

## Commentary

# Disease Management Outcomes: Are We Asking the Right Questions Yet?

Gordon K. Norman, M.D., M.B.A.

**D**ESPITE COMPOUND ANNUAL GROWTH RATES of 30% to 40% for the disease management (DM) industry over the past decade, the question “Does disease management work?” has persisted. (The older term, “disease management,” and its abbreviation, “DM,” are used here to denote the wide array of health and DM programs currently referred to as the “population health improvement model” by DMAA: The Care Continuum Alliance.<sup>1</sup>) Raised to greater public awareness by a skeptical Congressional Budget Office report<sup>2</sup> in 2004, the debate was further fueled by a 2007 RAND report<sup>3</sup> that surveyed the literature and concluded that the evidence of savings from published DM studies is inconclusive—due partly to the paucity of published evidence, criticisms with study design and rigor, and a mixture of outcomes across different programs, conditions, and populations. Few question the ability of disease and care management programs to improve processes of care, adherence to evidence-based guidelines, and clinical outcomes; even the authors of the RAND review concede the evidence for these DM outcomes is consistently positive.

What remains in dispute is whether DM programs consistently produce return on investment (ROI), meaning short-term net savings for the health plans, self-insured employers, and public sector sponsors who typically pay for these services. The Centers for Medicare and Medicaid Services (CMS) recently made national headlines<sup>4</sup> and prompted congressional concern<sup>5</sup> over the decision to curtail the Medicare Health Support (MHS) program for fee-for-service Medicare beneficiaries<sup>6</sup> based on a preliminary analysis that showed no cost savings for the first 6 months of the project.<sup>7</sup> This action—taken mid-stream during a 3-year, randomized, controlled trial (RCT) designed to evaluate financial and other outcomes<sup>8</sup>—appears to reflect a rush to judgment by CMS. Most DM practitioners appreciate that outcomes take time to achieve, particularly for a population of older and sicker patients, and so would have expected exactly what the early MHS evaluation described. This puzzling action

prompted one seasoned observer to pose, “Is CMS the enemy of disease management?”<sup>9</sup>

While financial results for DM programs are important, they are not the only outcomes with which we should be concerned. It is time we start asking the right questions about DM outcomes and seeking answers in ways that reflect a much deeper understanding than just whether DM “works” to save money. Ultimately, there are fundamental things we want to know about any DM program:

- Is it beneficial? What are the specific benefits? For whom, when, and why?
- Are the benefits predictable and durable? Are some early, some later?
- Are the benefits replicable across different contexts, populations, and time?
- Can these benefits be achieved cost-effectively? Where, when, and how?
- Are cost savings possible? How much? When are savings likely? When not?
- What about the intervention seems most directly related to benefits? Is it more about who, what, how, or when it is delivered?
- What contextual factors influence the likelihood of benefits?
- What can we do better to improve these benefits?

These questions do not seem profound or even complicated, so why should they be so challenging to answer? The answer lies partly in the complexity and diversity of these programs, and partly in the different paradigms used by academia and business for evaluating programs and their respective standards for “proof.” Academic researchers and policy makers usually seek proof “beyond a reasonable doubt,” analogous to the criminal standard of proof with strong internal validity; nothing less than rigorous experimental designs are satisfactory for causal proof with strong internal validity, hence the strong pref-

erence for RCTs for evaluating DM programs. In contrast, private employers, some public sector entities, and most health plans apply a “preponderance of evidence” threshold, analogous to the civil standard of proof; quasi-experimental or careful nonexperimental actuarial methods may suffice to compute outcomes with sufficient credibility for their purposes.

Another important difference is the conclusion scope and time frame that different entities apply when evaluating DM outcomes. Policy makers strive to make generalizable assessments based on all the available evidence about DM programs, as they use these judgments to guide policy decisions on the assumption that future performance is best predicted by retrospective evidence of past performance. In essence, they are addressing the complex question, “Do (and will) DM programs always work?” By contrast, private health plans or employers are interested in favorable results for their own experience rather than assurance that all users of similar programs achieve comparable results. (From a competitive perspective, they may actually wish the opposite.) When local analysts, actuaries, and/or chief financial officers have evaluated their own DM programs and found that cost, utilization, quality, and satisfaction measures all correlate in a favorable direction, they are generally satisfied they have answered the question “Do my DM programs work?” in the affirmative. This is not a subtle distinction if one believes that differing DM programs for different conditions applied to different populations with different interventions will necessarily generate different outcomes.

If substantial heterogeneity exists in the makeup of DM programs and the outcomes from these programs, how are we to know what is best under what circumstances? Are RCTs a panacea for resolving the debate, “Does DM work?” Long-time health care executive Scott MacStravic, PhD, recently pointed out the limitations of scientific studies in assessing DM outcomes.<sup>10</sup> Excerpted from his World Health Care Blog entry from April 14, 2008 is the following:

The same error in logic keeps turning up in evaluations of proactive health management (PHM) efforts, with recent examples relative to both disease management and prevention. It is the attempt to reach an overall conclusion about PHM in general, as well as its various components, as if each is a uniform “solution” to a single “problem.” Since this is simply not the case, all such efforts are doomed to failure from the start, but manage to capture headlines when they are reached, nevertheless.

It is an unfortunate reality of the model for evaluation used in such cases that its very rigor in stipulating what is “scientific” works to limit the probability of success in what is evaluated. The best way to succeed in either DM specifically, or prevention in general, is to treat both as a “marketing” challenge—identify which people are the best prospective “customers” for using the proposed “product,” specifically in terms of their potential for success. Then, customize the intervention to match both what such

prospects are most likely to “buy” and what is most likely to deliver a positive return on the investment involved.

Such a matching process, ie, making the intervention fit the individual prospect—in terms of both likelihood of their “buying” or engaging in it, and probability of doing so realizing as much as possible of their individual potential for delivering savings—would optimize the return on investment. Unfortunately, it would also violate the “rules” of scientific evaluation from beginning to end. It would not provide a single form of the intervention, but one that varies by individual, and not to a randomly assigned group, but one composed of people who were “self-selected” to different interventions by design.

Even the frequent reports of disappointing or equivocal results from DM and prevention include examples of particular interventions that do work, along with those that do not. It is only the overall picture that is reported, whereas it would make more sense to be delighted by even a few examples that work, in order to determine how to improve interventions, rather than condemn all with the same brush. If the same logic were extended to identifying which are the best prospects, and customizing interventions, the overall picture would probably be considerably more positive.

Moreover, we would learn more about what does work, instead of denying ourselves potential gains because a majority of science-restricted interventions do not. Presumably the idea is to gain success, in reducing sickness care use and expense, improving the quality of life of consumers, saving money and improving performance for employers. The aim for success should be the dominant concern when planning, implementing, and evaluating DM and preventive interventions, not scientific restrictions that reduce the chances of such success.

MacStravic’s observations, while perhaps unorthodox to some, are neither radical nor Luddite in nature. They simply point out that methods commonly used for scientific proof in medicine have limitations when applied to complex, multifactorial interventions that are more rooted in epidemiology, quality improvement, and social science than biology or medicine. His insights echo similar themes raised by eminent health quality guru Don Berwick, M.D., in a recent *Journal of the American Medical Association* commentary on “the science of improvement.”<sup>11</sup> DM is often characterized as the application of quality improvement principles to population health. Berwick is critical that slavish adherence to principles of evidence-based medicine may not be in our best interest when it comes to improving patient safety and quality of care. In his editorial, Berwick points out why RCTs may be limiting in the quality improvement realm and makes a plea to reevaluate the design and evaluation of DM programs for better learning.

Berwick also cites the work of Pawson and Tilley in *Realistic Evaluation*,<sup>12</sup> which presents the view that for evaluating complex programs (social programs in particular), RCTs seeking to disprove the null hypothesis with high internal

validity based on average measures are less useful than what they label the “realistic evaluation” approach that seeks to understand what works for whom under what circumstances, and places equal emphasis on external validity, generalizability, and cumulative learning from evaluating programs. DM is very much a social intervention, the benefits of which will vary by program, by population, by circumstance, and by individual—how to design evaluations that discover what works best for whom and under what conditions is our primary challenge for learning, even as purchasers and policy makers continue to stress short-term ROI “proofs” to justify their investments. Again, classic experimental designs seem ill suited to both tasks, and perhaps less valuable for either one than once expected.

As a physician, quality improvement practitioner, and DM proponent, I have long regarded the question, “Does DM work?” to be as nonsensical as questions such as, “Does medication work?” or “Does surgery work?” The answer to each is, “It all depends.” The right medicine taken for the right condition in the right dosage for the right time period surely can improve health (though saving money is highly unlikely). Surgery performed by the right surgeon for the right indication at the right time for the right patient can yield better health (but rarely cost savings).

More relevant and interesting questions are: “For whom does DM produce what outcomes?” “For what conditions?” “Under what circumstances?” DM strategies are not like a pill or a procedure; they are not uniform, standardized health interventions, but rather a complex series of tools and interactions aimed at influencing health behaviors, often personalized down to the individual level, to achieve maximal impact for each participant. Obviously, people with chronic conditions or serious health risks are as heterogeneous a group of experimental subjects as one can find, so differing responses to any intervention would be expected, if not guaranteed.

A strong case can be made that measurement for learning and measurement for judgment, at least for DM, require different perspectives. How we reconcile this with academicians’ skepticism, as well as industry efforts to build consensus on pragmatic approaches to evaluating DM, poses a significant challenge. With due respect to the classical hierarchy of evidence used in proving new medical science, it is not clear that the application of the most rigorous of these methods will resolve the ongoing debate; it seems incumbent upon the industry to point out the logical fallacy that only RCTs can prove DM value, while providing a road map for better methods that promote industry learning and quality improvement.

In most cases, DM does not introduce new science. Rather, it strives to narrow the gaps between actual practice and evidence-based medicine by applying best practices to chronic condition management and health improvement. Various pundits have offered the observation that *doing what we know better* often trumps *knowing better what to do* in terms of value to society.<sup>13</sup> Despite this, the National Institutes of Health budget for science discovery is greater than 88-fold of that for the Agency for Healthcare Research and Quality to evaluate the delivery of health services. And what CMS spends on research, demonstration, and evaluation is less than 1.5% of the amount it spends for operations (largely claims pay-

ments). It seems our national research priorities are skewed away from learning how best to use existing knowledge of care delivery. We also are relatively unsophisticated in our knowledge of how best to measure health delivery programs.

Most of what we spend in the name of health care is intended to produce a health benefit outcome—longer life, better quality of life, or both. Despite conventional wisdom and recent political rhetoric to the contrary, most prevention and routine health interventions are not cost saving,<sup>14</sup> yet they are widely perceived as beneficial and worthwhile because they improve health in a cost-effective fashion. Cost saving was not the original intent of the DM pioneers in the early 1990s, either; rather, their goal was to apply quality improvement principles to see if patients at risk for avoidable morbidity could lessen those risks by close adherence to evidence-based medicine, better compliance with physician treatment plans and prescribed medications, and more healthful lifestyle behaviors. It just so happened that when targeting certain high-risk, high-cost populations, these interventions not only increased quality of life, but the avoided morbidity costs led to a net reduction in health care cost for payers. Thus was born the expectation for DM ROI, which continues to the present.

Better value for health spending is what we seek, and what our fragmented health care non-system is lacking. Assuming the common representation of value = quality/cost, value is improved whenever we raise quality for the same cost, raise quality and reduce cost simultaneously, or reduce cost for the same quality. Even when DM does not produce net savings for a given circumstance, is it likely that it still achieves notable health quality benefits with attractive cost-effectiveness relative to other health interventions? This is not a question I hear posed by academics or policy makers often, which is surprising, given the preponderance of evidence that demonstrates DM programs increase health care value by raising quality. I suggest that DM should not be held to a different standard than other common health care interventions. CMS covers new interventions shown to be effective without consideration of cost savings or even cost-effectiveness. The FDA approves new medicines on the basis of safety and effectiveness, not cost or cost-effectiveness. Yet resources are limited, and for a society facing health care costs that now consume more than 16% of the gross domestic product, we cannot ignore cost. It is appropriate to evaluate DM cost-effectiveness; in most cases I believe it will be shown to be highly cost-effective, and under many circumstances it will produce cost savings. That is more than can be said for the vast majority of health interventions for which we now pay without hesitation.

Thanks to Wennberg’s prolific research, we have known for more than 35 years that profound and unwarranted variation in the care of patients is ubiquitous in this country.<sup>15</sup> We have been shown repeatedly by McGlynn<sup>16</sup> and others that when it comes to the consistent delivery of health care services known to be beneficial, we are just plain lousy. On average, Americans receive about half of recommended care processes, and the gap between what we know works and what actually is done is substantial and troubling. In their latest Dartmouth Atlas,<sup>17</sup> which fo-

cuses on the care chronically ill seniors receive in the last 2 years of life, Wennberg and team have demonstrated yet again that rampant unwarranted variation in care also pervades that segment of the population. Among their recommendations for addressing this problem, the authors suggest care management systems, such as disease registries and DM protocols. Given this dismal status quo and good, albeit imperfect, evidence that DM can raise care quality by closing these gaps in evidence-based care, we must be thoughtful and deliberate about separating the DM “baby” from the chronic care “bathwater.” We can ill afford to be cavalier about evaluating such programs or drawing casual inferences, given the promise these approaches hold for efficiently improving population health.

So back to the original question, “Does DM work?” Averaging the successes and failures of different programs for diverse conditions over heterogeneous populations to answer that question seems unhelpful, if not misleading. More important knowledge comes from answering the question, “Does DM ever work?” If DM savings result from any programs, then the circumstances of those programs can be scrutinized to develop a deeper understanding of why, for population X, the specific interventions led to changes that resulted in net savings. Then, by examining more instances of such specific linkages between interventions and their causative mechanisms, applied to specific contexts and their resulting outcomes, we can develop better insight about how to approach novel situations to replicate these outcomes. For social program evaluators, this paradigm is quite familiar and routine, but it is a nuanced and fundamentally different way of viewing the evaluation conundrum for those who engage in the DM ROI debate.

Questions that challenge the DM industry and preoccupy those working across the industry for answers through DMAA: The Care Continuum Alliance include:

- How do we reconcile the demands for more rigorous DM evaluations with the methods we have developed in our consensus Outcomes Guidelines editions that attempt a pragmatic balance between suitability and acceptability?
- If a number of key industry stakeholders continue to maintain that only RCTs constitute compelling evidence for causation, to what extent should the industry be willing to conduct more RCTs, despite the shortcomings described by Berwick, Pawson & Tilley, and others, and despite the disinterest of most current DM purchasers to fund such studies?
- Will more RCTs increase the probability of publication and dissemination of credible outcomes to help resolve the debate, or are they likely to perpetuate “nothing works consistently” conclusions by virtue of an inherent misalignment with the nature of DM as a social experiment with high contextual influence?
- When we cannot conduct RCTs, should we seek “realistic evaluation” approaches using pluralistic quasi-experimental methods as rigorously as we can, then strive for greater dissemination and discussion of those results with iterative learning going forward?
- Is it necessary to divide our thinking about measurement for judgment and measurement for learning? If

cumulative learning requires different evaluation approaches from what external stakeholders are expecting for “proof,” how do we best resolve this dilemma?

Turning the dialog to these more illuminating questions just may begin to generate more light than heat, unlike the present debate. It is time to start asking better questions in lieu of perpetuating the stalemate over “Does DM work?”

## References

1. DMAA: The Care Continuum Alliance. Advancing the population health improvement model. Available at: <[www.dmaa.org/phi\\_definition.asp](http://www.dmaa.org/phi_definition.asp)>. Last accessed June 26, 2008.
2. Holtz-Eakin D. An analysis of the literature on disease management programs, Congressional Budget Office report to Congress, October 13, 2004. Available at: <[www.cbo.gov/doc.cfm?index=5909&type=0](http://www.cbo.gov/doc.cfm?index=5909&type=0)>. Last accessed June 26, 2008.
3. Mattke S. Evidence for the effect of disease management: is \$1 billion a year a good investment? *Am J Manag Care*. 2007;13:670–676.
4. Abelson R. Medicare finds how hard it is to save money. *New York Times*. April 7, 2008. Available at: <[www.nytimes.com/2008/04/07/business/07medicare.html?pagewanted=1&r=1&hp](http://www.nytimes.com/2008/04/07/business/07medicare.html?pagewanted=1&r=1&hp)>. Last accessed June 26, 2008.
5. The Henry J. Kaiser Family Foundation, kaisernetwork.org. Senators ask CMS to reconsider decision to delay second phase of disease management pilot program. Available at: <[www.kaisernetwork.org/daily\\_reports/rep\\_index.cfm?DR\\_ID=51013](http://www.kaisernetwork.org/daily_reports/rep_index.cfm?DR_ID=51013)>. Last accessed June 26, 2008.
6. Centers for Medicare and Medicaid Services. Completion of phase I of Medicare health support program FAQs. Available at: <[www.cms.hhs.gov/CCIP/downloads/MH-SEOPexfaqsfm012808\\_FINAL.pdf](http://www.cms.hhs.gov/CCIP/downloads/MH-SEOPexfaqsfm012808_FINAL.pdf)>. Last accessed June 26, 2008.
7. McCall N, Cromwell J, Bernard S. Report to Congress: evaluation of phase I of Medicare health support (formerly voluntary chronic care improvement) pilot program under traditional fee-for-service Medicare. Available at: <[www.cms.hhs.gov/Reports/Downloads/McCall.pdf](http://www.cms.hhs.gov/Reports/Downloads/McCall.pdf)>. Last accessed June 26, 2008.
8. Centers for Medicare & Medicaid Services. Medicare Health Support. Available at: <[www.cms.hhs.gov/CCIP/](http://www.cms.hhs.gov/CCIP/)>. Last accessed June 26, 2008.
9. MacStravic S. Is CMS the enemy of disease management? World Health Care Blog. Available at: <[www.worldhealthcareblog.org/2008/02/24/is-cms-the-enemy-of-disease-management/](http://www.worldhealthcareblog.org/2008/02/24/is-cms-the-enemy-of-disease-management/)>. Last accessed June 26, 2008.
10. MacStravic S. Science vs. success in evaluating health management. World Health Care Blog. Available at: <[www.worldhealthcareblog.org/2008/04/14/science-vs-success-in-evaluating-health-management/](http://www.worldhealthcareblog.org/2008/04/14/science-vs-success-in-evaluating-health-management/)>. Last accessed June 26, 2008.
11. Berwick D. The science of improvement. *JAMA*. 2008;299:1182–1184.
12. Pawson R, Tilley N. *Realistic Evaluation*. London, England: SAGE Publications, Ltd; 1997.
13. Woolf SH. The break-even point: when medical advances are less important than improving the fidelity with which they are delivered. *Ann Fam Med*. 2005;3:545–552.
14. Cohen JT, Neumann PJ, Weinstein MC. Does preventive care save money? *N Engl J Med*. 2008;358:661–663.
15. Wennberg J, Gittelsohn A. Small area variation in health care delivery. *Science*. 1973;182:1102–1108.

16. McGlynn EA, Asche SM, Adams J, et al. The quality of health care delivered to adults in the United States. *N Engl J Med*. 2003;348:2635-2645.
17. Wennberg JE, Fisher ES, Goodman DC, Skinner JS. Tracking the care of patients with severe chronic illness. The Dartmouth Atlas of Health Care 2008. Available at: <[www.dartmouthatlas.org/atlas/2008\\_Chronic\\_Care\\_Atlas.pdf](http://www.dartmouthatlas.org/atlas/2008_Chronic_Care_Atlas.pdf)>. Last accessed on June 26, 2008.

Address reprint requests to:  
Gordon K. Norman, M.D., M.B.A.  
Chairman-Elect, DMAA: The Care Continuum Alliance  
700 Pennsylvania Ave. N.W.  
Suite 700  
Washington, D.C. 20004-2694  
E-mail: [cgraziano@dmaa.org](mailto:cgraziano@dmaa.org)