DISEASE MANAGEMENT
Volume 7, Number 2, 2004

Using an Empirical Method for Establishing Clinical Outcome Targets in Disease Management Programs

ARIEL LINDEN, Dr.P.H., M.S.,[1] JOHN L. ADAMS, Ph.D.,[2] and NANCY ROBERTS, M.P.H.[3]

ABSTRACT

The disease management (DM) industry is being scrutinized now more than ever before, with programs being asked to demonstrate improvement in clinical quality in addition to the expected reduction in medical costs. In healthcare, clinical improvement targets are often set at levels considered to be clinically meaningful. This difference may or may not be statistically significant. The term "effect size" refers to the smallest difference that could be detected statistically. This paper proposes a simple empirical method for determining the minimum expected improvement level for DM clinical outcome measures in which two proportions are being compared. This method is useful in situations where the outcome measure does not lend itself to be determined by the subjective judgment of medical expertise. Graphical displays are provided for the reader to use to help determine appropriate effect sizes for studies in lieu of, or in addition to, the statistical calculations. (Disease Management 2004;7:93–101)

INTRODUCTION

THIS IS THE THIRD ARTICLE in a series in which the authors present methodological issues pertaining to the evaluation of disease management (DM) program effectiveness. Linden et al.[1] provided a comprehensive analysis of the "total population approach" which is currently the most widely used method for assessing DM program efficacy. Due to the multitude of limitations and threats to validity of this design, in a follow-up paper, Linden et al.[2] described an alternate and more appropriate methodology for evaluating DM program effectiveness, using time series analysis on utilization data. In the current paper, the authors shift focus to clinical outcomes measurement—an equally important component of any DM program assessment.

The DM industry is being scrutinized now more than ever before, due to the newly introduced accreditation programs of the National Committee for Quality Assurance (NCQA) and the American Accreditation Healthcare Commission (better known as URAC); the introduction of demonstration projects sponsored by the Centers for Medicare and Medicaid Services (CMS); and the heightened awareness of purchasers as to what they should expect and demand from DM programs.[3] As such, these programs are now being asked to demonstrate improvement in clinical quality in addition to a reduction in medical costs.

In a recent collaborative effort, Johns

[1]Linden Consulting Group, Portland, Oregon.
[2]RAND Corporation, Santa Monica, California.
[3]Providence Health System, Portland, Oregon.

Hopkins Outcomes Verification Program and American Healthways, Inc.[4] created a report that details several standard outcome metrics for five disease states—asthma, diabetes, ischemic heart disease, congestive heart failure, and chronic obstructive pulmonary disease. The intent of this effort was to develop a set of standardized metrics that could be used across the various DM settings. The majority of clinical measures included in this report are similar to, or included in, the Health Plan Employer and Data Information Set (HEDIS®), which is another set of clinical measures widely used in the managed care industry for quality improvement and benchmarking purposes.[5] The data generated from this measurement design are used to compare proportions of two independent groups, with the results normally expressed as the proportion of patients receiving "X" in one measurement period compared to the proportion of patients receiving "X" in a subsequent measurement period. In some cases, "X" is the receipt of a particular test or service (regardless of the level), and in others "X" is achieving a specific desired goal level. For example, an outcome metric for patients with diabetes might be the proportion receiving a glycosylated hemoglobin test (HbA1c) in each measurement period. Alternatively, the outcome metric might be defined as the proportion of patients with diabetes with an HbA1c at or below 7%. In both cases the denominator is all diabetics. In the first example, the numerator is the number who received the test. In the second example, the numerator is those achieving the goal level.

While this preliminary work in developing clinical outcome measures is a great first step, establishing appropriate and achievable targets for these studies should be the logical next milestone.

In healthcare, clinical improvement targets are often set at levels considered to be clinically meaningful. This difference may or may not be statistically significant. The term "effect size" refers to the smallest difference that can be detected statistically. Some outcome measures do not lend themselves to be determined by the subjective judgment of medical expertise. What may be considered clinically significant for an individual patient cannot in all cases be expected for a larger group or population. For example, a physician should strive to have the glycosylated hemoglobin (HbA1c) level of a patient with diabetes below 7%.[6] On a population basis, however, it would be unreasonable to expect that every patient with diabetes will have an HbA1c level below 7% (equivalent to a rate of 100%). Moreover, there are well-documented regional variations in practice patterns[7,8] that may limit the generalizability of one standard predetermined effect size. Where, then, should the target be set?

Some accreditation and regulatory bodies have set predetermined outcome levels while others require organizations to set a performance target before the study begins and report how that objective was chosen. For instance, Medicare's Quality Assessment and Process Improvement (QAPI) projects once required a 10% reduction in the performance gap between year 1 (proportion 1) and year 2 (proportion 2). In other words, if a population had a 50% performance level in year 1 (meaning a performance gap of 50%), and then moved that performance level to 56% in year 2, then the performance gap was reduced from 50% to 44%, which is a 12% reduction in the gap. Recently this requirement was changed to allow a contracted Managed Care Organization to select its own performance targets, as long as the organization can provide acceptable justification.[9]

In the case of DM programs, an improvement goal of clinical outcome measures is usually ascertained by reviewing the results of similar studies conducted on internal historical data across other purchasers, or vis-à-vis established external benchmarks. However, in many cases, goals are determined simply through "eyeballing" the baseline level and then establishing a mutually agreeable target with the client organization. Increasingly, there are financial consequences associated with missing clinical outcome targets. Therefore it is in the interest of both the DM program and the purchaser of these services to establish meaningful, reasonable and achievable performance targets.

In studies where valid clinical judgment is either unavailable or provides little, if any, direction into the assignment of an appropriate population-based outcome objective, estima-

tion of effect size can be determined statistically. This paper proposes a simple empirical method for determining the minimum expected improvement level for DM clinical outcome measures in which two proportions are being compared. Additionally, graphical displays will be provided for the reader to use in helping determine appropriate effect sizes for studies in lieu of, or in addition to, the statistical calculations.

MODEL PARAMETERS

There are four interrelated parameters that have an effect on the conclusions that are attained from a typical statistical test. When any three of the following parameters are defined, the forth component can be calculated mathematically: (1) sample size, or the number of observations, subjects, or cases under study, (2) significance level, or alpha—the probability that the observed result is due to chance alone, (3) power, or the probability that a difference will be observed when it actually occurs, and (4) effect size—the magnitude of change between two groups or within one group, pre and post intervention.

While the purpose of this paper is to provide the model for calculating effect size, it is important for the reader to understand each of the four parameters and their interrelationships. This, in turn, will enable the reader to determine the appropriate values to assign each of these parameters in order to derive the appropriate effect size for their clinical outcome measure.

Sample size

In general terms, increasing sample size will concomitantly increase the power to detect a true effect, as well as decrease the effect size needed to reach statistical significance. In DM programs though, typically the number of members with clinical outcomes data available is limited. This is especially true in programs where clinical indicators can be obtained only from those members enrolled in the nursing intervention component. Therefore, the DM program should always strive to acquire clinical data on as many members as possible to maximize the potential for achieving the target outcome levels.

Significance level

In simple terms, alpha refers to the probability of finding a difference between two proportions by chance alone. For example, an alpha of 0.05 indicates that 95 times out of 100 when there was no effect, we will not conclude there was one. Conversely, we can commit a type I error by erroneously concluding that five times out of 100 there was an effect, when in fact, there was not one (a false positive). By setting the alpha level lower (eg, 0.01) we are making the test more conservative, indicating that we are willing to be wrong only one in 100 times in saying there was a difference, when in fact, there was none. While lowering the alpha decreases the chance of committing a type I error, it also reduces the chances of concluding that the DM program had an effect. In research, the alpha is typically set at 0.05.

Power

Power is used to determine the likelihood that the results of the study will yield a significant effect when there truly is one. In other words, a power of 80% suggests that 80 times out of 100 when there is a true intervention effect, we'll identify it as such. Conversely, we can commit a type II error by erroneously concluding 20 times out of 100 that there was no effect, when in fact there was one (a false negative).

As indicated above, power increases with an increase in the sample and effect sizes, as well as choosing a larger alpha (eg, 0.10 as opposed to 0.05). The rule of thumb in research is to set the power level at 80% or higher.

Effect size

As described earlier, effect size refers to the smallest difference detected between the two proportions under study. The interrelationship with the other three parameters is such that a larger sample size is required to detect a small effect size, and a larger effect size will result in higher power. Since the objective of this analysis is to establish the effect size, we only need to define alpha, power and sample size, and the effect size will be completely determined.

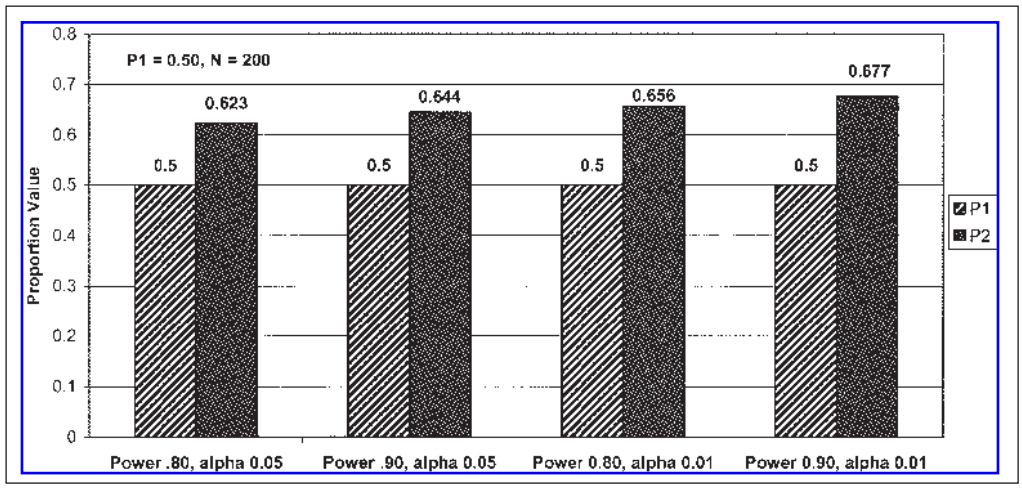

FIG. 1. The impact of different power and alpha levels on effect size, holding proportion 1 (P1) to 0.50 and sample size (N) to 200, with a one-tailed test of significance.

Effects of manipulating model parameters

Figure 1 illustrates the impact that different power and alpha levels have on effect size, when holding N and proportion 1 values constant. As shown, changing power from 0.80 to 0.90 results in an increase of the effect size by an absolute 2.1%. Similarly, changing alpha from 0.05 to 0.01 results in an increase of the effect size by an absolute 3.3%.

Figure 2 illustrates the impact of N on effect size at different Proportion 1 values, when holding power and alpha constant. As clearly demonstrated, smaller N's require a much larger effect size to meet these criteria of power and significance. Another fact that is worthy of note is that, due to the parabolic nature of the effect-size curve (because of the non-linear mathematical equation), a Proportion 1 value of 0.40–0.50 will require a larger effect size than any other proportion level. Therefore, a DM program that has a pre-intervention clinical measure within the range of 0.40–0.50 will have to demonstrate a larger impact on the enrolled members than had the proportion 1 value been either higher or lower than this range of values.

In summary, this section established that there are three parameters under control of the

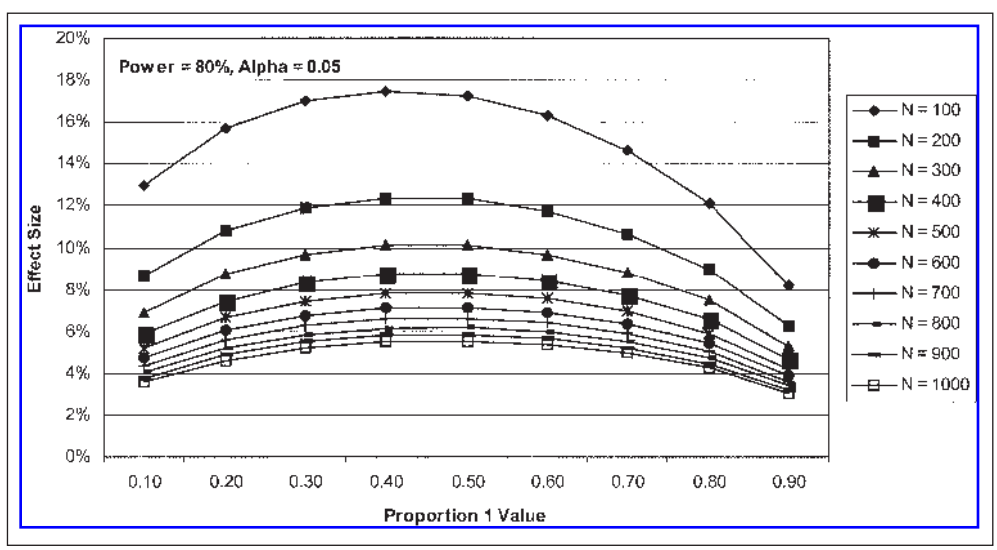

FIG. 2. The impact of sample size on effect size, holding power to 80% and alpha to 0.05, with a one-tailed test of significance.

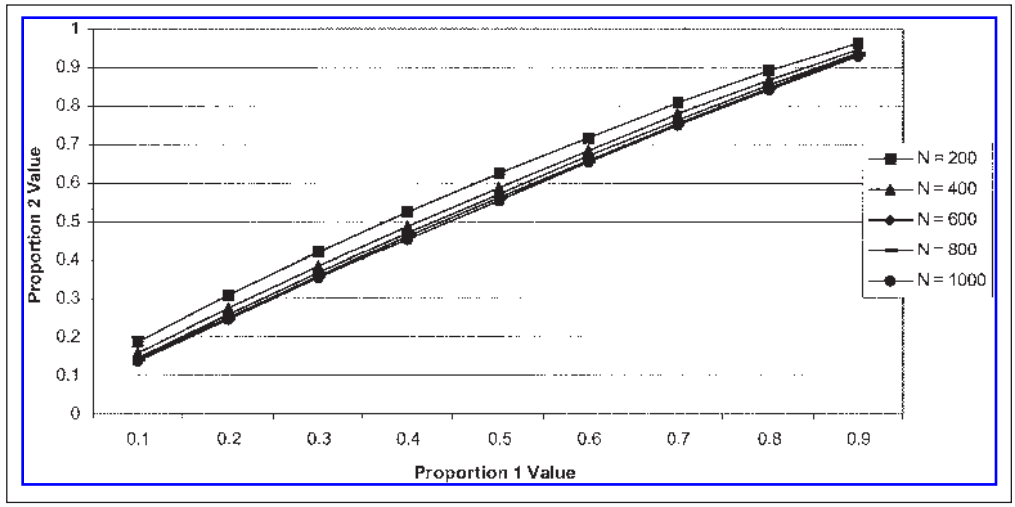

FIG. 3. Determination of Proportion 2 with power = 80%, alpha = 0.05, and variable sample sizes (N), with a one-tailed test of significance.

DM program analyst that will impact effect size; N, power, and alpha level. Additionally, it has been illustrated that the beginning pre-intervention proportion level will have an impact on the size of the effect needed to demonstrate a significant improvement in the program. While the ability to detect and adjust for type I or type II errors are important in concept, they are not germane to the topic of this paper (because we are solving for effect size, we have pre-established what level of significance and power will be required to meet these criteria). Nonetheless, interested readers should refer to Donner,[11] Lachin,[12] and Moher et al.[13] for a more comprehensive discussion on this subject matter.

DETERMINING A MEANINGFUL EFFECT SIZE

In the appendix, formulae are provided for determining effect size mathematically. Similarly, Figures 3 and 4 provide graphic displays to assist with the determination of an appropriate and meaningful effect size for different N's at varying Proportion 1 values. Figure 3 holds power constant at 80% and alpha at 0.05

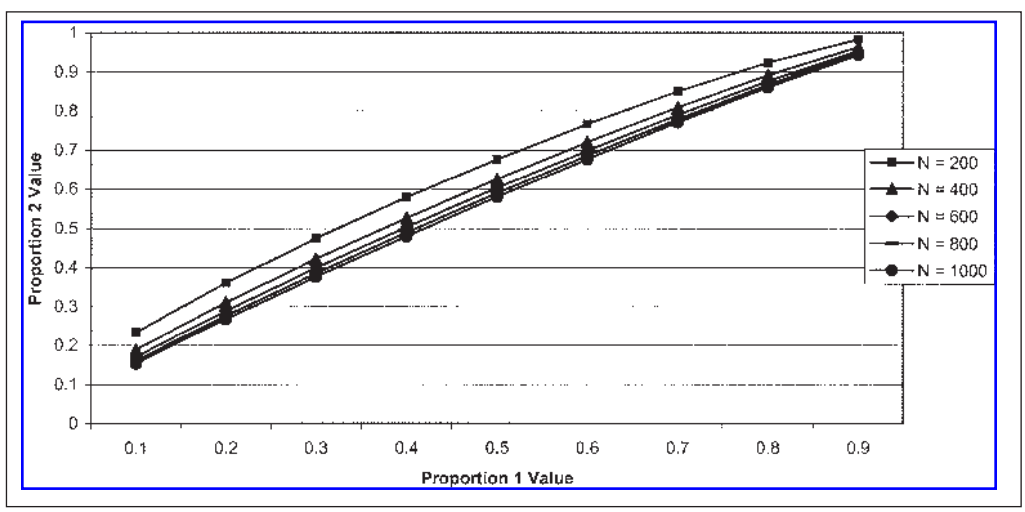

FIG. 4. Determination of Proportion 2 with power = 90%, alpha = 0.01, and variable sample sizes (N), with a one-tailed test of significance.

(less strict criteria), while Figure 4 holds power constant at 90% and alpha 0.01 (more strict criteria).

There are several important factors that one should consider when determining the appropriate effect size. This section will focus on the process for making those determinations.

Time period under study

In most cases, DM programs operate on a 12-month contract year. However, there are situations in which the contract period will be either shorter or longer than 12 months. For example, a contract originally implemented sometime within a calendar year may be shortened to end in December, to allow the following contract year to extend for a calendar 12 months. Conversely, this same contract period may be extended to include the first few months of the initial year and the complete following calendar year. In either case, the effect size should be adjusted to reflect the period of time in which the DM program can impact the clinical measures of the patient population in the intervention.

In the case of a shortened contract period, the parties may choose to agree upon a smaller effect size target. This can be determined by setting either a lower power (eg, 0.80 or less) or a higher alpha level (eg, 0.05 or higher), or both. For a longer contract period, the reverse should be assumed (higher power, lower alpha, or both). These are the only two variables that can be manipulated, since both N and the initial proportion (P1) are predetermined by the available data.

For example, let's assume that a DM program's start date is September 1st of the present year. Both the health plan and the DM program administrators agreed that the first contract period would extend to the end of the following year. Therefore, the clinical metric would be evaluated after 16 months. With an N of 800, the proportion 1 value was found to be 0.30 for the measure. If the power would be set at 80% and alpha at 0.05, the program would establish a target proportion 2 value of 0.36 (equaling an effect size of 6%). Since the DM program has 4 months longer than a traditional contract period to impact this metric, the effect size target could be made more stringent by establishing tighter controls of power, alpha or both. In this case, effect size can be increased by another 1% by changing power from 80% to 90%, and increased by another 2% by changing alpha to 0.01 from 0.05. Therefore, by using more stringent statistical criteria, the new proportion 2 target was set to 0.39 (or equaling an effect size of 9%).

This example obviously brings up the need for establishing acceptable ranges for power and alpha levels, since they are the only two variables that can be manipulated in this fashion. As a general rule, power levels are usually set at 80–90%, and alpha levels are usually adjusted to 0.10, 0.05, or 0.01. In any of these cases, the criteria must be established at the outset of the program, and agreed upon by both parties. The effect size, if achieved, will be valid in that the results were statistically significant and had power to detect a true effect.

Sample size

One useful assumption that is made when developing a model for determining effect size is that the sample size is identical in both measurements (Proportion 1 and Proportion 2). Therefore, in the case of a pre-post intervention, it is to the benefit of the DM program to augment the N as much as possible for the initial measurement (since by the nature of the effect size equation, this will reduce the effect size needed to meet the power and alpha requirements).

Another valuable reason for maximizing N is that it provides the analyst with considerably more flexibility in developing the effect size model. For example, let's assume that a DM program tracks the percentage of patients presenting in the Emergency Department (ED) for avoidable acute exacerbation of chronic obstructive pulmonary disease (COPD). If the program intends to prevent these types of ED visits by providing educational efforts to members for identifying triggers for acute exacerbations, this measure would be a valuable clinical outcome metric of the program.

Let's assume that the proportion 1 value was 0.60, and the N = 1651 (representing the total number of patients presenting to the ED for

acute exacerbations in the initial baseline year). Setting alpha at 0.05 and power at 80%, we estimate the proportion 2 value to be 0.56 (indicating an effect size of only 4%). Both DM program and client would probably agree that this target is too low. However, if the analyst developed the model based on monthly values of ED visits instead ($1651/12 \approx 138$), holding the power at 80% and alpha at 0.05, the effect size would now be 15% (0.60–0.45 for proportions 1 and 2, respectively).

The rationale for choosing this new effect size for this measure is based on the following logic; ED visits is a metric that can be measured frequently and impacted by an intervention within a short period of time. Thus, the DM program should focus on the variability associated with patients presenting to the ED for avoidable acute exacerbations on a monthly basis (or even weekly or daily), as opposed to aggregating the data to annual values. This is in contrast to certain clinical indices where change can only be identified after longer time periods. Mammography screening rates is a good example of this, since it is recommended that women receive these screenings only every other year. As a result the N will be much smaller on clinical markers that require much longer time periods to acquire.

In summary, there are two issues that must be considered when determining how N will be used: (1) the frequency of data collection and (2) the impact of time for the specific clinical measure under study. Therefore, the DM program and client organization should judge the value in manipulating the N and time period under study, and agree to the appropriate effect size.

Setting a reasonable and achievable target

Determining an effect size using either clinical judgment or statistical modeling requires a reality check. Several factors must be considered before the outcome target is defined and the program held accountable for meeting it.

As established earlier in the paper (Figs. 3 and 4), a small N, in and of itself, will impact the equation heavily, creating a high effect size needed to meet power and statistical significance criteria. As a result, the effect size may be such that the ability to achieve it is unlikely.

Similarly, the proportion 1 value plays a significant role in achieving a reasonable effect size target. A low proportion 1 value indicates that there is much room to improve, especially if other similar organizations have achieved better scores. However, a very high proportion 1 value would make a significant effect size very difficult to achieve, (as well as investing a great deal of resources into an effort to achieve little return).

For example, let's assume we have a clinical measure with an N of 200, and we set power to 80% and alpha at 0.05. If the proportion 1 value is 0.10, the resulting proportion 2 target will be 0.19 (equal to an effect size of 9%). However, if the proportion 1 value was 0.90 (all other parameters being equal), then the resulting effect size would only be 6%. From both a practical and resource perspective, it would be much more difficult to improve the score from 0.90 to 0.96, than 0.10 to 0.19.

Ultimately, these variables must be considered in the context of the whole picture. To establish meaningful, reasonable, and achievable target effect sizes, both parties must agree on the criteria, given the information that they have available to them at the time, and based on the other elements of the contract that must be weighed accordingly. While this statistical model can churn out an effect size for any outcome metric, it is only valuable when used in conjunction with appropriate human judgment.

Similarly, the results must be acceptable to both parties as determined during the initial establishment of the targets. In other words, either the DM program meets the target, or it does not. The logic should be similar to that established in other research endeavors. If the p value is set to 0.05, and the outcome measure does not meet that criterion, the value is reported as not statistically significant. In DM programs this procedure should be followed as well in order to maintain the integrity of the measurement and evaluation system, and minimize the likely introduction of subjective interpretation.

CONCLUSION

This paper introduced the concept of using an empirical statistical model for establishing

effect size for clinical outcome metrics in DM programs. It was shown that there are four parameters used in the equation, some of which are within the control of the program analyst, and some that are not. For example, the analyst can establish power and significance levels (alpha), but has no control over proportion 1 values and little control over sample size (N).

A consequence of having this many model variables is that the equation can be manipulated to achieve the desired effect. Therefore, it is imperative that the parameter values chosen are logic-based, and mutually agreed upon by both the DM program and the client organization. The resulting model should establish an effect size that is meaningful, reasonable, and achievable.

REFERENCES

1. Linden A, Adams J, Roberts N. An assessment of the total population approach for evaluating disease management program effectiveness. Disease Management 2003;6:93–102.
2. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: an introduction to time series analysis. Disease Management 2003;6: 243–255.
3. Linden A, Roberts N, Keck K. The complete "how to" guide for selecting a disease management vendor. Disease Management 2003;6:21–26.
4. American Healthways Inc., John Hopkins. Standard outcome metrics and evaluation methodology for disease management programs, 1st ed. Palm Desert, CA: 2nd Annual Disease Management Outcomes Summit, 2002.
5. National Committee for Quality Assurance (NCQA). HEDIS 2003 Technical Specifications, Volume 2. Washington, DC: NCQA, 2002.
6. American Diabetes Association. Position Statement: Tests of glycemia in diabetes. Diabetes Care 2003;26: S106–S108.
7. Fisher ES, Wennberg DE, Stukel TA, et al. The implications of regional variations in Medicare spending. Part 1: The content, quality and accessibility of care. JAMA 2003;138:273–287.
8. Fisher ES, Wennberg DE, Stukel TA, et al. The implications of regional variations in Medicare spending. Part 2: Health outcomes and satisfaction with care. JAMA 2003;138:288–298.
9. Medicare + Choice Quality Review Organization for the Centers for Medicare and Medicaid Services (CMS). Quality Assessment Performance Improvement (QAPI): instructional guide. Bethesda, MD: CMS, 2001.
10. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Controlled Clinical Trials 1981;2:93–113.
11. Donner A. Approaches to sample size estimation in the design of clinical trials—a review. Statistics in Medicine 1984;3:199–214.
12. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272:112–124.
13. Fleiss JL. Statistical methods for rates and proportions, 2nd ed. New York: Wiley & Sons, 1981.
14. Feigl PA. Graphical aid for determining sample size when comparing two independent proportions. Biometrics 1978;34:111–122.
15. Cochran WG, Cox GM. Experimental designs, 2nd ed. New York: Wiley & Sons, 1957.
16. Cohen J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1969.
17. Snedecor GW, and Cochran WG. Statistical methods, 6th ed. Ames, IA: Iowa State University Press, 1967.
18. Haseman JK. Exact sample sizes for use with the Fisher-Irwin test for 2 × 2 tables. Biometrics 1978;34: 106–109.

Address reprint requests to:
Ariel Linden, Dr.P.H., M.S.
Linden Consulting Group
6208 NE Chestnut St.
Hillsboro, OR 97124

E-mail: ariellinden@yahoo.com

APPENDIX

Formulae for determining power and effect sizes

There are several ways to determine effect size given an assumed initial proportion (P1), sample sizes, alpha, and power.[13–17] Perhaps surprisingly, all of them require iterative procedures. One common method is to use a formula for the power when comparing two proportions. A brief derivation is included for the interested reader. This formula is based on normal approximations and should not be used if the sample size in either the pre or post period is below 30. In these cases, the exact calculations can be done with most sample size calculating statistical packages.

$$
\begin{aligned}
&\textit{Power} = \\
&= P\left((\bar{p}_2 - \bar{p}_1) - (p_2 - p_1) > z_{1-\alpha/2}\sqrt{p_1(1-p_1)/n_1 + p_1(1-p_1)/n_2} - (p_2 - p_1)\right) \\
&= P\left(\frac{(\bar{p}_2 - \bar{p}_1) - (p_2 - p_1)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} > \frac{z_{1-\alpha/2}\sqrt{p_1(1-p_1)/n_1 + p_1(1-p_1)/n_2} - (p_2 - p_1)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}\right) \\
&= 1 - \Phi\left(\frac{z_{1-\alpha/2}\sqrt{p_1(1-p_1)/n_1 + p_1(1-p_1)/n_2} - (p_2 - p_1)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}\right) \\
&= \Phi\left(\frac{(p_2 - p_1) - z_{1-\alpha/2}\sqrt{p_1(1-p_1)/n_1 + p_1(1-p_1)/n_2}}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}\right)
\end{aligned}
$$

Where Φ is the cumulative normal distribution function. Conceptually to determine the minimum proportion in the second period that would have adequate power you would solve this equation for proportion 2 (P2). This is actually quite difficult. In practice the analyst can try different values for P2 until the required power value is obtained, or create either a function or sub-routine using a "Do Until Loop" in a programming language such as Visual Basic. The following equation provides an example of how this can be developed in the case of equal sample sizes for the pre and post period:

$$
\textit{CommonN} = \frac{Z\alpha \times \sqrt{2 \times \mathrm{P1} \times (1 - \mathrm{P1})} - Z\beta \times \sqrt{\mathrm{P1} \times (1 - \mathrm{P1}) + \mathrm{P2} \times (1 - \mathrm{P2})^2}}{(\mathrm{P2} - \mathrm{P1})^2}
$$

Where: α = user defined, β = user defined, target N = user defined, $Z\alpha$ = normal inverse of $(1 - \alpha)$, $Z\beta$ = normal inverse of $(1 - \beta)$, and P2 = 0 + 0.00001 (loop until common N = target N).

Somewhat more precise formulae are available using more elaborate approximations or directly inverting the chi-squared test from 2 × 2 contingency tables.[18] Differences from these approximations are not substantial unless sample sizes are small or proportions are near zero or one. However, it is important that both parties agree on the method to be used.

In this paper, we use effect size to describe the difference in the scale of measurement (eg, difference in proportions). Note that some authors use effect size to refer to the difference divided by its standard error. We think of the use of that method as something of a holdover from before the wide availability of power software. This enabled the use of tables of power as a function of sample size and effect size to be relatively compact. We prefer the simple difference of proportions definition since it is easier for those with less familiarity with statistical power issues to understand. From a contractual point of view it also prevents a DM company from achieving significant results with larger sample sizes rather than an improvement in performance.