DISEASE MANAGEMENT
Volume 11, Number 2, 2008

DOI: 10.1089/dis.2008.1120019

Sample Size in Disease Management Program Evaluation: The Challenge of Demonstrating a Statistically Significant Reduction in Admissions

ARIEL LINDEN, Dr.PH., M.S.[1,2]

ABSTRACT

Prior to implementing a disease management (DM) strategy, a needs assessment should be conducted to determine whether sufficient opportunity exists for an intervention to be successful in the given population. A central component of this assessment is a sample size analysis to determine whether the population is of sufficient size to allow the expected program effect to achieve statistical significance. This paper discusses the parameters that comprise the generic sample size formula for independent samples and their interrelationships, followed by modifications for the DM setting. In addition, a table is provided with sample size estimates for various effect sizes. Examples are described in detail along with strategies for overcoming common barriers. Ultimately, conducting these calculations up front will help set appropriate expectations about the ability to demonstrate the success of the intervention. (*Disease Management*. 2008;11:95–101)

INTRODUCTION

Prior to implementing a disease management (DM) strategy, a needs assessment should be conducted to determine whether sufficient opportunity exists for an intervention to be successful in the given population. Under the assumption that a DM intervention will target a reduction in acute utilization, there are 3 analyses out of a broader possible set[1] that should be routinely conducted: A time series analysis of historic utilization rates in the target chronically ill population to determine if the level and trending patterns suggest an opportunity for reduction,[2] a number needed to decrease analysis to determine the number of hospital admissions that must be reduced by the program in order to achieve a return on investment,[3,4] and a sample size analysis to assess whether, given the expected reduction in acute utilization, the population is of sufficient size to allow for this program effect to achieve statistical significance.

While the first 2 analyses have been well described in the DM literature and are often used in practice, the third is commonly overlooked by those planning to implement a DM intervention. Conducting this calculation up front will help set appropriate expectations about the ability of the intervention to demonstrate successful outcomes. This paper offers readers the tools required to conduct sample size calcula-

[1]President, Linden Consulting Group, Hillsboro, Oregon.
[2]School of Medicine and School of Nursing, Oregon Health and Sciences University, Portland, Oregon

tions in advance of program implementation. It begins by describing the parameters that comprise the generic sample size model and their interrelationships. Sample size formulae modified for the typical characteristics of a DM population are then presented. A table is provided to help determine whether a DM program's estimated reduction in hospitalizations will reach statistical significance given the size of the population under management. Finally, examples are provided with discussion to assist readers in applying this methodology to their particular situation or setting.

MODEL PARAMETERS

In general, sample size is a function of 4 interrelated parameters: *significance level*, or alpha—the probability of finding a significant difference when there truly is none (ie, false-positive); *power*—the probability of not finding a significant difference when there truly is one (ie, false-negative); *effect size*—the magnitude of change between 2 groups or within 1 group, pre and post intervention; and the *standard deviation* (SD)—the degree to which observations are spread out around a mean in a given data set. While the first 2 parameters, significance level and power, are set by the researcher, effect size and standard deviation values are a function of the intervention under study. This section describes each of these parameters and their relevance to sample size.

Significance level (alpha)

In a test of statistical significance, we start with the assumption that there is no difference between 2 groups (or 1 group—pre and post intervention). Once the intervention is completed and a difference is observed, we need to assess whether there is convincing evidence that this is a statistically meaningful difference or simply due to chance. The Greek letter "alpha" refers to the probability of rejecting the null hypothesis when the null hypothesis is actually true (a type I error).[5] Therefore, an alpha of 0.05 (or 5%) indicates that 5 times out of 100 a difference is found between 2 groups (or 1 group pre and post intervention) that is due to chance—not the intervention. By setting the alpha level lower (eg, 0.01), we make the test more conservative, indicating that we are less willing to be wrong. Thus, while lowering the alpha decreases the chance of committing a type I error, the more stringent threshold criteria reduces the likelihood of concluding that the DM program had an effect. In population-based studies, the alpha is typically set at 0.05.[6]

Power

Like the significance level, power is established by the researcher independent of the intervention under study. While the alpha level is chosen to protect against false positives, sufficient power protects against false negatives (ie, concluding that there is no difference between the 2 groups when in fact there is one).[5] In other words, power is the likelihood that the results of the study will show a significant effect when there truly is one. A power of 80% means that 80 times out of 100 when there is a true intervention effect, we will identify it as such. Power increases with an increase in the sample size and effect size, as well as with 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.[6]

Effect size

Effect size refers to the expected difference in outcome measure (ie, year-over-year hospitalization rates) as a result of the intervention. The smallest effect size that can be detected statistically (ie, shown to be statistically significant) is determined by the size of the population under study, such that a larger sample size is required to detect a smaller effect size and vice versa. Similarly, a larger effect size will result in higher power.[6] Established DM vendors should be able to provide a good estimate of what effect size can be expected for commonly measured outcomes in a given population.

Standard deviation

Standard deviation is the most commonly used measure to denote the variability of data about the mean. When 2 group means are compared, the variability is typically expressed as

the *within group standard deviation,* an average of the variability within the 2 groups. Large SDs suggest the presence of extreme values and small SDs indicate that data tend to cluster around the mean. In data sets that are normally (or randomly) distributed, as the sample size grows the SD tends to get "pulled in" closer to the mean. That is, extreme values no longer appear as outliers in the presence of many more randomly distributed data points. Thus, larger, normally distributed samples have smaller SDs, which in turn increases power to detect statistically significant effect sizes.

However, rare event outcomes such as hospitalizations or emergency department visits do not follow a normal distribution (ie, the majority of values are amassed close to zero with only a few outliers). In using statistical methods designed for normal distributions to evaluate such outcomes we may inadvertently increase the likelihood of committing a type II error. Fortunately, outcomes expressed as a *rate* (ie, a number of events occurring in a population over a given period of time) generally follow a Poisson distribution[7] and we can use this distribution instead of the normal distribution to calculate SD. A benefit of the Poisson distribution is that the variance is equal to the mean (therefore the SD equals the square root of the mean). As will be described later, this convenient relationship allows us to determine the sample size necessary to achieve various effect size estimates for rare outcomes.

Effects of manipulating model parameters

To demonstrate how these parameters are related, Table 1 illustrates the independent impact of each model parameter on sample size. Column A represents a base case in which a 10% reduction from the starting value of 0.10 is expected with a within group SD of 0.308. By setting alpha at 0.05 and power at 80%, a sample size of approximately 14,903 is derived. Columns B through F show that, holding all else constant, a larger sample size is needed when accompanied by any of the following: lower starting value, smaller effect size, larger SD, or more stringent alpha and power levels. Note that Columns B and C result in an identical sample size. This is simply due to the fact that a 9% reduction from 0.10 is equal to a 10% reduction from 0.09.

In summary, this section established that there are 4 interrelated parameters that contribute to the determination of sample size: the effect size, calculated as the change in value of 1 group in a pre-post or the difference in mean value of the treatment group compared with the control; the standard deviation, the variability around each group's mean, calculated using the appropriate distribution (eg, normal, Poisson); and the levels of alpha and power, which are characteristics of the test of statistical significance and are determined by the researcher. Readers interested in a more comprehensive discussion of the individual parameters and their interactions can refer to Donner,[8] Lachin,[9] and Moher et al.[10]

DETERMINING SAMPLE SIZE

In the appendix, formulae are provided for determining sample size mathematically. Similarly, Table 2 provides population size estimates based on various baseline admission

Table 1. The Impact of Changing Individual Model Parameters on Sample Size*

	A	*B*	*C*	*D*	*E*	*F*
Starting value	0.10	0.09				
Effect size (%)	10.0		9.0			
Standard deviation	0.308			0.339		
Alpha	0.05				0.01	
Power	0.80					0.90
Sample size	14,903	18,398	18,398	18,019	22,175	19,950

*Column A represents the base case and all other columns reflect the changed parameter and its influence on sample size.

Table 2. Estimated Sample Sizes Based on Starting Admission Rate Per Person and Predicted Effect Size (Δ), Assuming Alpha of 0.05, 80% Power and a Two-Tailed Test

	Starting admission rate per person									
Δ	*0.10*	*0.09*	*0.08*	*0.07*	*0.06*	*0.05*	*0.04*	*0.03*	*0.02*	*0.01*
1%	1,561,917	1,735,464	1,952,397	2,231,310	2,603,195	3,123,834	3,904,793	5,206,391	7,809,586	15,619,172
2%	388,510	431,677	485,637	555,014	647,516	777,019	971,274	1,295,032	1,942,548	3,885,096
3%	171,793	190,881	214,742	245,419	286,322	343,587	429,483	572,644	858,966	1,717,933
4%	96,139	106,821	120,173	137,341	160,231	192,278	240,347	320,463	480,694	961,388
5%	61,211	68,012	76,514	87,445	102,019	122,422	153,028	204,037	306,056	612,112
6%	42,287	46,985	52,858	60,409	70,478	84,573	105,717	140,955	211,433	422,866
7%	30,905	34,339	38,631	44,150	51,508	61,810	77,262	103,016	154,524	309,048
8%	23,536	26,152	29,421	33,623	39,227	47,073	58,841	78,455	117,682	235,364
9%	18,498	20,553	23,122	26,425	30,829	36,995	46,244	61,659	92,488	184,976
10%	14,903	16,558	18,628	21,289	24,838	29,805	37,256	49,675	74,513	149,025
11%	12,249	13,610	15,312	17,499	20,416	24,499	30,624	40,831	61,247	122,494
12%	10,237	11,374	12,796	14,624	17,061	20,473	25,592	34,122	51,184	102,367
13%	8,674	9,638	10,843	12,392	14,457	17,349	21,686	28,914	43,372	86,743
14%	7,438	8,264	9,297	10,626	12,396	14,876	18,595	24,793	37,189	74,379
15%	6,443	7,159	8,054	9,204	10,738	12,886	16,107	21,476	32,214	64,429
16%	5,631	6,256	7,038	8,044	9,384	11,261	14,077	18,769	28,153	56,307
17%	4,959	5,510	6,199	7,085	8,266	9,919	12,398	16,531	24,797	49,593
18%	4,398	4,887	5,498	6,283	7,330	8,796	10,995	14,660	21,991	43,981
19%	3,924	4,360	4,906	5,606	6,541	7,849	9,811	13,081	19,622	39,244
20%	3,521	3,912	4,401	5,030	5,868	7,042	8,803	11,737	17,605	35,211

rates (per person per year) and predicted effect sizes, assuming an alpha of 0.05 (two-sided test), 80% power, and a within group SD calculated using formula 3. This table can serve as a reference to assess whether a DM program has sufficient opportunity to demonstrate a statistically significant decrease in admissions based on the size of the available population.

To demonstrate the applicability of this model to DM, admission rate estimates are used from the 2004 National Hospital Discharge Survey[11] for 5 of the primary conditions traditionally managed by DM: acute myocardial infarction, congestive heart failure, asthma, chronic obstructive pulmonary disease, and diabetes. The baseline discharge rate for this set of diagnoses is 116.4 per 10,000 population, or 0.0116 discharges per person per year. Assuming an effect size of 10% (a 0.0012 decrease), the target discharge rate per person is 0.0105, and the within group SD is 0.105 (per formula 3 in the Appendix). Using the typical alpha (.05) and power (.80) levels, the sample size formula is:

$$n \approx 2(0.105)^2 \frac{(1.96 + 0.84)^2}{0.0012^2} = 128{,}028$$

When a program is implemented and evaluated at the population level, the sample size is synonymous with the population size. When a program is implemented and evaluated using a treatment and a control group, the sample size calculation estimates the number of subjects needed in each group. Thus, in this example a population of 128,028 is required for a 10% reduction in admissions per person to be statistically significant. In a pre-post design, this means that the population size must be 128,028 in both the pre and the post periods; in a design with a concurrent control group, the treatment group must be 128,028 and the control group must also be 128,028.

ADDITIONAL CONSIDERATIONS IN DETERMINING SAMPLE SIZE

If the population size for a given intervention is fixed or is smaller than that which the formulae produce, statistical significance can still be preserved by manipulating the baseline rate. Three approaches are discussed in this section along with some additional detail re-

garding implementing the sample size calculation in practice.

Unit of measure

In contrast to the *mean* (ie, a value expressed solely in relation to the count of the observations), a *rate* is a value expressed relative to a population size that is defined by the investigator (eg, per thousand members, per hundred thousand population). Given this latitude in how a rate can be expressed, selecting the appropriate unit of measure is of utmost importance. There are 2 primary reasons why it is recommended that sample size calculations be performed at the individual level (eg, admissions per person per unit of time): (1) in population-based analyses this basis of measure is less influenced by population turnover[12] and length of enrollment (for this reason, health insurers report on a per-member-per-month or per-year basis), and (2) this is the standard method of reporting outcome comparisons between 2 or more groups (cohorts).

Baseline rate and effect size

Unless the DM program has identified opportunities to improve its ability to identify and intervene upon participants at high risk of a near-term hospitalization, it is difficult to manipulate effect size. However, as indicated in Table 2, for any given sample size a larger baseline rate requires a smaller effect size to achieve statistical significance. One strategy is to calculate estimates for the disease-specific population as opposed to creating estimates for the entire population. This distributes the same number of admissions over a more narrowly defined population, resulting in a larger baseline rate, which in turn requires a smaller effect size to achieve statistical significance.

Another reasonable approach to increase the baseline rate is to include *complications* of the primary diagnosis[13] in the calculation of the outcome of interest (ie, admissions). A complication refers to any diagnosis that is physiologically related to the primary chronic condition and thus can theoretically be impacted by the program. Conversely, including admissions completely unrelated to the conditions under management introduces bias into both the sample size calculation and the subsequent program evaluation.

A final approach to increase baseline rate is to extend the measurement period. DM programs generally operate under 1- to 3-year contracts. Thus, assuming a baseline admission rate of 0.05 per person per year and a 5% projected decrease, a population size of 122,422 would be required for this decrease to be statistically significant (Table 2). However, by holding the effect size at 5% and extending the duration from 12 to 18 months (and hence the rate from per person per year to per person per 18 months) the baseline rate is approximately 0.08 (0.05 spread out over 18 months instead of 12 months), which requires a substantially smaller population of 76,514.

In this article, the term effect size is used to describe the difference in the scale of measurement (eg, difference in means or rates). Some authors use effect size to refer to the difference divided by its standard error.[14] The use of such a method is a holdover from the days in which statistical software to calculate power was not widely available. This method enabled the use of streamlined tables of power as a function of sample size and effect size. The simple difference of means is the preferred definition as it is easier for those with less familiarity with statistical power to understand and, more importantly, it ensures that significant results are not achieved purely as a result of large sample sizes.[6]

CONCLUSION

Sample size estimation is an integral component of the needs analysis that should be conducted prior to the implementation of any DM strategy. There are several interrelated parameters that contribute to the determination of sample size of which only alpha and power are under the control of the researcher. The baseline admission rate and its respective SD are characteristics of the population under study, and the effect size is a function of the intervention. If the population is found to be too small, the baseline admission rate can be increased by restricting the population to those with the disease, expanding the list of admit-

ting diagnoses to include complications of the primary chronic condition under management, or extending the study time frame in which the effect size is intended to be achieved. Ultimately it is imperative that the formula parameters are logic based and the ramifications understood by both the DM program and the client organization.

While more precise formulae are available,[15,16,17] they do not differ substantially from the approximations described above and require statistical expertise to implement.

ACKNOWLEDGMENT

The author would like to thank John Adams, Ph.D., from the RAND Corporation and Julia Adler-Milstein from the Ph.D. program in Health Policy at Harvard University for their invaluable assistance in conceptualizing the issues and editing the manuscript.

The author has no conflicts of interest to disclose.

REFERENCES

1. Rossi PH, Freeman HE, Lipsey MW. *Evaluation: A Systematic Approach.* 7th ed. Thousand Oaks, CA: Sage Publications; 2003.
2. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: An introduction to time series analysis. *Dis Manag*. 2003;6:243–255.
3. Linden A. What will it take for disease management to demonstrate a return on investment? New perspectives on an old theme. Am J Manage Care. 2006; 12:217–222.
4. Linden A, Biuso TJ. In search of financial savings from disease management: Applying the number needed to decrease (NND) analysis to a diabetic population. Dis Manage and Healt Outc. 2006;14:197–202.
5. Vogt PW. *Dictionary of Statistics and Methodology: A Non-Technical Guide for the Social Sciences*. 2nd ed. Thousand Oaks, CA: Sage Publications; 1999
6. Linden A, Adams J, Roberts N. Using an empirical method for establishing clinical outcome targets in disease management programs. Dis Manag. 2004;7:93–101.
7. Fleiss JL, Levin B, Myunghee CP. *Statistical Methods for Rates and Proportions.* 3rd ed. Hoboken, NJ: John Wiley & Sons; 2003.
8. Donner A. Approaches to sample size estimation in the design of clinical trials—a review. *Stat Med.* 1984; 3:199–214.
9. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Control Clin Trials.* 1981;2:93–113.
10. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA*. 1994;272:112–124.
11. DeFrances CJ, Podgornik MN. *2004 National Hospital Discharge Survey. Advance Data from Vital and Health Statistics. No 371*. Hyattsville, MD: National Center for Health Statistics; 2006
12. Linden A, Goldberg S. The case-mix of chronic illness hospitalization rates in a managed care population: Implications for health management programs. J Eval Clin Pract. 2007;13:947–951.
13. Linden A, Biuso TJ, Gopal A, et al. Consensus development and application of ICD-9 codes for defining chronic illnesses and their complications. Dis Manage and Healt Outc. 2007;15:315–322.
14. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Hillsdale, NJ: Erlbaum; 1988.
15. Dupont WD, Plummer WD. Power and sample size calculations: A review and computer program. *Control Clin Trials.* 1990;11:116–128.
16. Pearson ES, Hartley HO. *Biometrika Tables for Statisticians.* Vol I, 3rd ed. Cambridge, UK: Cambridge University Press;1970.
17. Ott RL. *An Introduction to Statistical Methods and Data Analysis.* 4th ed. Belmont, CA: Duxbury Press; 1993.
18. Gail M. Power computations for designing comparative Poisson trials. *Biometrics.* 1974;30:231–237.
19. Shiue W-K, Bain LJ. Experiment size and power comparisons for two-sample Poisson tests. *Appl Stat*. 1982; 31:130–134.
20. Thode CH. Power and sample size requirements for tests of differences between two Poisson rates. *Statistician.* 1997;46:227–230.

APPENDIX

Equation (1) illustrates a generalized sample size formula for comparing 2 independent samples. Due to the algebra, the equation may take on a slightly different form depending on the reference source[18,19,20]

$$n = 2\sigma^2 \frac{(Z_{\alpha/2} + Z_{\beta})^2}{\Delta^2} \qquad (1)$$

where σ denotes the *within group standard deviation*, $Z_{\alpha/2}$ is the Z-score from the normal distribution representing the critical value of alpha ($Z = 1.96$ for an alpha of 0.05 using a two-tailed test), Z_{β} is the Z-score from the normal distribution representing the power ($Z = 0.84$ for a power of 0.80), and Δ is the effect size.

In novel prospective studies the SD is not known and therefore must be estimated. Often, the researcher relies on values derived from similar studies or may perform a pilot study to provide estimates. Many times only 1 SD is known and that value is then used as a proxy for both groups' estimates. As described earlier, count or rate data typically assume a Poisson distribution in which the variance is equal to the mean value. This allows the estimate of SD to be:

$$\sigma = \mu \tag{2}$$

where μ is the mean value for a given group. Therefore, the within group SD can be expressed as the average between the 2 standard deviation estimates:

$$\bar{\sigma} = \frac{\sqrt{\mu_1} + \sqrt{\mu_2}}{2} \tag{3}$$

where the subscript references the group being compared (or time period under a pre-post assessment).

Address reprint requests to:
Ariel Linden, Dr.PH., M.S.
President, Linden Consulting Group
6208 NE Chestnut Street
Hillsboro, OR 97124

E-mail: alinden@lindenconsulting.org