out of the program. Once patients are selected by some criteria, the percent of selected patients that agree to participate (opt-in vs. opt out) should be reported. Also – again related to the reference group issue – an assessment of the differences/similarities of selected patients that participate vs. those that do not participate (also called "non-response" or "refusal bias") must be conducted. To examine this one could compare general characteristics (e.g. age, sex, risk factors, claims costs, etc.) of participants to non-participants in either a claims data set or by selecting a random sample of non-respondents to assess this non-response bias. Alternatively, one could adjust for differences, for example, using direct standardization of other more sophisticated statistical tools that are briefly discussed later.

There is one key issue that will be discussed at this juncture as it is directly related to selection of participants for a DM program: Regression-to-the-mean (RTM). RTM can cause extreme bias, a bias that is likely more common in certain selection strategies (e.g., selection of the high cost patients) than in other selection strategies (e.g. selection of the entire population with a certain diagnosis).

Here is a conceptual example of how RTM works. Say you toss a 6-sided die and you roll a "6" on your initial roll; the probability that you will roll something less than "6" on the next roll is 5/6 or 83%. Similarly if you roll a "1" on the first roll, the probability that on the next roll you will roll something greater than a "1" is 5/6 or 83%. These are all examples of regression-to-the-mean, a phenomenon technically related to randomness. The key is to be aware of the natural tendency of a population to change over time (e.g., in this case the population is a 6 sided die) before one makes a judgment that more than 80% of the "6s," so

to speak, improved because of your intervention. To put it another way, they would regress to the mean, the mean or average in this cases being 3/6.

In health care, the “6s” are equivalent to the highest cost (i.e., top 18%) of patients in a population. Thus, if random fluctuations were totally responsible for changes over time we would expect that these “Health Care 6s,” so to speak, would regress to some mean just like multiple throws of the die regress to the average. To put it another way, if we chose a threshold value of “6”, and we then selected only the die (or patients) that were “6s” the first time around 100% of the population, by definition, would be “6s.” The next time we measured these die (patients), only 18% (plus or minus some margin of error) would be “6s.”

In health care, results from a population that was a non-representative sample (including that chosen by referral, predictive modeling, convenience sample) will most likely be subjected to regression-to-the-mean, and therefore, misinterpretation, if the sample was biased toward individuals at the high (or low) end of a distribution. For example, sub-groups formed from a stratification algorithm into high, medium, and low risk based upon current or prior utilization are subject to RTM. This is most evident when the extreme cases are chosen because they are extreme. For example, a study presented to the AAHP Building Bridges Conference in 2002 showed dramatic change in six separate conditions from the sub-set of the population with the highest costs in the first 30 day time segment compared to interval between day 31 and day 60.<sup>12</sup> Figure 4 graphically depicts this change. These examples are comparable to the second “die” example discussed above, a threshold value was selected and the percent of individuals above that threshold was measured in two time segments. The first measure per condition is, by definition, always 100%. The second, which varies by condition, “regresses to the mean.”

It is important to understand that changes seen over time in populations could be due to either the random effect of “regression-to-the-mean” or non-random effects. These latter changes are technically

not regression-to-the-mean. Obviously, there are “natural,” non-random impacts on sickness. Indeed understanding and treating these non-random effects are the reason we have a health care system.<sup>13</sup> An important future question in DM would be to separate out classic RTM (random error) from changes over time due to the natural history of disease and medical interventions.

## **(D) METRICS**

Metrics can be used to measure many things; however, in a causal sense we need to be explicit about the different kinds of metrics and where they fit in an intervention or causal pathway. We have already discussed population risk factor metrics (e.g. age, sex, etc.). These are used to help understand the equivalence between the DM population and the reference population, between the DM “opt-ins” and the DM “opt-outs,” for example.

Many important efforts are being made to standardize metrics in managed care organizations (e.g., HEDIS)<sup>14</sup> and specific to certain diseases (e.g., American Diabetes Association). These laudable efforts are not the subject of this paper. From the perspective of determining causality what is important is that the metric chosen must be calculated correctly and stated precisely. For example, any kind of population-level standard metric should specify the time period in which it was conducted: “X people out of X+Y people with diabetes in the year 2002 had an admission to an inpatient facility in a certain geographic area” is one example of a precise definition.

Some numeric measures, for example those used as codes for variables (e.g. ICD-9 diagnostic codes), are “nominal” and obviously should not be summed up or averaged. Other numeric measures, e.g. claims dollars, likely have what is called a skewed distribution. One way to determine the skewness of a distribution is to compare the average with the level of the 50<sup>th</sup> percentile, also called the median. If these two numbers are not in close agreement

— a very common occurrence in health care economic data – the distribution is likely skewed and could lead to misinterpretation if averages alone are used. As an example, one may report that the average monthly cost of a congestive heart failure (CHF) patient is \$1,000, while the 50<sup>th</sup> percentile level is \$450. If we were unaware of the skewness of the distribution, we may conclude that a representative group of people randomly chosen from the CHF population would have an average cost of \$1,000, when in fact, that likely would not be the case. In a skewed distribution the average is very much influenced by a very small percentage of the population that were outliers with cost values above, for example, \$500,000. By removing these outliers, we find that the average is only \$500, a number closer to the median. Thus, in cases like this the median or some other percentile level may provide a more accurate picture of a population’s economic profile and is used in well-done research studies.<sup>15</sup>

Alternatively, sophisticated methods are available to “transform” skewed data into distributions that are not skewed: The simplest is to select a “cut-off” point in the distribution and assign a “yes” to those individuals above this cut-off point and “no” to those at or below it. With this “dichotomization,” a simple proportion (or percentage) can be calculated as a summary measure that will not suffer from skewness. This could be expanded to multiple cut-off points (this is often done when viewing distributions of age groups). Other more sophisticated methods are available (e.g., logarithmic transformations), however, and experts should be consulted prior to conducting these transformations.

One more issue that is very important in the reporting of summary measures is related to the extent to which individuals in the population vary from each other. This variation is a key component of any method to test statistical significance. It can be easily calculated in a simple proportion without having access to person-level data, but to calculate standard deviations from averages requires individual level data.

Assuming metrics are measured correctly and stated precisely, from the perspective of the evaluation of a program, it is very important to ensure (or validate) that they are measured the same way in both the DM and the reference population. This concept of comparability is depicted in Figure 5. This figure also introduces the terminology we will use throughout the paper regarding the three metrics types.

We define these metric types separately to help differentiate where each lies along the intervention or causal pathway.

“Type I” metrics are essentially measures of the operation of the DM program. These “program process metrics” gauge “What was done?” These could include the number of patients enrolled, the number of attempted phone calls and completed phone calls, the questions that were asked during the interview, the number of guidelines sent to providers, the number of feedback reports sent to providers, and so forth.

**“TYPE I” METRICS ARE ESSENTIALLY MEASURES OF THE OPERATION OF THE DM PROGRAM. THESE “PROGRAM PROCESS METRICS” GAUGE “WHAT WAS DONE?”**

A “Type II” metric is the proximate reaction or response that the patient or provider has to the intervention captured by a Type I metric. These “proximate outcome” metrics can also be a more formal intermediate outcome that may be an early indicator of the success of the DM program. Examples of these proximate outcomes include the measures showing the extent to which a patient actually received a recommended screening procedure, another would be based upon adherence to a medication regimen. In a congestive heart failure program, for example, it might be the frequency with which a patient weighed himself or herself on a daily basis, in a diabetes management program it might be the frequency with which a self-test of blood glucose was

conducted. In essence, a Type II metric is the thing which changed which then in turn caused a change in the Type III metric.

**A “TYPE II” METRIC IS THE PROXIMATE REACTION OR RESPONSE THAT THE PATIENT OR PROVIDER HAS TO THE INTERVENTION CAPTURED BY A TYPE I METRIC. THESE “PROXIMATE OUTCOME” METRICS CAN ALSO BE A MORE FORMAL INTERMEDIATE OUTCOME THAT MAY BE AN EARLY INDICATOR OF THE SUCCESS OF THE DM PROGRAM.**

“Type III” metrics or ultimate outcome metrics are measures of the principle goals of a DM program. We have divided these kinds of metrics into three broad categories: Health (both clinical and quality of life), Economic, and Satisfaction.

**“TYPE III” METRICS OR ULTIMATE OUTCOME METRICS ARE MEASURES OF THE PRINCIPLE GOALS OF A DM PROGRAM. WE HAVE DIVIDED THESE KINDS OF METRICS INTO THREE BROAD CATEGORIES: HEALTH (BOTH CLINICAL AND QUALITY OF LIFE), ECONOMIC, AND SATISFACTION.**

A Type III health metric, again including both clinical and quality of life issues, is a very broad category and is based on the definition of health that the World Health Organization proposed in 1978 in the famous Alma-Ata Declaration:

*The [Alma-Ata] Conference strongly affirms that health, which is a state of complete physical, mental, and social well-being, and not merely the absence of disease or infirmity, is a fundamental human right and that the attainment of the highest possible level of health is a most important world-wide social*

*goal whose realization requires the action of many other social and economic sectors in addition to the health sector.*<sup>16</sup>

Thus, in this sense, the broad term health includes both clinical and humanistic outcomes. Examples of the former are the actual HbA1c level assessed during a screening for diabetics or the results of a blood pressure test or lipid profile for cardiovascular patients. Numerous other clinical metrics are Type III metrics if they are the ultimate target of the DM program; they might include tumor size, low birth weight babies, incident events such as a heart attack or a fall in a nursing home, mortality, new co-morbidities, and so forth. Examples of humanistic” outcomes included functional status, general health-related quality of life measures, for example the SF-36 or SF-12 questionnaire,<sup>17</sup> years of life gained<sup>18</sup>, disability-adjusted life years (DALYs)<sup>19</sup>, and quality-adjusted life years (QALYs). QALYs are of special interest as they combine mortality related measures (years of life gained) and morbidity measures (quality during those years of life gained) into one measure. These metrics are “a measure of health outcome which assigns to each period of time a weight, usually ranging from 0 to 1, corresponding to the health-related quality of life during that period, where a weight of 1 corresponds to optimal health, and a weight of 0 corresponds to a health state judged equivalent to death; these are then aggregated across time periods.”<sup>20</sup> Proponents argue that QALYs can be used to help allocate limited health resources in a rational way, however, it must be said, they are not without controversy.<sup>21</sup> They have been used to assess the value of primary prevention in diabetes.<sup>22</sup> In addition, the value of medical management, for example, in asthma<sup>23</sup> and multiple sclerosis<sup>24</sup> have been assessed using QALYs. We are aware of several groups in the United States working to incorporate QALYs into everyday population health policy. Thus, we recommend that DM leaders work with experts in developing health related quality of life measures, like QALYs, and wherever possible to incorporate QALYs into DM programs evaluations.

Type III metrics representing economic factors

include raw financial information or standard utilization data and would include, for example, total claims costs during a specified time period, number of inpatient admissions during a specified time period, return-on-investments metrics, market share and so forth.

The satisfaction Type III metrics are based on the results of provider and/or patient satisfaction surveys.<sup>25</sup>

The Disease Management Association of America is very explicit about the need to base any DM program intervention on evidence-based guidelines. Thus, every attempt should be made to include some of the standard metrics that are specific to the disease of interest. These metrics, listed in standard guidelines such as the American Diabetic Association (ADA)<sup>26</sup> or the National Heart Lung Blood Institute (NHLBI) standards for asthma<sup>27</sup> should be consulted. We strongly stand with the DMAA, that these standard metrics, Type II and Type III in our parlance, are always monitored in DM programs. This should occur even in the absence of a specific study to assess “cause” and “effect.” These metrics can be part of the surveillance effort for quality assurance and quality improvement. The classic work of Berwick’s Institute of Health Care Improvement is worth reviewing in this light.<sup>28</sup>

As a concluding thought, metrics by themselves are not directly related to cause and effect, however, if they are compiled incorrectly or otherwise invalid, they will undermine any determination of cause-effect in a formal study.

**AS A CONCLUDING THOUGHT, METRICS BY THEMSELVES ARE NOT DIRECTLY RELATED TO CAUSE AND EFFECT, HOWEVER, IF THEY ARE COMPILED INCORRECTLY OR OTHERWISE INVALID, THEY WILL UNDERMINE ANY DETERMINATION OF CAUSE-EFFECT IN A FORMAL STUDY.**

## **(E) INTERVENTION OR CAUSAL PATHWAY**

To evaluate causality in a DM program, it is important to clearly spell out the causal pathway. Using the nomenclature introduced above, a typical hypothesis would be that a Type I metric (DM program based on evidence-based guidelines and feedback mechanisms to providers and patients) leads to a Type II metric (proximate outcome) which leads to a Type III metric (ultimate outcome). Thus, the hypothesis would be in this form depicted in Figure 6.

If we take “components” directly from the DM definition, we see four basic hypotheses that could be tested in a DM program. These hypotheses all start with the DM program but arrive at the outcome through different “components.”

The four types are as follows:

- 1) DM (Level I) → *Patient* → Patient-based outcome (Level II & III: Health, Economic and/or Satisfaction).
- 2) DM (Level I) → *Patient* → *Provider* → Patient → Patient-based outcome (Level II & III: Health, Economic and/or Satisfaction (e.g. where patient informs doctor of symptoms, new drugs available, etc., which leads to patient behavioral change).
- 3) DM (Level I) → *Provider* → *Patient* → Patient-based outcome (Level II & III: Health, Economic and/or Satisfaction).
- 4) DM (Level I) → *Provider* (Level III: Satisfaction).

Although a “full service” DM program to include both patient and provider interventions is ideal, it is possible that in certain circumstances both components will not have been implemented. From the perspective of an evaluation this incomplete rollout in a certain market could be very valuable. It is akin to a “natural experiment” and this population can serve as a reference population for the markets where there was complete rollout. So,

an understanding of the different ways your DM program “hypothetically” reaches desired outcomes could be a huge advantage when evaluating impact of one or more components or sub-components.

From the perspective of the patient or the purchaser, “value” does not necessarily hinge completely on the causal determination from these studies. In many cases, the existence of a program at all is valuable enough to argue for its continuation, thus a Type I metric by itself showing that phone calls were made may be sufficient, at least initially, to a customer. Similarly, a description of a HEDIS metric improving each year (a Type II or Type III metric) may be enough to justify a DM program. That said, we believe the long-term viability of a DM program will hinge on its ability to suggest “causation.” Any analysis of return on investment (ROI), for example, should take these concepts of equivalence and comparability very seriously or outcome measures like ROI will be inaccurate. Indeed, the very existence of a paper broadly addressing causality specific to DM is testimony to the need in population-based health care to move toward that ultimate goal. The remainder of this paper is focused on that task.

**THAT SAID, WE BELIEVE THE LONG-TERM VIABILITY OF A DM PROGRAM WILL HINGE ON ITS ABILITY TO SUGGEST “CAUSATION.” ANY ANALYSIS OF RETURN ON INVESTMENT (ROI), FOR EXAMPLE, SHOULD TAKE THESE CONCEPTS OF EQUIVALENCE AND COMPARABILITY VERY SERIOUSLY OR OUTCOME MEASURES LIKE ROI WILL BE INACCURATE.**

**(F) BIAS/CONFOUNDING POTENTIAL**

What exactly is bias? Simply stated, it is non-equivalence between reference and intervention population on risk factors or non-comparability in

metric methods. In the language of evaluation, bias is a “threat” to the internal validity of the study. This bias can take many forms including selection bias, random error bias, and confounding.

Without any comparison to a reference population, the “interpretation” of the outcome metrics collected to the DM program implemented is more and more difficult as you go from Type I, Type II, and Type III. Not surprisingly, the control the DM program managers will decline incrementally as one moves down the causal pathway from Type I to Type II to Type III metrics, while at the same time the degree of potential bias will likely increase.

Metric Type:	Type I	Type II	Type III
Level of Direct Control by DM	High	Middle	Low
Level of Potential Bias	Low	Middle	High

Factors that can vary between the reference group and the intervention group include:

**1. Economic Factors**

One solution to this problem is to adjust economic variables (e.g., claims payments) for general inflation, based upon a special kind of external “reference population,” i.e., the Bureau of Labor Statistics. This may not adjust perfectly as inflation could be “disease-specific” or could be different in one region compared to another. Other economic factors to be considered include a restating of all future costs and saving in terms of their “present value.”<sup>20</sup> We recommend consulting an expert in economics or finance to help “adjust” for these economic factors.

**2. Statistical issues (random error)**

“Statistical significance” can be dealt with using statistical inferential techniques, available in any statistics textbook.<sup>29</sup> It should be cautioned that a “statistically significant p-value” is not related to causality by itself, it must be interpreted within the framework of the selected study design and other biases (i.e., other than random error or “precision”

which it is designed to deal with) that will likely be found in any study. In addition, different kinds of statistical tests are designed from different kinds of data (for example, a t-test should not be used on data that is highly skewed). We recommend consulting an expert in epidemiology or statistics to ensure the use of an appropriate statistical test and its interpretation.

### 3. Lost to follow-up

The Consolidated Standards of Reporting Trials (CONSORT) check list<sup>30</sup> used in scientific medical research recommends the all trials follow, among other things, intent-to-treat methodology. This means that all individuals enrolled at the beginning in the trial must be accounted for in the evaluation at the end of the trial. Even in randomized control trials, this principle is not always followed.<sup>31</sup> From the perspective of DM, this means that all individuals selected for possible inclusion in the program, including those who agreed to participate but were later lost-to-follow-up need to be accounted for. Patients who met entrance criteria but are not longer participating in the program in the follow-up period can be a very challenging issue from an evaluation perspective. The initial approach is to determine if these patients were substantially different from those remaining in the program. Lost to follow-up could be due to a change in doctors, moving away, changing health insurance companies, changing doctors, dropping out for no known reason, becoming too ill to be in DM and being admitted to a nursing homes or a hospice, lack of motivation on the part of the patient, and mortality.

Some of the comparative analysis introduced in section B (population selection) may be useful to determine if those lost-to-follow-up were different than those who remained in the program. If the two groups were different in the aggregate, methods are available to “adjust” for this problem. A few are introduced below.

### 4. Confounding and Methods to Deal with Non-Equivalence

One of the major questions encountered in evaluating DM programs is what is the residual

program effect when we take into account something else that might have influenced the outcome. Something else may include plan-specific issues (e.g., benefit changes, contracting changes, membership changes); disease-specific issues: (e.g., new technologies, results of new studies, natural course of disease); or patient-level issues: (e.g. independent risk factors for the outcome, co-morbidity or age). Statisticians have phrases for these including “stratifying to examine the influence of something,” “controlling for something,” or “adjusting for something.”

What do we really mean, for example, when we “control for age” in the comparison of the effects of a program on two groups who differed in their age patterns? Simply this: if two groups had the same age patterns, what would be the difference in the effect of the program (less, the same, greater)? And that’s exactly what a statistical control does: it gives us the effect of a predictor (program participation) with at least one other predictor (for example, member’s age) held constant.

Some researchers also would call age a confounding variable or a confounder in examining the relationship between participation in the program and an outcome effect that is being compared. No matter what statistical adjustments are done the issue of time ordering (did the difference in the groups occur before or after the DM program intervention?) will limit your ability to say that a causal impact occurred from participation in the DM program. With this caveat in mind, how can we seek to take age into account and make a more fair comparison? Methods to make adjustments include, but are not limited to, stratification, standardization, multivariate modeling, and matching.

- **Stratification:** This involves comparing the program effect on people, for example, of the same age categories in each group. Statistical tests can be used to determine if the differences between the groups in each category are not likely to be due to chance.

- **Standardization:** Rates, proportions, or averages (means) can be calculated that take into account the effects that might occur due to the differences in ages of members in the two groups. For example, the “direct standardization” uses an outside population’s distribution (e.g., for the United States) to remove misleading results due to uneven distribution of certain demographic factors among populations being compared. Other methods are available as well; for instance, age and/or sex distribution factors are commonly adjusted using statistical modeling when doing rate comparison.
- **Multivariate Modeling:** If more than one confounding variable is of concern in making the comparison between the groups, then multivariate modeling (or stratified analysis) can be used to seek to control for the effects of variables. Statistical tests can again be used to determine if the participation in the program had an effect not due to chance after other factors have been statistically taken into account.
- **Matching:** One or more confounding variables can be matched when comparing the two groups. For example, when comparing the DM program effect, the difference in the program between members who got the program to members not getting the program who were individually matched to have the same age and gender could be compared. Age and gender would not be variables confounding this comparison of program effect if this matching was done.

The technical methods, tools, and the assumptions necessary to legitimately conduct these quantitative adjustments are far beyond the scope of this paper. A good introductory biostatistics or epidemiology text should be consulted.<sup>32</sup> We strongly recommend seeking the advice of an expert in quantitative methods in the health care field, for example, a biostatistician or an epidemiologist when planning an evaluation study and before conducting these analyses.

## (G)STUDY DESIGNS FOR DM

In this section we will introduce the seven different study design categories (numbered with roman numerals I—VII). Each category includes one or more study designs. In total, seventeen separate designs are discussed. This list of seventeen is only introductory; it is not by any means a comprehensive list of study designs available to DM. The framework is in outline form and will easily allow for the addition of other pertinent study designs in the future. Figure 7 is a conceptual summary of the general issues we have been discussing throughout this paper. It highlights that comparable metric methods between the DM population and the reference population are needed. To assess impact (or effect) we need to examine the difference between these metrics – while at the same time taking into account the potential impact of alternative “causes” of the outcome. The test of the hypothesis that DM has impact can be accomplished by the use of credible study designs.

Table 1 lists the seventeen different study designs. It summarizes the components of the designs using three categories: the Reference Population (patient as their own control, DM exposure variation, no exposure to DM, and benchmark), Random Allocation (Yes or No), and the Unit of Measurement (Group or Individual), and Examples. In this table, we have included a limited number of article or book references (based upon a computerized search among the journals in Index Medicus using the term “disease management” and “cost”) and/or references to books that discuss the details of some of the methods. These references are not comprehensive by any means. We strongly recommend that a follow-up article with the explicit purpose of cataloging a comprehensive list of published and non-published articles that have used one of the 17 study designs outlined here.

In all cases, the internal validity of these designs is dependent on the equivalence of the two (or more populations) and comparability of the metric methods used. The issue of equivalence

and comparability must be addressed whenever these designs are adopted. We cannot state with any degree of certainty which is better in all circumstances. In general, the level of credibility has to do with their level of control (in design or evaluation) that the evaluation team has over the intervention as well as all other factors which could independently impact outcome (i.e. confounding factors).

As these study design are introduced, we will illustrate each based upon a case study of a “real world” (but simulated) congestive heart failure program. These examples will appear in TEXT BOXES for all the 17 study designs.

### I. Post-Test Studies

These are the simplest kind of study and the type we would generally not recommend because there is no explicit reference group. These studies are often termed “Post only” as patients are treated, managed and observed over time. Patients could be measured at one point in time or at multiple points in time. They can be “high risk” strategy (1.1 in Table 1) or a “population” strategy (1.2 in Table 1).<sup>33</sup> One important difference between these two strategies is that regression-to-the-mean is much diminished in a population strategy when compared to a high-risk strategy. Of course, the panel of patients enrolled in these kinds of studies can be compared to a reference population. By doing so, the investigation would advance from this category to a Benchmark or other design, something we highly recommend.

In the history of medicine, post-only studies were very common and can bring great value in some circumstances. For example, think of a trauma unit at a ski resort that treats fractured bones. Most people would not subject those injured skiers to any kind of study except post-only. However, that said, when studying chronic diseases, the cause-effect relationship is not as clear-cut as the impact of a doctor successfully setting a broken femur.

### II. Benchmark Studies

In these studies the participants in a DM program and their Type II or Type III results are compared to

a benchmark population from secondary literature. We have differentiated these benchmark studies into three sub-categories.

“II.1”. National/Regional Benchmark: In this design the patient panel results are compared to a national or regional benchmark (e.g. HEDIS measures, SF12 metrics, ADA guidelines for diabetes, or the NHLBI guidelines for Asthma cited earlier).

“II.2”. Peer-reviewed study as benchmark: In this design, the patient panel results are compared to the more rigorous results from a peer-reviewed experimental or observational study.

“II.3” Statistical Prediction as benchmark: In this design the patients are selected based upon their “probability” of being high cost in a future time period. Moreover, a prediction of the estimated costs can be made prospectively and compared to the actual results.

The credibility of these designs in assessing causality is dependent on “external validity” of the reference population, i.e., the equivalence between the DM population and the reference population and the comparability of the metric methods between the two groups.

### III. Quasi-Experimental

Quasi-experimental designs are studies that have some allocation of an “exposure” but this allocation is not based upon a randomization scheme. We cite three examples of quasi-experimental studies:

“III.1” The Pre-Post Study: This is where the DM intervention occurs at the beginning of the “Post” period. The reference group often becomes the time period of the same patients prior to this enrollment (but it could be the entire population in the pre period, even if they were not enrolled in the post period); thus, in its most common form, this design has the “patient as their own control.” The validity of this kind of study, all other things being equal, is totally dependent on the correctness of the

assumption that the experience in the “before” period is a good predictor of the experience of the population in the “after” period. If that assumption is true, the differences in the results observed during the “after” period compared to the results “predicted” from the “before” period are valid. In that case, an accurate assessment of the “causality” of the DM program can be derived. Some purchasers of DM, however, may be skeptical of the validity of the assumption. They may ask: “Was the ‘before’ period in this population a good predictor of outcomes that would have occurred in the “after” period in the absence of DM? How do we know?” These people could logically state that advancements in medicine and technology, even the adoption of the principles of DM by community physicians in the “after” year, were responsible for the improvement seen. This is a legitimate argument and needs to be addressed by the evaluators of the DM program that use the pre-post methodology.

“III.2” Time-Series: A time series is essentially an expansion of the pre-post study into multiple time periods and could include, perhaps, multiple interventions over time.

“III.3” “Regression-Discontinuity” may be a powerful methodology to evaluate DM, but to our knowledge, has not been used extensively. The regression-discontinuity design chooses a “cut-off” point (high cost, for example) where patients above the cut-off point receive the DM intervention and those at or below the cut off point do not receive the DM intervention. The relationship between the pre-post measure of those below this cut-off is compared to the linear relationship of those above the cut-off point, if there is a “discontinuity” in the regression lines (and all assumptions are met) the program is thought to be causal.

The validity of these quasi-experimental designs is very dependent upon the equivalence of the intervention and reference population. In these designs it is likely that the metric methods are

comparable. The specific methodology of each is discussed in Cook and Campbell.

#### **IV. Ecological**

Ecological designs are unique in that the group is the unit of measurement. There is no information on the relationship between a specific exposure and a specific outcome at the individual level. These studies are very useful in the initial test of hypotheses and can be used as justification for a more rigorous design. This study type is listed as “IV.1” in Table 1.

#### **V. Cross-Sectional**

A cross-sectional study is a sample (ideally a random sample) at one point in time (sometimes called a “snap shot”) of a population. Its purpose is to examine the association, though usually not causal, between a specific “cause” and a specific “effect.” Because it is taken at a single point in time, it is often difficult to determine temporality; i.e. that a specific cause preceded a specific effect. One advantage of a cross-sectional study is that the “dose” of a DM program can be associated with an outcome in each person sampled. Thus, an initial dose-response table could be created that may be used later as the basis for a case-control or follow-up study. Importantly, with the addition of a second sample from a different place or different point in time (similar to a pre-post study, but each patient is NOT his or her own control), where a DM program was not as well established, could lead to some important conclusions about the impact of a DM program.

V.1. One sample.

V.2. Two or more samples

#### **VI. Case-Control**

A case-control study is where individuals are selected because they have already experienced (or just experienced) an outcome of interest. This design, with the notation “VI.1” in Table 1 is one of the great achievements of epidemiology as it allows for an examination of risk factors where the incidence of an outcome is very rare. In DM one could adopt this design by selecting “cases” that

experienced, for example, an in-patient hospital stay and then compare their “exposure” to a factor from a prior time period that was thought to be related to the outcome (a risk factor, participation in DM program, etc.). For example, cases and controls could be examined for compliance to medication and if the two populations are equivalent, any difference (while controlling for other variables) in medication compliance may be shown to be related to the hospital admission. If this were the case, a DM program would be well served to improve their strategies for management of medication. There are many variations to this design and it is recommended that an expert epidemiologist be consulted before engaging in a case-control study.

### VII. Follow-up Design

In general, these are studies where a group of individuals are recruited and the exposure is measured at baseline (this could also include measurements a pre- period in both the intervention and reference groups) and outcomes are measured prospectively, often at multiple points in time. These studies could be “observational” or “experimental” and measurements are made at the individual patient level. In the observational cohort (V1.1), the “control” for confounding is extremely important. In the other designs, random allocation can be done at either the group or the individual level. These “experimental” designs reduce the likelihood that the intervention and reference populations are not equivalent, but, as said before, are no guarantees. In these designs equivalence could be achieved at baseline, but if the reference population is affected by newly introduced DM-type intervention “external” to the DM managers, the findings could be severely biased. However, if this “threat to internal validity” does not occur, causality determination is the highest in these experimental follow-up designs.

Table 1 highlights five kinds of follow-up studies.

VII.1 Cohort (e.g., a pre-post study with a reference group)

VII.2.i. Field randomized control trial (FRCT) by

Place. (e.g., Provider, Hospital, Region, Market)

VII.2.ii. Field randomized control trial (FRCT) by Time: Structured roll-out

VII.3.i Randomized control trial (RCT): Individual not-blinded

VII.3.ii. Classic Randomized control trial (RCT): Individual blinded

We want to address a study design in DM that meets the observational follow-up design criteria. In a form of this design, the results of participants (“opt-in”) in a DM program are compared to the results garnered from non-participants (“opt-out”) during the same time period. The adoption of this design, in at least one respect, deflates the argument that the effect seen in the DM population was due to time-based technological advancements as both groups were followed-up concurrently. That said, the equivalent assumption must be seriously questioned. Were the results of the non-participant’s truly credible predictors of the results based upon the DM participants, if the latter had not participated in the DM program? Again, a discerning customer might argue that patients who agree to participate in a health management program are quite different from those who do not agree to participate (e.g., age, gender, severity of disease, level of satisfaction or dissatisfaction with current care, willingness or unwillingness to assume self-management responsibilities over current disease processes, and a host of other factors.) To test the equivalence assumption it is essential to compare important risk factors in both groups using comparable metrics and metric methods. This could be done via a simple random sample of non-participants compared to a random sample of participants to test the accuracy of the key assumption of this study design. If it turns out that the non-participants are not equivalent to the participants; attempts should be made to try identify a group of people who are equivalent to the participants. This may prove to be difficult, but is not necessarily impossible if one can “match” on the variable willingness to participate.

We also want to reiterate that the higher forms of follow-up studies, those using some form of randomization, also require tests of the equivalence assumption. Randomization is no guarantee of equivalence at the beginning of study not or at the end as patients are “lost to follow-up.” Standard adjustment methods used in less rigorous designs must be used to take into account non-equivalence in these designs too. An important factor that can occur during the follow-up period for all of these follow-up designs (except VII.3.ii, the “blinded” form) is the problems that emerge when a reference group becomes acutely aware of their “control group” status and chooses (either by themselves, or by their providers, etc.) to adopt the “interventions” of the DM population. If DM interventions were truly effective and investigators were unaware of this important development – there may be no observed differences in the outcome metrics between the intervention and reference populations. In this situation, the incorrect conclusion that DM was NOT effective could be propagated to the detriment of the health of the population.

This section has highlighted various study designs that can be used as measures of assessing causality. We recommend that all studies that are done in DM programs be submitted for publication in peer-reviewed journals.

## **ASSESSING CAUSALITY: OTHER VIEWS:**

We have attempted to introduce the basic elements in assessing causality in DM. Other investigators have rated the strength of evidence that is based on the type of study design adopted. For example, a well-known grading system for “level of evidence”<sup>48</sup> used in medical research grades RCTs as an “A,” however, as we have stated, in many cases a classic double-blinded RCT is not possible in DM programs. The follow-up and case-control designs are given a “B” in this scheme and the post-test only is give a “C.” Quasi-experimental and

benchmark design are not graded in this system. Thus, the the “levels of evidence” system cannot be perfectly adopted by the DM industry. Apart from the post-test only, we would give all of our designs a “B.” This grade assumes that in the description of the evaluation, the level of equivalence and metric comparability be included.

Given this high likelihood that DM will not be judged on the results of a double-blind randomized control trial, it is beneficial to learn from the “masters” of observational research. Based upon his work on establishing the link between smoking and lung cancer shortly after World War II, Sir Austin Bradford Hill listed several criteria that should be considered when assessing the relationship between cause and effect. These criteria were developed with “disease” as the outcome; we have modified them a bit to make them more relevant to DM.<sup>49</sup>

- 1) Strength of Association: If we see an association between DM and an outcome, what is the magnitude? The higher the magnitude the more likely the relationship is causal.
- 2) Consistency upon repetition: To what extent has a similar relationship been observed in different places, circumstances, and times?
- 3) Specificity: Is the cause found when the outcome is present? Does the outcome usually result when the cause is present?
- 4) Time Sequence: Does the causal factor or “exposure” precede the outcome? This is the most important criteria of all. It may sound obvious, but it must be established that the intervention (Type I) preceded in time the proximate outcome (Type II) which, in turn, occurred before the ultimate outcome (Type III).
- 5) Biological gradient: If we change the term from “biological” to “Type I, Type II, or Type III gradient,” this criterion is relevant. In question form: Does more exposure to various components of DM, on the one hand, or higher amounts per unit time in DM, on the other, lead to more pronounced Type II or Type III

outcomes? In other words, what is the “dose-response” relationship?

- 6) Plausibility: Is the observed association consistent with biological ideas, e.g. consensus guidelines, the known process of disease?
- 7) Coherence of Explanation: Does the entire set of observations fit together?
- 8) Experiment: Does removal of the DM exposure (or sub-components of DM) result in a change in the outcome? How does treatment affect different groups in experimental studies?
- 9) Analogy: Are similar known patterns of cause and effect found in other areas of observational epidemiology?

Although it is not critical that all these lines of evidence be presented to support the notion of causality, the more that are supported, the more the case of causality is supported.

## **(H) SUMMARY/CONCLUSION SESSION**

The paper was developed to provide guidelines in advancing the rigor of DM program evaluation. Ways of thinking and methods to assess causality of disease management programs were introduced. We summarize our key points and recommendations below.

- 1) The DM program must explicitly state the metrics used to measure both the actual intervention (“cause”) and the resulting impact (hypothesized) of that intervention (“effect”).
- 2) The DM program metrics must be compared to metrics from an independent source, broadly defined as a “reference population.” At a minimum, this referent metric should be based on benchmark values from commonly accepted practice guidelines.

- 3) Statements that the DM program “caused” specific outcomes (e.g. cash savings) must be based on investigations that have adopted study designs (e.g. quasi-experimental, case-control, or follow-up) more rigorous than that of a post only design. Based upon “levels of evidence” criteria, the study designs should be at least at a grade “B” level and should acknowledge and “control” for factors, where possible, that may have impacted the outcome (so-called “confounding variables”).
- 4) It is unlikely that one study will provide definitive proof of DM program value, thus, we recommend that studies be on-going, at multiple points in time, and at multiple sites. Moreover, if feasible, more than one kind of study design should be used.
- 5) Finally, all studies purporting to show the “value” of DM interventions should be capable of passing peer-review such as through a submission to a peer-reviewed journal or an outcomes validation program.

It is our belief that the establishment of any degree of methodological sophistication will greatly enhance the reputation of the important and growing efforts in disease management. However, it must be said that this paper is only an introduction to some of the issues involved in assessing causality and value of a DM program. We strongly urge the readers to consult experts in epidemiology and statistics as well as textbooks and articles on the methods discussed here.



## FIGURES, TEXT BOXES AND TABLES

Figure 1: Validity Issues: Equivalence & Comparability

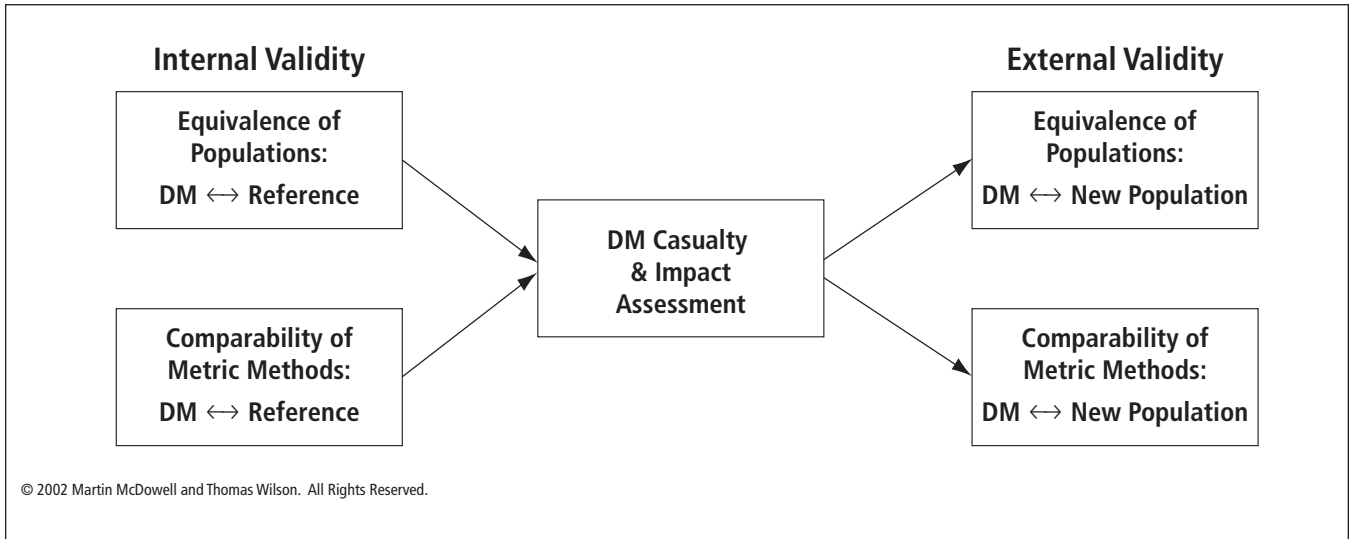


Figure 2: Schematic of "DMAA Components" of DM Programs Showing Intervention Pathway

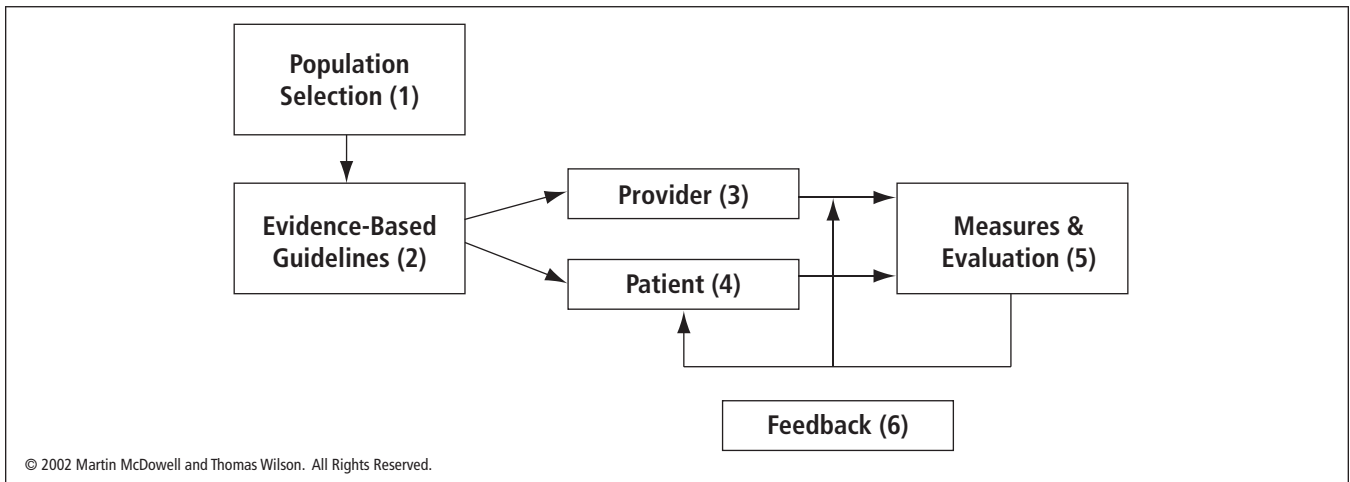


Figure 3: Equivalence between Populations

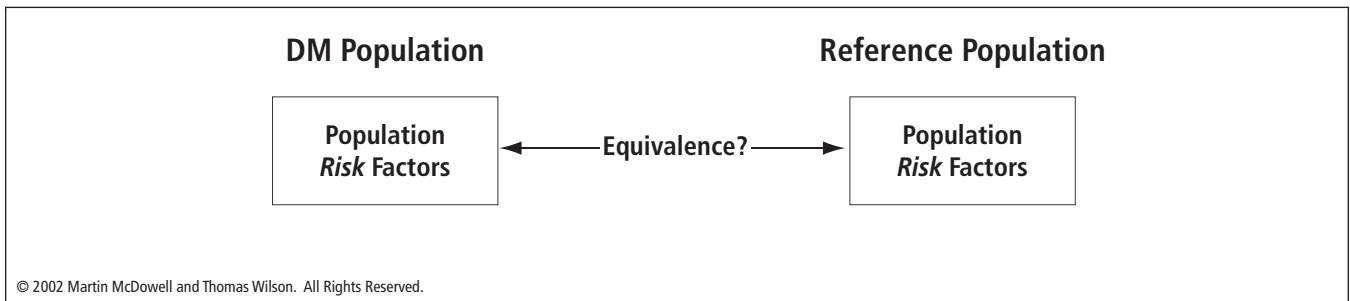
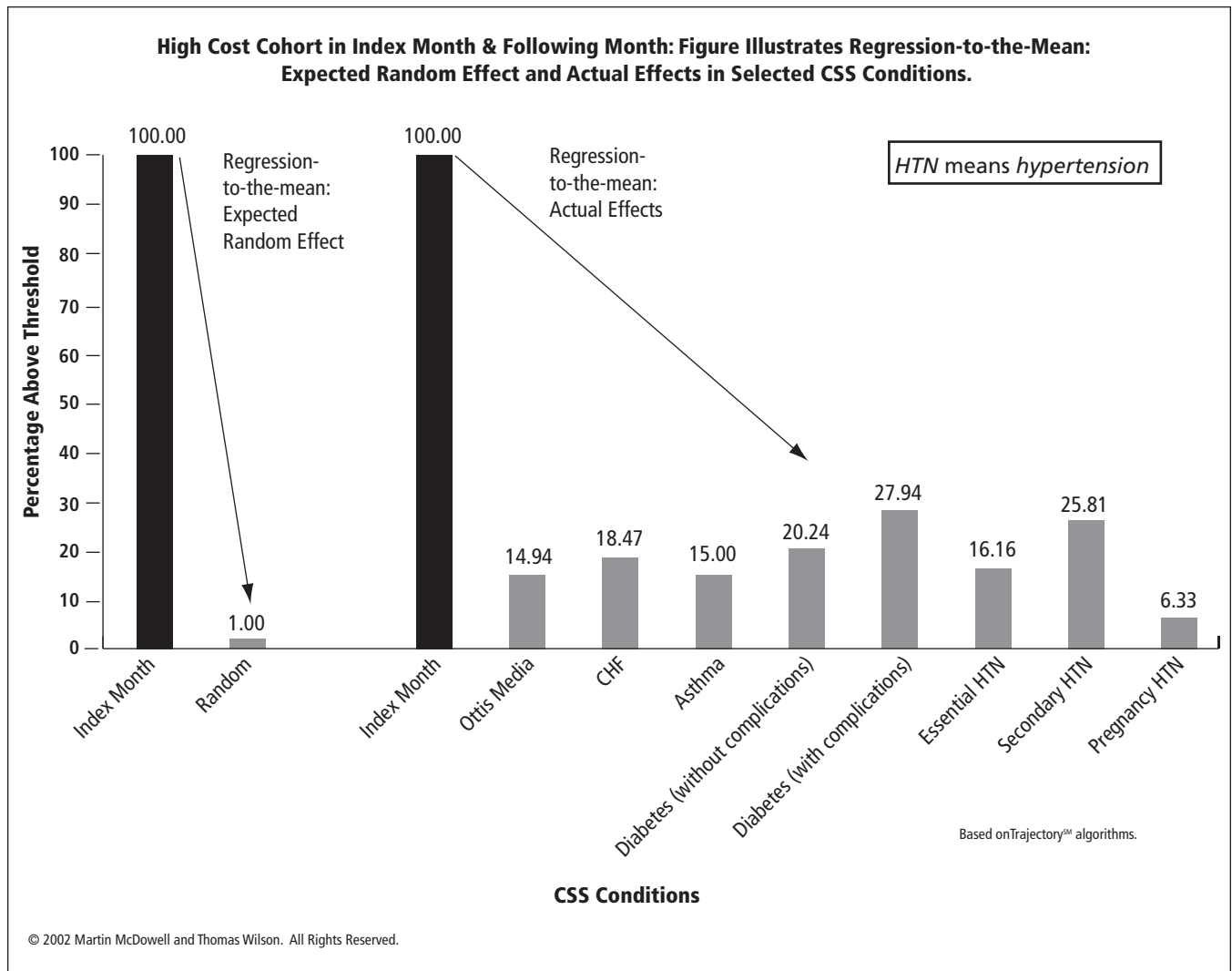
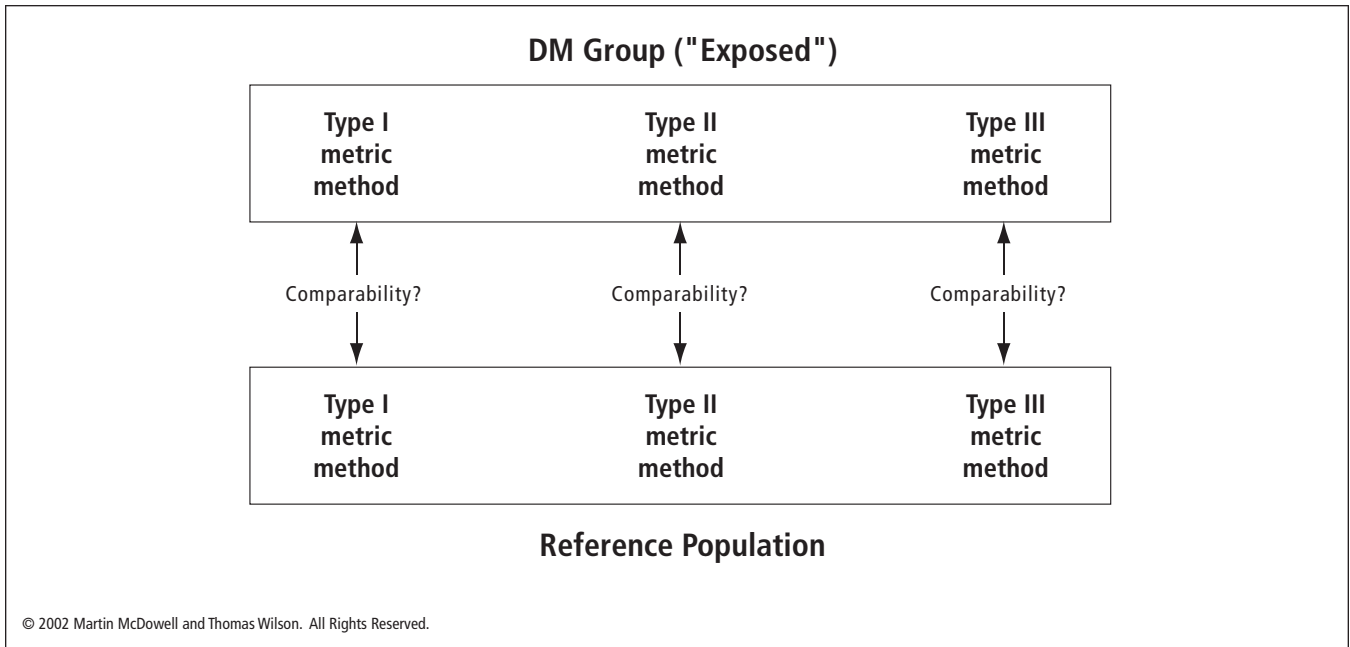


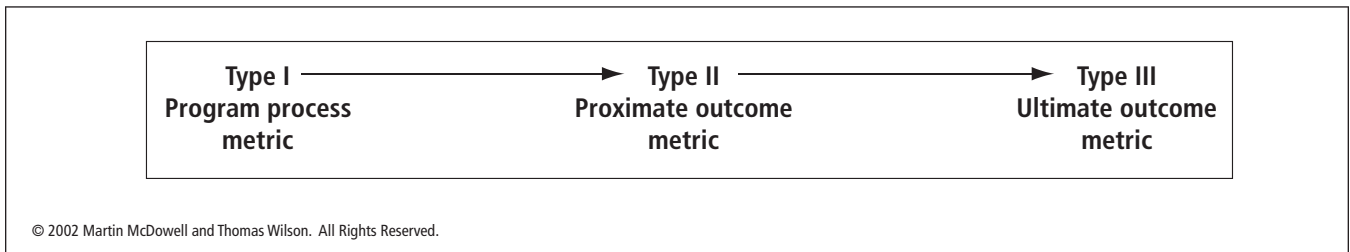
Figure 4: Illustration of Regression-to-the-Mean



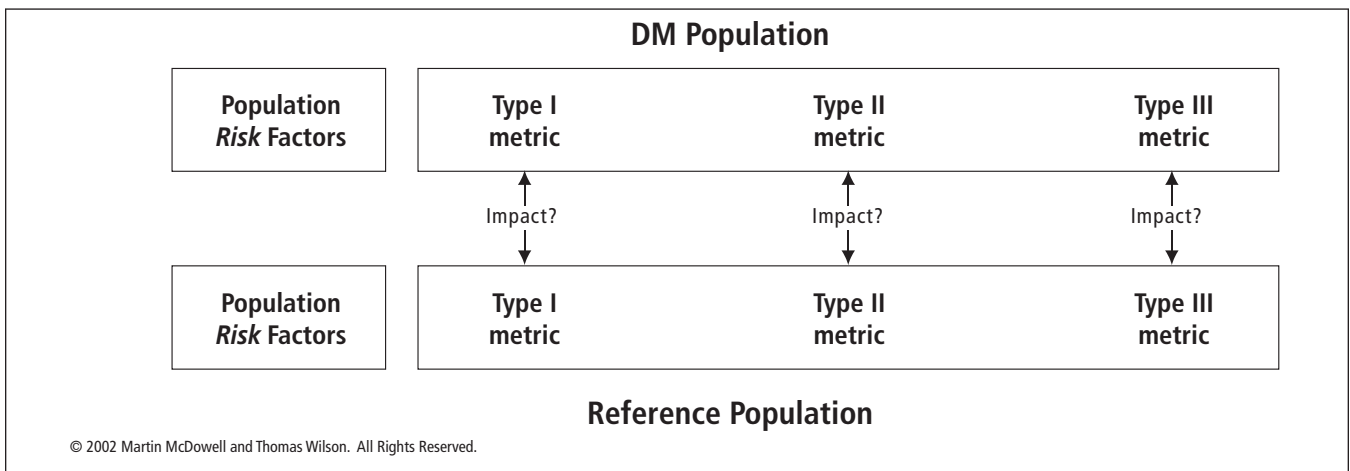
**Figure 5: Comparability of Metric Methods**



**Figure 6: Conceptual Testable Hypothesis in DM**



**Figure 7: Causality: Assessing DM program impact between "equivalent populations from "comparable" metrics methods.**



## Text Box 1: Basic components of CHF simulated case study

### Scenario in a managed care company:

- A) Congestive Heart Failure Patients are selected (various strategies depends on study design).
- B) DM Program: Evidence-based guidelines are followed. Interventions are conducted, potentially, with both patients and providers
- 1) Patient Level Intervention: Telephone Calls to remind participants to monitor body weight, sodium intake, and to adhere to medication regimen. Mailed reminders to monitor symptoms of fluid retention and to contact DM to discuss timely adjustments to medications.. Patients are provided feedback
  - 2) Provider Level Intervention: Evidence-based guidelines are disseminated. DM program can periodically provide aggregate-level and patient level feedback.
- C) Metric Types available
- |                          |                                                                                                                                                                                                                                                                                                                  |
|--------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Population Risk Factors: | Age, Sex, Co-morbidities, Cognitive ability.                                                                                                                                                                                                                                                                     |
| Type I:                  | Based on Calls to Patients over some specified time period.<br>Based on periodic mailed reminders to take medication.<br>Based on Guidelines Distributed to Providers.<br>Based on number of Feedback Reports to doctors over some specified time period.                                                        |
| Type II:                 | ACE inhibitors filled (from claims and from interview).<br>Body weight and dietary salt intake (from interview)<br>Cognitive ability (from interview)                                                                                                                                                            |
| Type III:                | A) Health: Ejection Fraction (from provider)<br>B) Economic:<br>Total Claims Dollars over some specified time period (from administrative claims)<br>Total admits over some specified time period (from administrative claims)<br>C) Satisfaction: Patient Satisfaction, Provider Satisfaction (from interview). |

The authors' comments are presented as guidelines that one may consider when conducting these kinds of studies. The comments are not comprehensive in any way, they are only suggestions, and should not be construed as recommendations. We strongly advise consulting with experts before initiating any of these study designs.

### Key to terms:

- "Comparability": This is related to the sameness of "metric methods" used on each population. Were clinical and financial metrics measured in the same way, over the same time period? Were the "models used were appropriate to "adjust" for non-equivalence?
- "Equivalence": Related to the sameness of important " risk factors" for the outcome between the intervention and the reference population?
- "Internal validity": Where comparability and equivalence were both "high" in the study.
- "External validity": Related to internal validity and the degree to which the study population were "representative" of a larger population?

**Text Box 2: Post-Test**

	Study Design	Methods	Validity
I.1	High Risk: Post Test	<p>Description: Enroll participants and treat their illness.</p> <p>Patient Selection: Ad lib.</p> <p>Measurement: Could be very precise, however, virtually impossible to interpret the meaning of these metrics in the absence of a comparison population.</p> <p>Comparability: Not applicable.</p> <p>Equivalence: Not applicable.</p> <p>General: If there is an "opt-out" or "opt-in" provision, selection or "participation" bias may be an issue. Almost impossible to interpret findings. If selection is based on "above" a cut-off point, this design is very susceptible to regression-to-the-mean.</p>	<p>Internal Validity: Extremely low.</p> <p>External Validity: Extremely low.</p>
I.2	Population: Post Test	<p>Description: Enroll participants and treat their illness.</p> <p>Patient Selection: Ad lib.</p> <p>Measurement: Could be very precise, however, virtually impossible to interpret the meaning of these metrics in the absence of a comparison population.</p> <p>Comparability: Not applicable.</p> <p>Equivalence: Not applicable.</p> <p>General: If there is an "opt-out" or "opt-in" provision, selection or "participation" bias may be an issue. Almost impossible to interpret findings. . If selection is based on "above" a cut-off point, in many circumstances, this design is less susceptible to regression-to-the-mean than design I.1.</p>	<p>Internal Validity: Very low.</p> <p>External Validity: Very low.</p>

**Text Box 3: Benchmark Studies**

	Study Design	Methods	Validity
II.1	National Benchmark	<p>Description: This design could the same patient population as I.1 or I.2 except the results are compared to the national benchmark (e.g. ACE inhibitors use).</p> <p>Patient Selection: Ad lib.</p> <p>Measurement: Same as national benchmark</p> <p>Comparability: If the same measurement methods were used to measure the DM program, the two populations have a high degree of metric method comparability.</p> <p>Equivalence: If the same criteria to select patients for the DM program was exactly the same and was done in the same calendar time periods these two populations likely have a high degree of equivalence. However, other unmeasured risk factors such as cognitive impairment, severity of disease, tobacco use, etc. may make the two populations non-equivalent.</p>	<p>Internal Validity: Depends on the rigor of the measurements and equivalence of the DM population to the benchmark. If equivalence and comparability are high then internal validity is quite good.</p> <p>External Validity: NA. The only study with any measure of external validity is the one that produced the benchmark.</p>
II.2	Peer-Reviewed Article	<p>Description: The design used here was to select patients and follow them for 90 days to assess the rate of readmission using same intervention as Rich et al (56.2 percent reduction with the intervention: 91/142 in reference group and 75/140 in intervention group).<sup>34</sup></p> <p>Patient Selection: Scenario #1: Ad-lib selection. Scenario #2: Follow Rich’s article</p> <p>Measurement: Rate of readmission.</p> <p>Comparability: This was very high, as the measures were the same in both populations.</p> <p>Equivalence: In scenario #1 it is difficult to achieve equivalence. Unfortunately, it may be also very difficult in scenario #2 as Rich identified 1,306 and only enrolled 282 (see below). The existence of unconsidered and/or unmeasured risk factors may vary between the two populations.<sup>34</sup></p> <p>Note: “This study has several limitations, the first of which concerns the generalizability of the results. A total of 1306 patients fulfilled the criteria for a diagnosis of congestive heart failure, but only 282 (21.6 percent) were randomized. The distinguishing characteristics of the randomized cohort included advanced age (median, 79 years), a high prevalence of hypertension (75.9 percent), moderate functional impairment, and relatively well preserved left ventricular systolic function. The applicability of our findings to other patients with heart failure requires further study.</p>	<p>Internal Validity: Depends on the rigor of the measurements and equivalence of the DM population to the benchmark. If equivalence and comparability are high then internal validity is quite good.</p> <p>External Validity: NA. The only study with external validity is the one that produced the benchmark (see quote from study to the in the adjacent column).</p>

**Text Box 3: Benchmark Studies (continued)**

	Study Design	Methods	Validity
II.3	Statistical Prediction	<p>Description: Use existing software to “predict” costs of CHF population (including those with co-morbidities), enroll those predicted to be high cost, and compare “actual” to “predicted.”</p> <p>Patient Selection: Per predictive modeling algorithm.</p> <p>Measurement: Claims costs.</p> <p>Comparability: Unknown. Expected costs are derived from a linear regression model from another time period (and perhaps another place). Observed costs are calculated and subject to influences of factors that may or may not have influenced the model that predicted expected costs. Statistical model appropriateness depends on meeting the necessary assumptions of the model.</p> <p>Equivalence: Unknown. Expected costs are derived from a linear regression model from another time period (and perhaps another place). Observed costs are calculated and subject to influences of factors that may or may not have influenced the model that predicted the expected costs.</p>	<p>Internal Validity: Depends on the rigor of the measurements and equivalence of the DM population to the benchmark. Also varies by “fit” of model.</p> <p>External Validity: NA. The only study with external validity is the one that produced the benchmark.</p> <p>We would recommend validating the model with test of cross-classification (e.g. sensitivity, specificity, predictive value, etc.) and “residuals” in a reference population that is not exposed to DM.</p>

**Text Box 5: Ecological Studies**

	Study Design	Methods	Validity
IV.1	Ecological	<p>Description: Select patients in multiple plans, Score DM interventions on a linear scale of “dose” (using a panel to determine importance of DM sub-components to outcomes), compare to aggregate standardized criteria on a x-y graph.</p> <p>Patient Selection: All CHF patients by claims-based algorithm.</p> <p>Measurement: Per standard criteria for outcomes and per panel of experts to score strength of DM sub-components.</p> <p>Comparability: High on both intervention score and outcome score.</p> <p>Equivalence: On criteria being measured it is very high, however, there may be unknown and/or unmeasured risk factors that influence the results.</p>	<p>Internal Validity: Low. No causal relationship is established.</p> <p>External Validity: Low.</p> <p>Since a “causal relationship” is not established, the external validity of these results is very weak. However, it could lead to a more rigorous study of the causal hypotheses.</p>

**Text Box 4: Quasi-Experimental Studies**

	Study Design	Methods	Validity
III.1	Pre-Post Test	<p>Description: Select patients eligible for the program using some standard algorithm (or referral); determine Type III cost metric on prior calendar year and calendar year after introduction of DM program. Assess difference.</p> <p>Patient Selection: All CHF patients by claims-based algorithm.</p> <p>Measurement: Compare cost metrics before and after.</p> <p>Comparability: Should be high as we are using the same methods.</p> <p>Equivalence: Unknown: The patient as his or own control. This assumes the level of cost in the prior year (or other Type III metric) would have been the level of cost in the post year in the absence of DM. This is an untested assumption.</p>	<p>Internal Validity: Low (without verification of equivalence).</p> <p>We recommend one or more equivalent external reference populations are measured pre-post that did not have the DM intervention.</p> <p>External Validity: Dependent on (a) internal validity, i.e. the equivalence of the pre-post population and (b) the equivalence of the "new" population to which one is generalizing.</p>
III.2	Time Series	<p>Description: Select patients eligible for the program using some standard algorithm (or referral); determine trend of Type III cost metric during prior calendar year(s) and trend during calendar year(s) after introduction of DM program. Assess difference.</p> <p>Patient Selection: All CHF patients by claims-based algorithm.</p> <p>Measurement: Compare cost metrics multiple times before and after.</p> <p>Comparability: Should be high as we are using the same methods.</p> <p>Equivalence: Unknown: The patient is his or own control. This assumes the level of cost in the prior year(s) (or other Type III metric) would have been the cost trends in the Post year(s) in the absence of DM. This is an untested assumption.</p>	<p>Internal Validity: Higher than pre-post because of multiple measures over time, but need verification of the equivalence assumption.</p> <p>We recommend one or more equivalent external reference populations are measured pre-post that did not have the DM intervention.</p> <p>External Validity: Dependent on (a) internal validity, i.e. the equivalence of the pre-post population and (b) the equivalence of the "new" population to which one is generalizing.</p>

**Text Box 4: Quasi-Experimental Studies (continued)**

	Study Design	Methods	Validity
III.3	Regression-Discontinuity	<p>Description: Select patients with ejection fractions below some cut-off point, only intervene on patients above the cut-off point. We then compare pre-post relationship in individuals in both groups. If the hypothesis is correct (and certain assumptions are true) a discontinuity in slope among those above the cut-off compared to those below the cut-off will be observed. Outcome Measures: Claims costs and quality of Life.</p> <p>Patient Selection: All CHF patients by claims-based algorithm that have ejection fraction data.</p> <p>Measurement: Fitting linear regression (slope and intercept) of the pre-post measure in patients below the cut-off to the linear regression line of pre-post measure above the cut-off.</p> <p>Comparability: Metrics should be high as we are using the same methods. The appropriateness of the statistical model depends on factors, such as the distribution of data.</p> <p>Equivalence: Is the pre-post in linear regression “fit” in those below the cut-off equivalent to the pre-post linear regression “fit” in those above the cut-off point in the absence of a DM intervention?</p>	<p>Internal Validity: Varies by “fit” of model. Could be validated by inclusion of multiple cut-off points or the test of the model in a population without DM.</p> <p>External Validity: This depends on internal validity and the equivalence of the “new” population to the study population.</p>

**Text Box 5: Ecological Studies**

	Study Design	Methods	Validity
IV.1	Ecological	<p>Description: Select patients in multiple plans, Score DM interventions on a linear scale of “dose” (using a panel to determine importance of DM sub-components to outcomes), compare to aggregate standardized criteria on a x-y graph.</p> <p>Patient Selection: All CHF patients by claims-based algorithm.</p> <p>Measurement: Per standard criteria for outcomes and per panel of experts to score strength of DM sub-components.</p> <p>Comparability: High on both intervention score and outcome score.</p> <p>Equivalence: On criteria being measured it is very high, however, there may be unknown and/or unmeasured risk factors that influence the results.</p>	<p>Internal Validity: Low. No causal relationship is established.</p> <p>External Validity: Low.</p> <p>Since a “causal relationship” is not established, the external validity of these results is very weak. However, it could lead to a more rigorous study of the causal hypotheses.</p>

**Text Box 6: Cross-Sectional Studies**

	Study Design	Methods	Validity
V.1.	One Sample	<p>Description: A DM program is established in the managed care organization. CHF patients should be randomly selected and quizzed about their participation in a DM program and to what extent. Should also ask about their cognitive ability, dietary salt intake, body weight, and other outcomes.</p> <p>Patient Selection: Random selection of all CHF patients.</p> <p>Measurement: Analyze the cross-tabulation of DM participation and at "dose" of DM to outcomes to assess "association."</p> <p>Comparability: High</p> <p>Equivalence: On patient identification criteria quite high, however, the participants in DM could have different severity levels than others.</p>	<p>Internal Validity: Causal inferences are weak due to potential temporal ordering of DM and outcomes. That is did a DM intervention come before the outcome or after? Unknown without a better design.</p> <p>External Validity: Unless temporal ambiguity is addressed, the external validity is weak.</p>
V.2.	More than one sample	<p>Description: Could choose two areas of a catchment area, one that has a strong DM program and one that has a weak or non-existent DM program. Take a random sample of a defined population and compare Type I to Type II and III metrics.</p> <p>Patient Selection: Random selection of all CHF patients</p> <p>Measurement: Analyze the cross-tabulation of DM participation and at "dose" of DM to outcomes to assess "association."</p> <p>Comparability: High</p> <p>Equivalence: On patient identification criteria quite high, however, the participants in DM could have different severity levels than others.</p>	<p>Internal Validity: This could be stronger because it could be considered a replication in another population. Bradford-Hill criteria #2.</p>

**Text Box 7: Case-Control Studies**

	Study Design	Methods	Validity
VI.1	Case Control	<p>Description: We defined "cases" as CHF patients who were recently discharged from the hospital and "controls" as age-sex ejection fraction matched CHF patients who were not recently discharged from the hospital. The concern was that a DM program that was based on self-empowerment would not be effective in patients with cognitive impairment.</p> <p>Patient Selection: Based on age-sex-ejection fraction and outcome status (i.e. recent hospital discharge)</p> <p>Measurement: We test the hypothesis that DM participation in a program is confounded by cognitive ability.</p> <p>Comparability: Very high</p> <p>Equivalence: Very high (unless unmeasured risk factors, e.g. depression, influenced the results)</p>	<p>Internal Validity: High, assuming controls are an equivalent population to the cases.</p> <p>External Validity: High if "new" population is equivalent to the study population and study is replicated in other settings.</p>

**Text Box 8: Follow-up Studies**

	Study Design	Methods	Validity
VII.1.i	Observational Cohort Design	<p>Description: Among all CHF patients discharged from hospital one time in the last six months and are continuously enrolled. These patients are followed for 90 days to assess incidence of new admissions. We hypothesis that exposure to DM will lower the risk of the outcome, we also hypothesize that cognitive function will have an independent effect on this relationship. (An improvement would be to compare the internal cohort to an equivalent external reference group)</p> <p>Patient Selection: All patients with one hospital discharge. The reference group design was those with a low level of exposure to DM, the intervention group were those with a high exposure to DM (this exposure was based on a non-structured roll out by market). (An improvement would be to compare the internal cohort also to an equivalent external reference cohort)</p> <p>Measurement: Would stratify for cognitive function x DM participation vs. no DM participation. Adjust for potential non-equivalency, i.e. confounding (using an experienced statistician and/or epidemiologist).</p> <p>Comparability: High</p> <p>Equivalence: Unknown, will have to assess which risk factor would impact the outcome (e.g. level of depression or anti-depression medication) and attempt to adjust, using stratified analysis or some other form of multivariate modeling (In the enhanced version, we could compare independent risk factors in the internal cohort to the external cohort)</p>	<p>Internal Validity: High. Temporal ordering is excellent as the exposure to DM comes before the outcome (readmission).</p> <p>External Validity: High, if replicated.</p>
VII.2.i	Field Randomized Control Trial (by Place)	<p>Description: ID all CHF patients and rollout DM program in 4 different hospitals with random allocation to measure multiple Type II and Type III metrics.</p> <p>Patient Selection: All CHF patients by claims-based algorithm.</p> <p>Measurement: Standardized. Measure impact by groups.</p> <p>Comparability: High because all metrics are measured in the same way.</p> <p>Equivalence: Unknown without further inquiry. Social class, for example, and other risk factors could vary between patient populations from different hospital catchment areas.</p>	<p>Internal validity: Medium.</p> <p>External validity: Medium. Need to replicate in multiple settings.</p>

**Text Box 8: Follow-up Studies (continued)**

	Study Design	Methods	Validity
VII.2.ii	Field Randomized Control Trial (by Time)	<p>Description: ID all CHF patients and rollout DM program over 4 quarters with random allocation to measure multiple Type II and Type III metrics.</p> <p>Patient Selection: All CHF patients by claims-based algorithm.</p> <p>Measurement: Standardized. Measured subset impacted by DM in first quarter to everyone else. After 4 quarters all patients are enrolled. All four groups are followed for the 2<sup>nd</sup> year. In the second year, we can test the hypothesis is that the longer one is in the program, the greater the impact on Type II and Type III metrics.</p> <p>Comparability: High.</p> <p>Equivalence: Medium to High (unless unmeasured risk factors are distributed unevenly between the four groups or individuals react to the random assignment and it changes their behavior re outcomes.</p>	<p>Internal validity: Medium to High, but had to measure effects that occur following the first year.</p> <p>External validity: Could be high if replicated.</p>
VII.3.i	Randomized Control Trial: (Individual, not-blinded)	<p>Description: ID 400 CHF patients and randomly allocate to the intervention or reference population. The DM intervention to be tested is a mailed reminder to stay alert to the need to monitor risk factors, to adjust medication as needed, to call DM nurse if there is a problem. One half of the patients would receive it, the other half would not.</p> <p>Patient Selection: All CHF patients by claims-based algorithm.</p> <p>Measurement: Standardized. 1) Monitor calls to DM nurses by Group. 2) Measure Rx fill rate by groups using Rx claims data.</p> <p>Comparability: High.</p> <p>Equivalence: Assuming other independent risk factors are also randomized (this would need to be checked on some variables), this could be very high. However, if patients are aware of the group assignment and change behavior because of this awareness, equivalency in follow-up period could be compromised. Could also be a problem if nurse managers are aware of group assignment and alter their message based upon that awareness.</p>	<p>Internal validity: Medium to High.</p> <p>External validity: Medium to High. Need to replicate in multiple settings.</p>
VII.3.ii	Randomized Control Trial (Individual Blinded)	<p>Description: By DMAA definition, a disease management intervention requires patient-empowerment strategies; therefore, this kind of study design would be virtually impossible.</p>	<p>Highest internal validity, but very unlikely to occur in DM</p> <p>External validity: Would depend on the study population.</p>

**TABLE 1: Summary of Study Designs.**

Code	Description	Reference Population				Random Allocation		Unit of Measurement	Example in Literature
		Patient as own Control	DM Exposure Variation		Non-Exposure to DM	Benchmark	None: Observational	Yes: G=Group I=Individual	
I	Post test								35
I.1	"High Risk"	***					NA		
I.2	"Population"	***					NA		
II	Benchmark								
II.1	National					X	X		36
II.2	Peer-reviewed					X	X		34, 37
II.3	Statistical					X	X		8
III	Quasi-Experimental								38
III.1	Pre-Post	X					X		39,40
III.2	Time-Series	X					X		
III.3	Regression-Discontinuity				X		X		41
IV	Ecological								42
IV.1	Ecological		X		Optional		X		
V	Cross-Sectional								42
V.1	Cross-Sectional (1)		X				X		
V.2	Cross-Sectional (>1)		*		*		X		
VI	Case-Control								42
VI.1	Case-Control		**		**		X		
VII	Follow-up								42
VI.1.i	Observational Cohort	***	***		X		X		43, 44.
VI.2.i	Field RCT: Place	***	***		X			G	45
VI.2.ii	Field RCT: Time	***	***		X			G	46
VI.3.i	RCT: Not blinded	***	***		X			I	47, 1
VI.3.ii	RCT Blinded	***	***		X			I	

**KEY:**

X=Necessary component

NA=not applicable.

\* Special situation where a sample (preferably a random sample) is selected from a defined population and exposure and outcome status are determined.

\*\* Special situation where "cases" are selected as they have already "experienced" an outcome and controls are selected from the at-risk population that have not experienced the outcome (i.e. Type I and Type II metric). "Exposure" status is then determined in both "case" and "controls"

\*\*\* Could be done as an additional study criteria.



## REFERENCES

- <sup>1</sup>Krumholz HM, Amatruda J, Smith GL, Mattera JA, Roumanis SA, Radford MJ, Crombie P, Vaccarino V. Randomized trial of an education and support intervention to prevent readmission of patients with heart failure. *J Am Coll Cardiol* 2002; 29: 83-89.
- <sup>2</sup>Riegel B, Carlson B, Kopp Z, LePetri B, Glaser D, Unger A. Effect of a standardized nurse case-management telephone intervention on resource use in patients with chronic heart failure. *Arch Intern Med*. 2002; 162: 705-12.
- <sup>3</sup> Ofman JJ, Ryu S, Borenstein J, Kania S, Lee J, Grogg A, Farup C, Weingarten, S. Identifying patients with gastroesophageal reflux disease in a managed care organization. *Am J Health Syst Pharm*. 2001; 58: 1607-13.
- <sup>4</sup> Disease Management Association of America. [www.dmaa.org](http://www.dmaa.org).
- <sup>5</sup> Nelson EC, Splaine ME, Batalden PB, Plume SK. Building measurement and data collection into medical practice. *Ann Intern Med* 1998; 128:460-6.
- <sup>6</sup> Berkman LF, Kawachi I (eds) *Social Epidemiology*. New York: Oxford University Press, 2000.
- <sup>7</sup> Ash AS, Zhao Y, Ellis RP, Kramer MS. Finding Future High-Cost Cases: Comparing Prior Cost Versus Diagnosis-Based Methods. *HSR: Health Services Research*. 2001; 36: 194 – 206.
- <sup>8</sup> Cumming RB, Knutson D, Cameron BA, Derrick B. A Comparative Analysis of Claims-Based Methods of Health Risk Assessments for Commercial Populations. Society of Actuaries, 2002 ([www.soa.org](http://www.soa.org)) .
- <sup>9</sup> Brody KK, Johnson RE, Ried DL, Carder PC, Perrin N. A comparison of two methods for identifying frail Medicare-aged persons. *J Am Geriatr Soc*. 2002; 50: 562-9.
- <sup>10</sup> Pacala JT, Boulton C, Reed RL, Aliberti E. Predictive validity of the Pra instrument among older recipients of managed care. *J Am Geriatr Soc*. 1997;45 614-7.
- <sup>11</sup> Wilson TW. *Epidemiology of Value: How to Balance Clinical and Economic Value in Population Health Value Programs*. Wilson Research, LLC. Loveland, Ohio 2001.
- <sup>12</sup> Wilson TW. How I Learned to Stop Worrying and Love Regression-to-the-Mean: Case Studies from Population-Based Disease Management Programs. American Association of Health Plans, Building Bridges Conference VIII, Long Beach, California. April 10, 2002.
- <sup>13</sup> Porter R. *The Greatest Benefit to Mankind: A Medical History of Humanity*. NY: W.W. Norton & Co, 1997.
- <sup>14</sup> Schneider EC, Riehl V, Courte-Wienecke S, Eddy DM, Sennett C. Enhancing performance measurement: NCQA's road map for a health information framework. National Committee for Quality Assurance. *JAMA* 1999; 282:1184-90.
- <sup>15</sup> Whellan DJ, Gaudin L, Gattis WA, Granger B, Russell SD, Blazing MA, Cuffe MS, O'Connor CM. The benefit of implementing a heart failure disease management program *Arch Intern Med*. 2001; 161: 2223-8.
- <sup>16</sup> WHO/UNICEF: *Primary Care*. Geneva: World Health Organization, 1978.
- <sup>17</sup> Ware JE, Kosinski M. Interpreting SF-36 summary health measures: a response. *Qual Life Res* 2001;10: 405-13.
- <sup>18</sup> Heudebert GR, Centor RM, Klapow JC, Marks R, Johnson L, Wilcox CM. What is heartburn worth? A cost-utility analysis of management strategies. *J Gen Intern Med*. 2000; 15: 175-82.
- <sup>19</sup> Kominski GF, Simon PA, Ho A, Luck J, Lim YW, Fielding JE. Assessing the burden of disease and injury in Los Angeles County using disability-adjusted life years. *Public Health Rep* 2002; 117:185-91.

- <sup>20</sup>Gold MR, Siegel JE, Russell LB, Weinstein MC. Weinstein MC. Cost-Effectiveness in Health and Medicine. NY: Oxford University Press, 1996.
- <sup>21</sup>Drummond MF, O'Brien M, Stoddard GL, Torrance GW. Methods for the Economic Evaluation of Health Care Programs. 2<sup>nd</sup> edition. Oxford University Press, Oxford, UK. 1997.
- <sup>22</sup>Engelgau MM, Narayan KM, Herman WH. Screening for type 2 diabetes. *Diabetes Care* 2000; 23:1563-80.
- <sup>23</sup> Paltiel AD, Fuhlbrigge AL, Kitch BT, Liljas B, Weiss ST, Neumann PJ, Kuntz KM. Cost-effectiveness of inhaled corticosteroids in adults with mild-to-moderate asthma: results from the asthma policy model. *J Allergy Clin Immunol* 2001; 108:39-49.
- <sup>24</sup>Parkin D, Jacoby A, McNamee P, Miller P, Thomas S, Bates D. Treatment of multiple sclerosis with interferon beta: an appraisal of cost-effectiveness and quality of life. *J Neurol Neurosurg Psychiatry* 2000;68:144-9.
- <sup>25</sup>Algozzine T, Pannone R, Kozma CM. Opinions of disease management programs among medical directors of managed care organizations. *Am J Health Syst Pharm.* 1988; 55: 1029-33.
- <sup>26</sup> American Diabetes Association. Clinical Practice Recommendations 2002. *Diabetes Care* 25:S3, (entire issue) 2002.
- <sup>27</sup> National Asthma Education and Prevention Program. Expert Panel Report 2: Guidelines for the Diagnosis and Management of Asthma. NIH Publication No 97-4051, Bethesda, Maryland 1997.
- <sup>28</sup> Leape LL, Kabacoff AI, Gandhi TK, Carver P, Nolan TW, Berwick DM. Reducing adverse drug events: lessons from a breakthrough series collaborative. *Jt Comm J Qual Improv* 2000; 26:321-31.
- <sup>29</sup> Kleinbaum DG, Kupper LL, Muller KE. Applied Regression Analysis and Other Multivariable Models. 2<sup>nd</sup> Ed. Boston MA: PWS-Kent, 1988.
- <sup>30</sup> Begg C, Cho M, Eastwood S, Horton R, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA.* 1996; 276:637-639. ([www.consort-statement.org](http://www.consort-statement.org)).
- <sup>31</sup> Devereaux PJ, Manns BJ, Ghali WA, Quan H, Guyatt GH. The reporting of methodological factors in randomized controlled trials and the association with a journal policy to promote adherence to the Consolidated Standards of Reporting Trials (CONSORT) checklist. *Control Clin Trials* 2002 Aug;23(4):380-8.
- <sup>32</sup> Jekel JF, Katz DL, Elmore JG. Epidemiology, Biostatistics, and Preventive Medicine. 2<sup>nd</sup> edition. Philadelphia: Harcourt Health Services, 2001.
- <sup>33</sup> Rose G. Sick individuals and sick populations. *Int J Epidemiol.* 1985; 14: 32-38.
- <sup>34</sup> Rich MW, Beckham V, Wittenberg C, Leven CL, Freedland KE, Carney RM. A multidisciplinary intervention to prevent the readmission of elderly patients with congestive heart failure. *N Engl J Med* 1995; 333:1190-5.
- <sup>35</sup> Cook DT, Campbell DT. Quasi-experimental Designs for Research & Analysis Issues for Field Settings. Chicago: Rand McNally, 1979.
- <sup>36</sup> Knox D, Mischke L. Implementing a congestive heart failure disease management program to decrease length of stay and cost. *J Cardiovasc Nurs* 1999; 14: 55-74.
- <sup>37</sup> Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med* 1993; 329:977-986.
- <sup>38</sup> Cook DT, Campbell DT. Quasi-experimental Designs for Research & Analysis Issues for Field Settings. Chicago: Rand McNally, 1979.

- <sup>39</sup> Whellan DJ, Gaulden L, Gattis WA, Granger B, Russell SD, Blazing MA, Cuffe MS, O'Connor CM. The benefit of implementing a heart failure disease management program. *Arch Intern Med* 2001; 161: 2223-8.
- <sup>40</sup> Rubin RJ, Dietrich KA, Hawk AD. Clinical and economic impact of implementing a comprehensive diabetes management program in managed care. *J Clin Endocrinol Metab* 1998; 83: 2635-42.
- <sup>41</sup> Trochim WMK. *Research design for program evaluation: The regression-discontinuity approach*. Beverly Hills, CA: Sage, 1984.
- <sup>42</sup> Rothman K, Greenland S (eds.) *Modern Epidemiology*. 2<sup>nd</sup> ed. Philadelphia. Lippincott-Raven, 1998.
- <sup>43</sup> Sidorov J, Shull R, Tomcavage J, Girolami S, Lawton N, Harris R. Does diabetes disease management save money and improve outcomes? A report of simultaneous short-term savings and quality improvement associated with a health maintenance organization-sponsored disease management program among patients fulfilling health employer data and information set criteria. *Diabetes Care*. 2002; 25: 684-9.
- <sup>44</sup> Selby JV, Ray GT, Zhang D, Colby CJ. Excess costs of medical care for patients with diabetes in a managed care population. *Diabetes Care* 1997; 20:1396-402.
- <sup>45</sup> Philbin EF, Rocco TA, Lindenmuth NW, Ulrich K, McCall M, Jenkins PL. The results of a randomized trial of a quality improvement; intervention in the care of patients with heart failure. *Am J Med* 2000 109: 443-9.
- <sup>46</sup> Lorig KR, Ritter P, Stewart AL, Sobel DS, Brown B W Jr, Bandura A, Gonzalez VM, Laurent DD, Holman HR. Chronic disease self-management program: 2-year health status and health care utilization outcomes. *Med Care* 2001; 39: 1217-23.
- <sup>47</sup> McAlister FA, Lawson FM, Teo KK, Armstrong PW. Randomised trials of secondary prevention programmes in coronary heart disease: systematic review. *BMJ* 2001; 323: 957-62.
- <sup>48</sup> Sackett DL. Rules of evidence and clinical recommendations on use of antithrombotic agents. *Chest* 1986 89 (2 suppl.):2S-3S.
- <sup>49</sup> Bradford-Hill, Austin *The Environment and Disease: Association or Causation?* *Proc R Soc Med*. 1965; 58: 295-300.

