ACTEX

PUBLICATIONS

The Experts In Actuarial Career Advancement

Product Preview

For More Information: email Support@ActexMadRiver.com
or call 1(800) 282-2839

1 INTRODUCTION

This book covers the developing topic of care management interventions: design, management, and evaluation. Originally written for actuaries (the financial engineers of the healthcare payment system) I hope that it will be of value to anyone interested in these aspects of the management of healthcare intervention programs. The first edition addressed selected operational topics (such as the organizational structure and management of a disease management program) but its focus (appropriately for actuaries) was generally on cost, outcomes, and other financial issues. The topic of care management programs has expanded considerably since the publication of the first edition. In the 1990s and 2000s, private, commercial initiatives drove the industry, with government largely uninvolved. Government became more involved as the 2000s progressed with CMS sponsoring a large-scale test of disease management interventions (the Medicare Health Support initiative). Increasing disappointment with the results of some of the widely-implemented programs in the 2000s coincided with work of, among others, Don Berwick MD at the Institute for Health Improvement in Boston, which resulted in the coining of the well-known term "Triple Aim" of healthcare. Dr. Berwick was, however, not alone in developing innovative care management models. Many integrated health systems with access to complete medical records, such as the Mayo Clinic in Minnesota, the Geisinger Clinic in Danville, Pennsylvania and of course the Kaiser Permanente Medical Group became recognized as models for the movement toward co-ordinated care based on integration of data and systems. Many of the developments in both programs and financial incentives in the recent past have been introduced in an attempt to replicate within a non-integrated system the methods and infrastructure available within these integrated systems.

This book is for the most part analytical, objective and based on research. When the research began in 2003 that led to the first edition of this book, funded by the Society of Actuaries (SOA) and overseen by the SOA's Project Oversight Group, we had little idea of the scope and duration of the work that would ensue. Ten years is a long time in which to be engaged in a single project, albeit part-time and with the assistance of volunteers, co-authors, reviewers, and others and the publication of another book (about Risk Adjustment and Predictive Modeling) along the way. In total the SOA-sponsored study generated eight research papers. These papers, together with the addition of a number of topics that were not part of the original study formed part of the first edition of this book. Many practitioners, both actuaries and non-actuaries, have downloaded one or more of the original papers from the Society of Actuaries website, and have used some of the principles we developed in their own work. Some of the terms that we have coined in the course of the study, ("migration bias" for example) have found their way into day-to-day discussion of disease management (DM) outcomes. The popularity of the papers vindicates the Society of Actuaries Health Section's and the Committee on Knowledge Extension

Research's decisions to support the research, and confirms the increasing role that actuaries are playing in this new and exciting area of managed care.

Since the publication of the first edition, this area has seen an explosion in research, innovations in interventions, techniques and programs, and the first edition risked being out of date. Five years since the publication of the first edition, therefore, we have updated the original study with discussion of a number of new intervention programs that have been developed since the first edition, selected literature on programs previously discussed, new techniques (propensity matching and opportunity analysis, for example) and a more detailed discussion of other important topics, such as quality measurement and reporting.

While the subject area continues to develop ever-more rapidly, the focus of this edition (like the first) is to arm the reader with fundamental principles that can be applied within any type of population or intervention.

1.1 THE STATE OF THE UNION

In Chapter 5 we provide a detailed review of the literature on financial outcomes of different care management programs. Over the years that we have been engaged in this study, however, the world of Care Management interventions has not stood still. The history of evaluation of disease management outcomes is an example: it is interesting to consider what has been achieved and what has *not* been achieved in the last ten years. In 2004, the Disease Management Association of America[1] (DMAA) published "Principles for Assessing Disease Management Outcomes[2]." Far from establishing once and for all methodology and principles to be followed by practitioners, it is widely-agreed, including I believe by DMAA, that the guide fell short of the needs of the industry in this area. Accordingly, DMAA convened another work group in 2006 to tackle the subject again. The findings of this workgroup, entitled "DMAA Outcome Guidelines Report[3]" were published in December 2006. Because it is an industry consensus document, the DMAA workgroup report made a number of recommendations with which readers of this book may be familiar. In addition, the guidelines identified a number of potentially controversial issues, many of which were deferred for future consideration. Accordingly, DMAA convened a third series of work groups in 2007, which led to the publication of a second edition of Outcomes Guidelines[4] in the same year. The second edition addressed some of the gaps left by the first – for example, DMAA now recommends a particular method of selecting members for inclusion in a study population (which we discuss in greater detail in Chapter 11, and refer to as a re-qualification standard) when applying the adjusted historical control methodology, to overcome one of the more glaring areas of potential difference between comparison populations. The Guidelines also identify, but do not make recommendations for, the issue of minimum sample size for credible measurement. This and other topics are due to be covered in the 2008 volume of DMAA's guidelines. We have provided guidance in this book (see Chapter 14) that may assist users in this area.

[1] Now re-named Care Continuum Alliance, CCA.
[2] See Bibliography [62].
[3] See [154].
[4] See [155]

While its guidelines may help practitioners and purchasers, DMAA, as the industry trade association, was perceived by purchasers as representing an industry viewpoint, and at least somewhat suspect. The professional North American actuarial associations[5], Society of Actuaries, Canadian Institute of Actuaries, and American Academy of Actuaries, on the other hand, have a reputation for being objective. Recommendations from these professional actuarial bodies, therefore, will carry more weight, particularly given the increasing involvement of actuaries in the performance and review of studies. The American Academy of Actuaries released its paper "Disease Management Programs: What's the Cost?[6]" in 2005, and released a Practice Note for actuaries practicing in the field in early 2008. It is the nature of actuarial practice notes to be descriptive, rather than prescriptive, providing a compendium of acceptable approaches taken by actuaries in tackling a particular problem, rather than choosing a particular approach as the "best practice." Actuarial best practices in DM will be published in a Standard of Practice for DM, but, given the lack of maturity of actuarial practice in this area, publication is some years away. Since the publication of the Academy's practice note, the passage of the Affordable Care Act has diverted actuarial attention away from care management and evaluation issues. However, the introduction of new programs and reimbursement methodologies (some of which are discussed in Chapter 3) will require renewed focus on design and evaluation in the future.

Some peer-reviewed papers and other notable studies have been published since we began this study. Ariel Linden, a well-known researcher in this field, published a paper in 2006 that attracted considerable attention[7]. This paper addresses what the author calls "number needed to treat," and which may also be called (as we do in Chapter 6 of this book) the economics of care management. In addition, this paper draws attention to the need for identification of a causal relationship between any savings estimated or measured, and the underlying inpatient admission experience of the population (where the major portion of savings are to be found). Soeren Mattke, MD, and others from RAND published a paper with a provocative title: "Evidence for the Effect of Disease Management: Is $1 Billion a Year a Good Investment?[8]" The Congressional Budget Office and others published between 2008 and 2011 research into the results of the Medicare Health Support initiative. While these papers are covered in more detail in Chapter 5, the authors' conclusions will not come as a surprise to anyone who has read any of the literature, namely that there is some evidence that DM improves quality of care but that there is little reliable evidence of financial improvement. What remains puzzling is the absence of practical papers that examine the biases in measurement and the impact that these have on outcomes, as, for example, we have done in Chapter 12 of this study, and, for studies that fail to show successful outcomes, deeper analysis of what elements worked, did not work, and what could be changed.

Chapter 5 updates Chapter 4 of the previous edition (which is not reproduced here but is available in its entirety online at www.actexmadriver.com for any interested reader). Chapter 4 was written and published early in the life of the SOA project, and the revision takes ac-

[5] The Society of Actuaries mission is to provide education and research for North American Life, Health, Pensions and Investment Actuaries. The American Academy of Actuaries is the U.S. profession's interface with regulators, and is responsible for professional standards and accreditation. The Canadian Institute of Actuaries combines both educational and regulatory roles in Canada.
[6] See [44]
[7] See [119]
[8] See [130]

count (selectively) of newly-published articles as well as outcomes from studies of the newly-added programs discussed in Chapter 3. .

We have seen increased actuarial involvement in care management outcomes studies and audits since the publication of the original Society of Actuaries studies, including the inclusion of care management as a topic on the Society's health track Fellowship syllabus. The fundamental building blocks of studies – rigorous reconciliation of data and understanding of Per Member Per Month (PMPM) costs and trends for example – lend themselves to analysis by actuaries. We also suggest in Chapter 10 that a relatively new technique in the actuarial arsenal, but one gaining wide acceptance – risk adjustment – also has a role to play in ensuring equivalence between populations. This new edition contains a discussion of a related subject (Propensity Score Matching) that is used in health services research to generate comparable populations and which will be important in the future for actuaries working in this field.

To the extent that the prior edition of this book has helped educate actuaries and others about intervention program design, management and evaluation and equipped them to work with health services professionals and clinicians, the Society of Actuaries study will have made a contribution.

1.2 WHAT HAS CHANGED?

1.2.1 MEDICARE HEALTH SUPPORT

At the time of the publication of the first edition, the MHS program was in full swing. We looked forward to this program "finally provide(ing) the industry with the answers to two questions:

1. Does care management "work" (that is, produce a statistically-significant difference in financial and clinical results in the managed population)?
2. Potentially more important, how do the financial results measured by the randomized control methodology differ from results measured by a standard industry methodology (such as the actuarially-adjusted methodology described in Chapter 8)? While this comparative analysis is not part of the program, many researchers are anxiously awaiting the opportunity to perform just such a comparative analysis.[9]

The Medicare Health Support (MHS) program was introduced in 2005 under Section 721 of the Medicare Modernization Act of 2003 (MMA) (the same act that brought us Medicare Part D coverage for prescription drugs and expanded accessibility to health savings accounts). The act authorized development and testing of voluntary chronic care improvement programs, (later re-named Medicare Health Support) to improve the quality of care and life for people living with multiple chronic illnesses. This program applied to Medicare fee for service members with diabetes and/or heart failure. The Centers for Medicare and Medicaid Services (CMS) awarded eight different programs to disease managers in different regions. Three vendors subsequently withdrew from the program, and CMS reduced the savings target from [Fees Plus 5% of Total Chronic Claims] to just fees (break-even). These developments im-

[9] From Managing and Measuring Healthcare Intervention Programs, 1st edition.

plied that even during the program enrollment and savings targets were not being met, an impression later confirmed by outcomes evaluation.

As we discuss in more detail in Chapter 3, the MHS programs demonstrated some benefit in terms of improved quality of care but little by way of financial savings (although the conclusions were challenged by the DM industry based on assignment and lack of comparability of the patient populations). The lack of demonstrated success of the largely nurse call-center-based MHS model coincided with the increased interest in provider-based models, as we discuss later.

One may conclude, therefore that the MHS did provide an answer, at least to the first of our two questions. Perhaps less noticed is the fact that there is also an implicit answer to the second question: an intervention program that the industry accepted as being financially successful, based on a number of different evaluation methods, was demonstrated by a large randomized controlled trial as failing to provide a positive financial outcome. To our knowledge no study of the MHS program has ever compared the calculated outcomes based on more traditional evaluation methods and the randomized control method. Such a study, particularly if it were able to identify the sources of deviation between the randomized outcomes and those from other methods, would be valuable for identifying ways to improve our population studies (which will rarely be able to use a randomized method).

1.2.2 PLAUSIBILITY ANALYSIS

Practitioners have also contributed to advances in outcomes measurement, although the techniques have not been published in the peer-reviewed literature. Al Lewis, president of the Disease Management Purchasing Consortium International, recommends the use of what he calls "Plausibility Factors". These factors are not a substitute for a calculation of savings, but rather, a method for evaluating the reasonability of the published outcomes, and whether the calculated savings are possible ("plausible"), based on the underlying utilization of the population and what we know of the success of similar programs. Plausibility analysis requires the calculation of the following statistic (the plausibility factor) for the entire health plan:

$$\frac{\text{Disease-Specific Admissions/1000 (Program Year)}}{\text{Disease-Specific Admissions/1000 (Baseline Year)}}$$

USE OF PLAUSIBILITY FACTORS

The theory of plausibility factors is that they independently validate the measured financial results of a care management savings calculation, by demonstrating that actual utilization is reduced by the intervention, consistent with the financial measurement. Plausibility factors are generally utilization rates per 1,000 of the overall population for hospital admissions and emergency room visits for certain primary diagnoses. The primary diagnoses are: Diabetes, coronary artery disease, chronic obstructive pulmonary disorder, heart failure and asthma. The proposed interpretation of the Plausibility measures is that if the savings calculation results in positive savings but the utilization-based measures do not, the savings are not validated. Rather than reconciling the two contradictory results, the Plausibility factors are so dispositive that their results always trump any other outcomes calculation.

HOW VALID ARE UTILIZATION-BASED CALCULATIONS?

In order to be a valid test of the outcomes of a savings calculation, utilization-based measures must be calculated on the same basis as the savings. With plausibility factors this is not always the case. The plausibility factors may be a poor validator because:

1. In a population evaluation, the measurement population is carefully constructed to consist of members with sufficient eligibility to be enrolled and managed by the program and to exclude members and conditions that may confound the calculation. As calculated the plausibility factors bear only a tenuous relationship to the population being managed and measured. Their use implicitly assumes comparability between populations, but this comparability must be demonstrated and cannot be assumed.

2. Plausibility factors, because they apply to admissions and ER visits for primary diagnoses only, represent a very small percentage of all admissions and costs for chronic patients. For example, within a commercial population, these admissions and ER visits only account for 3% of the total claims costs for members with diabetes and the admissions only account for approximately 7 % of inpatient spend. Even a very successful program that avoided 25% of diabetes admissions could never demonstrate enough savings to warrant program costs under this methodology. Therefore, by definition, purchasers must be assuming that the program beneficially affects other utilization measures of the population, and indeed, programs aim to do precisely that. So failure to demonstrate reduction in the direct utilization measures does not necessarily imply lack of success with other types of utilization.

3. Plausibility factors do not take account of changes in population. Because the denominator is the entire population, a change in population size and composition will change the measured rate of chronic admissions per 1000, independent of any impact of a DM program (positive or negative).
4.
5. Plausibility factors do not take account of the risk profile of a population. It is entirely possible, for example, that a new group of relatively high-risk members may replace a relatively low-risk group, increasing the measured chronic admission rate per 1000, independent of any program effect.

6. The plausibility factors take no account of volatility in admission rates. As Table 1.1 illustrates, the standard deviation of admission rates per 1000 in even a population as large as Medicare can be fairly large. Thus one cannot accept a hypothesis that the program effect is positive unless the deviation from the prior year's admission rate is outside a confidence interval based on the standard deviation; conversely, one cannot reject the hypothesis that the program effect is positive simply because the difference between two rates is small (or even positive!).

7. The plausibility factors, unlike the underlying adjusted historical control methodology, take no account of existing trends in the population. As Table 1.1 illustrates, admission trends are frequently low (lower than overall trend). The results shown in Table 1.1 are significant in that they illustrate that (for Medicare members) discharge trends have generally been slightly negative for many chronic conditions, in an environment in which chronic prevalence has been increasing in the Medicare population.

To understand this last point, consider Table 1.1which illustrates the actual trends in discharges per 1000 for certain chronic conditions for Medicare patients for selected years between 1984 and 2000, and annually between 2000 and 2011. We can ignore the early years (because of program changes and changes in definitions of DRGs) and focus on the period 2000 to 2011. In 2008, DRGs were re-defined and data prior to 2008 will not always be comparable to those after 2007, although the trends remain obvious. Bronchitis and Asthma discharges are no longer reported every year because they do not meet the minimum threshold for reporting.

Trends in Discharges/1000 for Major Chronic Conditions[10]

This will have to be redone for better contract between the lines MJB

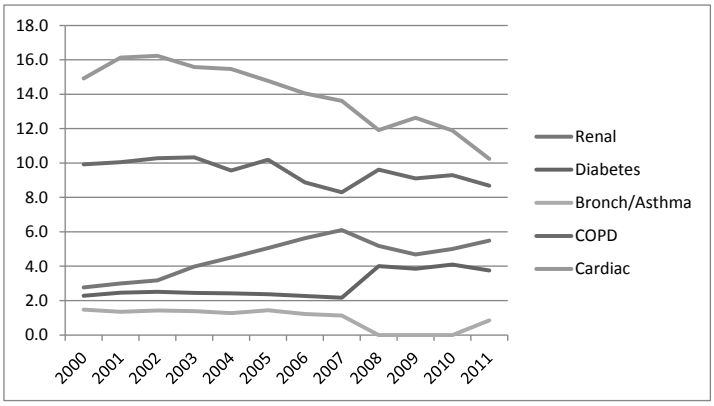

Figure 1.1

Over the 11-year period, the average trends in admissions for major chronic conditions (diabetes, heart conditions and COPD) while low, were negative. Even the overall trend in Medicare Admissions is negative as well. Admissions for Renal Failure, a condition that is generally omitted from studies of program outcomes, however, have been significantly positive during this period. Table 1.1 reports the trend (calculated as the coefficient of the admission rate in a simple linear regression fitted to the years 2000-2011 (2007 in the case of diabetes and bronchitis/asthma). The standard deviation is also shown. Figure 1.1 illustrates the same results in graphical form.

[10] Author's calculation from Medicare and Medicaid Statistical Supplements for the specific years. See: http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/ MedicareMedicaidStatSupp/index.html. Accessed June 2013.

TABLE 1.1 Medicare Discharges/1000 for selected Conditions

Year	Membership	Diabetes	Renal Failure	Bronchitis & Asthma	COPD	Heart	All Discharges
DRG		294*	316**	096***	088+	138, 139, 143-4++	
1984	29,996	4.717	1.547	5.937	7.084	11.906	363.213
1990	33,731	2.743	1.443	5.624	4.294	12.493	311.936
2000	39,211	2.280	2.768	1.470	9.925	14.914	298.895
2001	39,625	2.458	3.001	1.352	10.047	16.130	308.660
2002	40,079	2.516	3.174	1.428	10.275	16.229	314.563
2003	40,696	2.450	3.985	1.385	10.335	15.575	315.941
2004	41,391	2.425	4.498	1.276	9.564	15.459	312.100
2005	42,129	2.368	5.055	1.442	10.191	14.772	306.294
2006	42,975	2.267	5.632	1.217	8.878	14.050	288.170
2007	43,910	2.172	6.105	1.128	8.295	13.608	274.112
2008	45,067	4.008	5.175	n/a	9.616	11.910	262.294
2009	46,195	3.858	4.686	n/a	9.104	12.633	250.205
2010	47,316	4.096	5.002	n/a	9.295	11.888	260.817
2011	48,511	3.750	5.486	0.847	8.684	10.239	236.908
Mean (2000-)		2.367	4.547	1.337	9.517	13.951	285.747
Std. deviation		0.110	1.046	0.113	0.644	1.838	26.642
Trend 2000-2011^		-2.6%	8.8%	-2.5%	-1.3%	-3.3%	-2.3%

^ For diabetes and COPD, trends are calculated for 2000-7 only because of the re-definition in DRGs that occurred in 2008. Data for 2007 are reported for only 9 months and have been annualized.
* data are no longer reported for the Diabetes DRG and reported data from 2008 onward are for ICD-9 250.x. Prior to 2008, limited to age 35+.
** Now DRG 682/3
*** Now DRG 202; not reported in all years because of relative insignificance.
+ Now DRG 190/1/2
++ Arrythmia and Conduction disorders, chest pain and other circulatory disorders. Now DRG 308-10 and 313-4.

Some caution should be exercised when reading this table: underlying definitions may have changed during the illustrated periods and data are for Medicare only (comparable commercial data are not yet available, although we expect the new Health Care Cost Institute data to become available in the near future, which will make comparable Commercial reporting possible). Some conclusions may be drawn from the data, and should be kept in mind when reading the remainder of this book:

- At least in Medicare, while there are exceptions, there has been a downward trend in admissions for major chronic diseases for some years. The external trend in any population or measure must be considered in any study that considers a longitudinal population.
- Admission rates are subject to wide variability. Simple analyses that compare pre- and post- admission rates should consider this underlying variability in outcomes.

- The Medicare population experience illustrated here is likely to be more stable than commercial experience because the Medicare population itself is more stable; a commercial insurer or employer is subject to "churn" of both employees and groups, and we therefore cannot simply assume that the *underlying* population whose admissions we are measuring is comparable from year to year.

TOWARDS A UNIFIED THEORY OF UTILIZATION BASED MEASUREMENT

Tom Wilson's article[11] published some years ago highlighted the importance of demonstrating a causal pathway for any savings calculation. When measured appropriately to take account of underlying trends in the population, the reduction in utilization is a powerful demonstration of causality. In the future, with large volumes of both Medicare and Commercial data (both utilization and cost) becoming available, it is possible that we will be able to model expected utilization and cost for a specified condition-population, and compare this with that population's actual utilization and cost. A statistically-significantly lower utilization rate in a managed population would provide a powerful demonstration that an intervention or program had worked.

1.3 WHERE TO FROM HERE?

More work needs to be done to understand some of the areas we analyze in this book, and those discussed above. Other areas for future research include:

1.3.1 CONDITION IDENTIFICATION

In Chapter 12 we consider the effect on the measured results of changes in the way chronic members are identified. In Chapter 11 we also demonstrated that *when* the member was identified as having a chronic condition can have a significant effect on trend, and thus, on the estimated savings from a program. Understanding the impact of these issues on a study is not just an actuarial task and will require involvement of clinical and actuarial researchers.

1.3.2 TRANSITION STATES

We have discussed some of the implications of a transition state model earlier. If we understood chronic members' propensity to change states (particularly as their disease condition matures over time) we could perhaps do a better job of analyzing how and whether an intervention has changed that trajectory.

1.3.3 WHAT "WORKS" IN CARE MANAGEMENT?

Those of us who are practitioners in this area have been focused, because of the needs of our employers and clients, on assessing the impact of a program, particularly on financial outcomes. This focus has often been on program results at the expense of attempting to discern the impact of different types of intervention within sub-populations. For example, a typical

[11] See [220].

disease management program may include different types of interventions delivered to many different member sub-populations (with different conditions; co-morbidities; level of severity and risk). Programs often co-exist within a health plan, with case management interventions that apply yet more intensive management to a member's problems. My prediction for care management in the future is that we will see fewer, more intensive interventions targeted at smaller chronic populations, within integrated programs that include both intensive case management and broader population management (or wellness), often delivered through a more cost-effective medium such as the internet. This trend will increase our need to know what works, with whom. It will also increase the need for more accurate predictive models to be able to identify those members who match the "target" profiles. The Value Chain approach, outlined briefly in Chapter 5, may provide a basis for understanding program components. But it will require the care management companies to be willing to share much more detailed data if we are to answer questions like "what works?"

1.3.4 A "STANDARD" METHODOLOGY

The DM industry has struggled and failed for a number of years to agree on a standard measurement methodology. By default, most evaluations tend to be performed using a variant of the actuarially-adjusted historical control methodology. Given that a large percentage of industry evaluations are performed using a similar methodology, with variation being in the details (chronic definitions; timing; exclusions and inclusions), I have suggested above that a more potentially useful expenditure of the industry's resources would be in understanding the impact on the measured results of these definitions, as a pre-cursor to developing a common set of definitions. The industry has for too long struggled to respond to the demand for an absolute result (how much was saved), a problem that was answered by the Medicare Health Support program, rendering industry efforts to develop outcomes standards somewhat redundant. Instead, the industry should borrow a leaf from the National Council for Quality Assessment (NCQA) book and develop a set of measures *together with standard definitions* that health plans and those performing interventions could produce that would allow comparisons to be performed. I do not think that any user of NCQA's HEDIS measures would necessarily believe that these are an *absolute* measure of health plan quality or, for that matter, that they are the only measures of health plan quality. But the measures, imperfect as they are, have the advantage of being standardized, produced by all health plans, and therefore comparable. The care management industry could perhaps learn from the experience of NCQA and develop similar measures (and definitions) that would allow valid comparisons between programs and vendors.

1.4 WHAT HAVE WE LEARNED FROM OUR RESEARCH?

The key conclusions from the research can be summarized as follows:

1. The most important objective in any care management outcomes study is to ensure comparability between the intervention and comparison populations. The existing care management evaluation literature tends to encourage a belief that there are two "threats to validity" in studies: selection bias, which will be observed when participants are compared with non-participants, and regression to the mean. But as we show in Chapter 3, regression to the mean is an *individual*, not population concept (except in the default case of a population comprising similar individuals). As discussion throughout our re-

search suggests, the identification and correction of regression to the mean is a much larger and more complicated issue than some of the literature suggests, particularly when definitions of who is included in a population may not be clear.

2. The Economics of Care Management (as discussed in Chapter 6) is probably even more important than it was when the first edition was published. This is due to the explosion in the number of programs and program sponsors, particularly among agents who often are unsophisticated financially. An important question to ask about any program is whether the claimed (or projected, in the case of a proposed program) savings outcomes are plausible. Application of a simple economic model to the underlying population data allows users to estimate a range of likely outcomes, as well as test the sensitivity of those outcomes to different program components. More importantly, understanding the key variables of the financial model and their contribution to the overall financial outcome will allow analysis of individual proxy variables that can be directly measured (the enrollment rate, for example).

3. Population studies, a common study design in care management evaluation, may achieve comparability if the populations being studied do not change much from period to period. A major challenge for actuaries is to demonstrate this stability. Fortunately, actuaries understand the issues involved in ensuring comparability over time and the implication for PMPM costs when comparability is not achieved. Actuarial tools such as risk-adjustment make assessment of risk profiles over time and demonstration of equivalence simpler.

4. As discussed in Chapter 7, the actuarially-adjusted historic (pre-post) design, which is the most prevalent in the industry, offers a reasonable compromise between validity and practicality. Many would wish to use more scientifically-pure methods, but, as we discuss, these are seldom achievable. Instead, the popularity of the actuarially-adjusted historical control method in the industry is testament to the fact that a well-executed study is viewed as being reasonably reliable. The work that we and other researchers have done attempts to address some of the areas of sensitivity in outcomes, for example the identification of patients for different populations.

5. While the fundamental evaluation methodology does not vary much between practitioners, the assumptions and methods used to deal with data issues do vary considerably. Definitions matter. We cover in Chapter 8 many of the issues that are usually considered in a study – exclusions, inclusions, timing, and so on; principles that are widely applicable within many program evaluations.

6. Chapter 11, published in stand-alone form in the North *American Actuarial Journal* in October 2006, highlights the issue of chronic identification and its impact on chronic prevalence and trends. In a population study, the issues of *what* claims codes identify a chronic population, *when* those codes have to be observed, how frequently and over what time period, are crucial. As an industry we have only begun to scratch the surface of these issues, but it is probably the single most important issue for the industry to focus on in the future.

7. It is important to understand the impact or "value" of different assumptions on the final results of a study. It is surprising to me that many of the discussion in the literature remains at a theoretical level when many practitioners have access to data sets and could simply test out some of the issues that they debate. The industry would greatly benefit from it. It would make the current methodology more robust and would reduce the need for the industry to search for alternative methodologies. In Chapter 12 we examine some of the sensitivities of the results calculated using one such methodology for one client, under different assumptions. Much more of this type of analysis needs to be published, to gain knowledge about the methodology.

The care management industry continues to expand. One area of growth is Wellness and Worksite Health. Recognizing this, DMAA (formerly the Disease Management Association of America) was re-named "DMAA-The Care Continuum Alliance" in 2007. DMAA now covers the new, broader spectrum of interventions. The chapters that address some of the issues of Wellness and Worksite Health programs have been selectively updated to reflect changes in this area and some of the recent published literature.

1.5 A CHANGE IN EMPHASIS FOR ACTUARIES?

Traditionally, actuaries have focused on financial analysis and worked with aggregate data, often at a category of service level. An example of the change in emphasis in recent years is the need to deal with chronic populations, which requires that the actuary have a more detailed knowledge of the medical conditions, underlying services and treatments (and the claims that they generate) that a particular member requires. This, in turn, requires actuaries to have more clinical knowledge than was the case in the past. The Society of Actuaries has enthusiastically supported actuaries' involvement in broader healthcare topics. One indicator of this change in emphasis is the popular "Medical School for Actuaries" seminar that the Society of Actuaries hosts twice each year. Actuaries are increasingly fluent in topics such as risk adjustment that require knowledge of claims data and the medical conditions that generate them.

This change in emphasis turns the old "financial" analytical paradigm around, with the condition-population becoming the unit of interest and analysis. We still have much to learn about the behavior of traditional actuarial measures (for example cost PMPM and trend) when applied to sub-populations with common characteristics, such as a health condition. And because our clinical colleagues have barely begun to scratch the surface of what constitutes "clinical best practice" for members with conditions, we have a long way to go before we can begin to benchmark utilization and cost for these populations. Nevertheless, for the actuary interested in pursuing this area of practice, the techniques and tools described in this book (and in the companion volume, Healthcare Risk Adjustment and Predictive Modeling) are a place to start.

10 MEASURING CARE MANAGEMENT SAVINGS OUTCOMES

10.1 INTRODUCTION

Controversy over Disease Management outcomes has been part of the industry since its inception. Many authors and researchers have struggled to find a suitable methodology that will give results that are credible, reasonable, and acceptable to purchasers. Examples of discussions of general methodological principles for ensuring validity in Disease Management savings outcomes measurement may be found in several papers[1]. More recently the industry trade association, Care Continuum Alliance (CCA) has assembled a number of workgroups of industry experts to address methodological issues, resulting in the publication of three "Outcomes Measurement Guidelines"[2]. Several of these references, in addition to discussing measurement principles, provide lists of different study types and designs.

Since the first edition of this book was published in 2008, care management has undergone many changes. The Disease Management (DM) model that dominated the first decade of the 21st century has given way to a multitude of different programs and interventions, as we discussed in Chapter 3. Unfortunately the fracturing of the comprehensive DM model and increase in the number of smaller, more targeted models has made it more, not less, difficult to evaluate intervention programs (particularly when there is overlap or duplication between programs, as happens increasingly).

This Chapter will address both the theory of measurement design, and provide a practical evaluation of the most common designs for the practitioner.

10.2 EVALUATING A SAVINGS CALCULATION

The actuary may not always have the chance to design a measurement study, and will more frequently be called in to evaluate a vendor's or colleague's results. Of utmost importance is confidence in the validity of the study. Validity is important in research and in commercial applications because it gives an indication of how well a particular study addresses the nature and meaning of the variables involved in the research, and how much reliance may be placed on the results. Internal validity relates to whether the results of the study can be ascribed to the actions taken or whether they are the result of other factors. External validity asks whether the results of a study are reproducible and can be generalized to other populations or settings.

[1] See [219], [218], [59], [119], [62], and [64].

[2] See [155] and [156].

Three questions should be considered when evaluating results:

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

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

Evaluation requires two concepts defined previously – Causality (see Chapter 4.3.2) and Methodology (see Chapter 4.3.1).

10.3 PRINCIPLES OF MEASUREMENT DESIGN: WHAT CONSTITUTES A VALID METHODOLOGY?

Evaluation of a methodology is a different problem than the evaluation of the results of a study. The former is a question of conformance to evaluation principles, while in the latter case we evaluate whether or not the author's hypothesis is rejected. This chapter is not intended to be a review of the statistical principles of hypothesis testing, but a brief summary is provided in Appendix 7.2.

Whether designing a study from scratch, or evaluating a published study, the same principles determine whether a methodology is likely to be judged acceptable. The principles below are discussed in Chapter 4.

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

- The methodology must be one that a purchaser (or its consultant) is familiar with, or at least can grasp readily, and that should be perceived in the market-place as sound;
- The methodology must be documented in sufficient detail for another practitioner to replicate the study, and, if required, allow the client to be able to replicate the savings estimates themselves (or at least major components of the calculation);
- The results of the application of the methodology must be consistent with the client's savings expectations, and plausible overall;
- The application should lead to stable results over time and between clients, with differences between different studies and clients that can be explained; and

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

10.4 STUDY DESIGNS FOR DM: A SUMMARY

There is a principle in program evaluation, known as "hierarchy of evidence" that is worth keeping in mind when evaluating a study. There is general agreement about the ranking of different methodologies in terms of the validity of a study performed using them. Randomized trials rank above (what is referred to in by researchers as) observational studies. As we have noted elsewhere, randomized studies are rare (but not unheard of) in program evaluation, so the actuary is more likely to encounter some form of observational study. A ranking of different types of study is shown in Table 10.1.

TABLE 10.1

Hierarchy of Study Designs
Study Design
1. Randomized Control Trial
2. Cohort Studies
3. Case Control Studies
4. Cross Sectional Designs
5. Survey Studies
6. Case Studies

Some authors assign Meta Analyses a high weight in terms of credibility. We do not include this type of study in the list (despite the importance of this type of study) because Meta Analyses, by definition, are an aggregation of other studies. Despite the value that they add by showing consistency (or otherwise) of outcomes in an area, their credibility depends on the strength of the underlying studies.

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

- Control Group Methods
- Population Methods without control groups
- Statistical Methods

10.4.1 Control Group Methods

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

A control group may be:

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

- *Geographic* (comparing equivalent populations in two different locations). Unlike randomized controls, in which the control group is subject to the same forces as the intervention group, the risk profile and market forces present in different geographies may cause differences that obscure ("confound") the true difference in the intervention and reference populations. In many cases, these differences may be anticipated and adjustments made in the study. See, for example, "Actuarial Adjustment" in "Dictionary of DM Terminology[3]." This adjustment is easier to make when there is no dynamic effect on the reference population over time.
  Consider the example in Table 10.2:

  In this example, (assuming that the Intervention Population represents a population before and after the implementation of a program) there appear initially to be no savings: costs increased by $2 PMPM between the Baseline and Intervention periods. This result is due to the confounding effect of healthcare cost trend. Comparison with the Reference Population experience shows that trend at the rate of 16.7% was present in the Reference Population. The obvious adjustment to the Intervention population data is to multiply the Intervention period cost by \$100/\$90, or $\$105 \times 1.11 = \$116.67$. Then the estimated savings from the intervention would be: $\$116.67 - \$102.00$ or \$14.67.

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

[3] See [52].

TABLE 10.2

Application of Actuarial Adjustment to a Reference Population

	Claims Per Member Per Month	
	Reference Population	Intervention Population
Baseline Period Cost	$90	$100
Intervention Period Cost	$105	$102

- *Temporal* (also known as the Historical Control Design). This design compares equivalent samples drawn from the same population before and after the intervention program. This is the most common approach used in the Disease Management industry, and uses a medical Trend adjuster to project the historical experience to the same time period as the intervention data.
- *The Product Control Methodology* compares samples drawn from the same population at the same point in time, but differentiating between members who have different products, such as HMO vs. PPO, or indemnity vs. ASO. Clearly the introduction of product differences introduces the potential for confounding effects of product selection, different medical management, included benefits or providers (often a factor with ASO groups) and reimbursement, and this approach should be treated with caution. The mathematics of this methodology are similar to those of the Geographic methodology (see above).
- *"Patient as their own control" (Pre-post Cohort Methodology)* This method differs from the "temporal" method described above, in which the intervention and comparison populations are re-sampled in each period to ensure equivalence. Applying the same rules of identification to create an equivalent population in a different time period is somewhat analogous to the "with replacement" and "without replacement" problems with which actuaries are familiar from introductory statistics courses, where problems are often stated in terms of a number of red and black balls present in an urn. In the "Patient as their own control" method, the comparison group is the population as initially defined, but measured post-intervention. In this design, there is no equivalent reference population. One conclusion from our discussion in Chapter 4 is that regression to the mean is potentially present in the post-intervention population (often because candidates are identified for intervention based on recent claims experience). If the extent of the regression were known, an adjustment could be applied to the population. In most cases, however, the extent of regression is not known.
- *Participant vs. Non-participant studies.* In this method, the experience of those who voluntarily elect to participate in a program is compared with the experience of those who choose not to participate. The participants represent a group with potentially different risk-factors to that of the non-participants (we already know that they differ with respect to the important factor of willingness to take control of their own health by engaging in a program). Some authors appear to believe that it is possible to adjust the reference population (non-participants) to bring them into equivalence with the intervention population. It does not appear to us, however, that the effect of selection can be estimated. In any event, the existence of selection bias is known in the industry and this methodology is not assigned a high credibility, with or without adjustment.

- *Other types* of control group methods are cited in the literature. An example of a "staggered roll-out" method used in examining diabetes outcomes can be found in the literature[4]. These methods, however, can be de-constructed to fit one or more of the above five categories.

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

10.4.2 Non-Control Group Methods

- Among non-control group methods, *Services Avoided* methods are commonly used, particularly for case management applications., In the application of this methodology to case management or utilization review, the intended resource utilization of the member prior to the intervention is estimated because the member calls a health plan to pre-authorize a particular service. Savings are then estimated by performing a cost-estimate of the requested service, and comparing this estimate of cost with actual cost of services used after the intervention. In the specific example of case management, an estimate of the likely resource utilization of the member is compared with actual approved utilization (including any alternative services arranged or approved by the case manager), and the difference is counted as savings due to the case management program. Some applications track the utilization of members who report a change in intent (for example the intent to have surgery) for as long as 6 to 12 months post-intervention to ensure that the change in intent was not later reversed. Because of its wide-spread use, the methodology scores high on familiarity but lacks a reference population. The method also includes a high degree of subjectivity both in selection of candidates and in estimating what utilization would have been, absent of intervention. Case managers are also known to request more resources and services than required, in the expectation that some will be denied, thus over-estimating the savings. For these reasons, the validity of savings calculated by this method is questionable.

- *Clinical Improvement* methods have achieved reasonable market acceptability, and gained ground in recent years because of the increased focus on clinical improvement measures (see the Appendices to Chapter 8 for examples of the many different quality measures now required from plan sponsors). In such a method, the change in an objective clinical measure is first observed (for example, the rate of use of a particular

[4] See [205].

medication by members who have a diagnosis for which the medication is indicated, or a diabetic member's HbA1c score). The peer-reviewed clinical literature is searched for studies that indicate how health is improved (and resource utilization is decreased) in the population with the particular diagnosis, as a result of adherence to treatment. A dollar value is assigned to the reduction in resource utilization, which is then applied to the observed change in the clinical measure due to the program. This method appeals to some evaluators because it involves objective causal factors: unlike some other methods that measure changes in claims, this method can point to actual improvement in a clinical factor that can cause reduction in claims. Despite its appeal, the methodology rates relatively low because the results are achieved through a subjective process and often lack a reference group – in this case, the subjective element is the estimation of financial savings by inference from published clinical studies (somewhat akin to a benchmark method). To our knowledge, no study has ever been published that compares estimates of savings in a population using a clinical improvement approach with estimates in the same population using a randomized or other control group approach. We should also remember that our review of the literature in Chapter 5 found that the literature does not yet show a strong, demonstrated, link from clinical to financial effects.

10.4.3 Statistical Methods

The term "statistical methods" is used when purely statistical techniques are involved (for example, regression or benchmarks), rather than the construction of an explicit reference population. The term "Statistical Methods," however, should not be confused with the statistical tests that underlie hypothesis testing (see Appendix 7.2). Statistical tests of hypotheses should be applied to any study to determine whether the results are significant.

- *Time-Series Methods.* The objective of these methods is to fit a curve or series of curves to the data over time, and then to demonstrate a divergence from the best-fit line once the intervention is applied. This method is a generalization of the trend-adjusted historical control methodology, which focuses on just one historical period. The difficulty of fitting a curve to healthcare data over time appears to us to be almost insuperable, because of the need to capture in the model the effect of a multitude of factors, both endogenous and exogenous. Because of the difficulty of demonstrating a high correlation between the actual data and the fitted line, demonstrating divergence from that line and assigning causality to the DM program is complicated.
- *Regression Discontinuity*[5] This design may be considered as a special case of the Time Series method (above). At its core, this method looks for a statistically-significant difference between two similar sub-sets of the population. A regression line is first fitted to data that relates pre- and post-intervention experience. A dummy variable is included in the regression to capture the difference between the intercept of the Intervention Population's regression line at the "cut-off point" and that of the Reference Population. To understand the Regression Discontinuity method, consider Figure 2. In this example, we plot the relationship between an individual's risk-score in a baseline period (Year 1) and a cost per member per month in the follow-up period

[5] This section is adapted from publications of William M.K. Trochim ("Research Design for Program Evaluation" by W.M.K. Trochim. Sage, 1984) and Linden A, J.L. Adams and N. Roberts: "Evaluating Disease Management programme effectiveness: an introduction to the regression discontinuity design". Journal of Evaluation in Clinical Practice, 10 (2004). We wish to thank Dr. Ariel Linden for his helpful discussion on this section.

(Year 2). Each point in the scatter represents the pair of observations for a single member. The method requires an objective way to segregate those members eligible for, and those not eligible for, the intervention. The risk-score is a useful variable because intervention programs are frequently targeted at members whose risk-score exceeds some pre-determined minimum. (Risk-score is a preferable measure for the Year 1 variable than cost because, while the group targeted for intervention often includes high cost members, this is seldom the sole criterion for targeting,) An upward sloping regression line implies that members with high Year 1 risk score tend to have high Year 2 costs, as well. The closer this relationship, the closer the data points will be to the line. On the other hand, a line that slopes upward at less than 45° indicates regression to the mean (high-cost year 1 members tend to have lower year 2 costs; low-cost year 1 members tend to have higher year 2 costs).

A regression is fitted to the Year 1 vs. Year 2 data. An example of this regression is:

$Y_i = ß_0 + ß_1 X_i + ß_2 Z_i + \varepsilon_i$, where:

Y = Dependent variable (year 2 cost for the i^{th} person)

β_0 = Regression intercept

X_{1i} = Independent variable (year 1 cost for the i^{th} person)

β_1 = Regression coefficient for variable 1

Z = Dummy variable with value zero (if observation is in the Reference Population) or 1 (if in the Intervention Population)

β_2 = Regression coefficient for dummy variable Z

ε_i = Random error term for the i^{th} person

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

A significant value for the dummy variable regression coefficient (β_2) implies that there is a statistical difference between the intercept of the basic line (reference population line) and the intervention population line at the cut-off point. The value of this coefficient gives an estimate of the effect of the program at that point. The focus on the cut-off point may seem excessive, but an important feature of this method is that the effect is calculated at a point (the cut-off point) at which the reference and intervention populations are most similar. This overcomes a potential objection that there is not a reasonable "goodness-of-fit" throughout the entire population, particularly at the extremes of the distribution. Second, while this method may have been used to demonstrate a significant effect, clients who are purchasers of care management interventions require an estimate of the savings due to the program, and we are not aware of an actual savings calculation using this method. Savings could, however, be calculated by projecting the expected cost of the intervention population using the regression analysis, and subtracting the intervention population's actual expenses (as in the hatched area above). We have described this method at greater length than some others because there is a considerable interest in its potential application in commercial calculations, and we expect to see the method used more in the future.

Relationship Between Costs In Two Periods

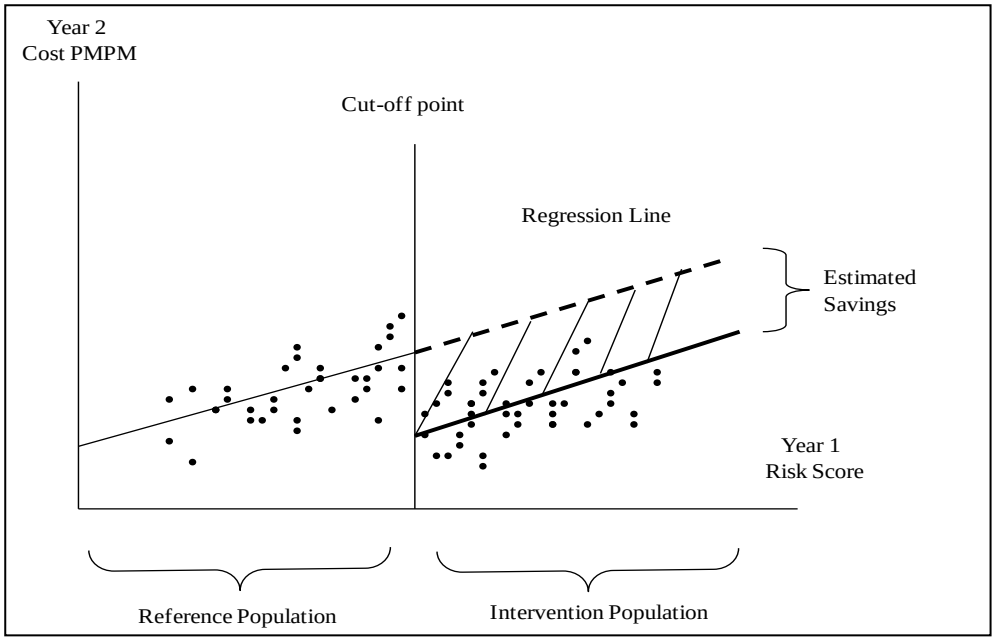

FIGURE 10.1

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

Statistical methodologies are restricted to the scientific community; we have not seen them used in wide-spread commercial application. The more commonly-used control group and non-control group methods are simpler to understand and the calculations are transparent. Statistical methods involve techniques with which most business users are not familiar and may regard with a degree of suspicion (for their "black box" aspect). The Regression-discontinuity method is often discussed favorably in the literature, but we have yet to find significant analyses using this method. Time series methods have one important advantage in that they draw attention to long-term utilization and cost trends in the population, which provides the evaluator with valuable information about what was happening in the population before the intervention program began. We question the practical usefulness of this method in a health plan environment, where so many variables change over time, making it virtually impossible to control for confounding. Benchmark methods are favored by some authors (and have some appeal to actuaries,

who are used to making the type of adjustments required to compare different populations). The sheer number of variables and risk-factors, however, (and lack of information about their values) that could potentially affect a benchmark study will make this another difficult methodology to apply in practice. Statistical methodologies may yet prove useful but none is developed to the point of being practical for implementation in a commercial environment.

Some authors suggest "Propensity Scoring" as a methodology, equivalent to others considered above. We recognize the importance of Propensity Scoring, a method of identifying or creating populations of the same degree of risk as the intervention population. (Propensity Scoring is potentially of importance to actuaries because of its similarity to Risk Adjustment. We address this technique in greater detail in Chapter 11.) Because we do not consider Propensity Scoring a methodology, but rather a technique for adjusting other populations or creating matched populations, we do not include it in our comparison[6].

Comparative Assessment of Methodologies

The table on page 130 summarizes our conclusions concerning effectiveness of different measurement methodologies in meeting our key criteria (above) validity:

- Inherent validity (lack of obvious bias);
- Scientific rigor;
- Familiarity (how commonly used is the methodology in the industry?);
- Market Acceptance (how is the method perceived in the market-place?);
- Ease of replication and auditability;
- Application (how is the methodology applied in practice?), and
- "Other Issues" (other important issues in the application of each methodology).

These criteria for assessing methodologies are our own, and reflect our experience as consulting actuaries in this area. Other actuaries, or practitioners from other disciplines, may have different criteria by which to judge methodologies. The point, however, is that methodologies are not equally valid, and results that are prepared according to a higher-scoring methodology should be given more weight.

10.6 THE PROBLEM OF MULTIPLE PROGRAMS

As we noted at the beginning of this chapter, the changes in the care management industry that have taken place in recent years have resulted in a multitude of programs, frequently addressing the same issues in the same patients. When Disease Management was the dominant model we could assess a single program; now we have to attempt to assign effectiveness to a number of different programs. Unfortunately there is no generally accepted industry method for doing so.

One approach that we have found helpful in the circumstances is to identify specific goals of a program, and members that are identified who meet the definition of eligibility for the program. This enables us to take an "intent-to-treat" approach to evaluation, which aids validity. We can then assess measures that match the program's specific goals (e.g. reduction in admissions or improvement in quality measures). An alternative approach is to measure a population in which we know multiple interventions have been applied and compare the effect to that in a population

[6] For more information on this technique, see [122].

in which all but the specific intervention are present. This allows us to estimate the marginal effect of the intervention.

Another alternative is to stagger the implementation in sub-populations so that there are (at least for a time) comparable populations with and without the intervention.

No evaluation method is entirely satisfactory in these circumstances, and better approach is to work with the program sponsor to hold out a "virgin" population in which no other interventions are present. If this can be done, even over a short period, the resulting outcomes will have much greater credibility.

10.7 CONCLUSION

A non-control group methodology is unlikely to be a satisfactory method for calculating care management savings results, except under unusual circumstances. In "The Principles for Assessing Disease Management Outcomes,"[7] the committee examining appropriate research methodologies concluded that the preferred method for any evaluation is a randomized control study (in our experience, easier to implement and more practical than is commonly believed). As our discussion shows, other forms of non-randomized control group can also be valid (provided equivalence is maintained and can be satisfactorily demonstrated). A non-randomized control group could be temporal, geographic, or product-based, but not based on self-selected members (such as non-participants). The achievement and demonstration of equivalence is an area where actuaries may make a contribution. Actuaries are qualified, through their experience in rating and underwriting of health risks, to perform the process of drawing conclusions and making projections in one population from data or experience of another population. In our next few chapters, we will address issues to consider in order to maintain control of the study data and the achievement and demonstration of equivalence.

In addition to the issues of equivalence that are important in choosing a valid methodology, the application of the methodology needs to be carefully controlled, or the results of the study will be invalidated. In Chapter 12 we turn to the issue of the controls that should be in place as we apply an important methodology, the actuarially adjusted historical control design.

[7] See [62].

TABLE 10.2

Comparison of Certain Commonly Used DM Savings Calculation Methodologies

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

TABLE 10.2 (Continued)

Comparison of Certain Commonly Used DM Savings Calculation Methodologies

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