

EVALUATING DISEASE MANAGEMENT PROGRAM EFFECTIVENESS: AN INTRODUCTION TO THE BOOTSTRAP TECHNIQUE

Ariel Linden, Dr.P.H., M.S.,¹ John L. Adams, Ph.D.,² Nancy Roberts, M.P.H.³

¹ President, Linden Consulting Group, Portland, OR. ariellinden@yahoo.com

² Senior Statistician, RAND Corporation, Santa Monica, CA. adams@rand.org

³ Regional Director of Integrated Performance/Six Sigma Champion, Providence Health System, Portland, OR. nancy.roberts@providence.org

Summary: The bootstrap technique is introduced and examples provided of how it can be used to determine DM program effectiveness and develop estimates for the population.

Corresponding Author:

Ariel Linden, DrPH, MS

President, Linden Consulting Group

6208 NE Chestnut Street

Hillsboro, OR 97124

503-547-8343

ariellinden@yahoo.com

Funding Source: None

Potential Conflicts of Interest: None

Condensed Running Title: Using Bootstrapping in DM Program Evaluation

EVALUATING DISEASE MANAGEMENT PROGRAM EFFECTIVENESS: AN INTRODUCTION TO THE BOOTSTRAP TECHNIQUE

Ariel Linden, Dr.P.H., M.S.,¹ John L. Adams, Ph.D.,² Nancy Roberts, M.P.H.³

¹ President, Linden Consulting Group, Portland, OR. ariellinden@yahoo.com

² Senior Statistician, RAND Corporation, Santa Monica, CA. adams@rand.org

³ Regional Director of Integrated Performance/Six Sigma Champion, Providence Health System, Portland, OR. nancy.roberts@providence.org

ABSTRACT

Introduction: Disease management (DM) program evaluations are somewhat limited in scope due to typically small sample sizes comprising important subsets of the treated population. Identifying subsets of the data that have differing results from the aggregate of the whole program can lend insight into where, when, and how the program achieves its results. Additionally, there is a very limited set of classical tools available for the smaller sample sizes typically encountered in DM. Without readily available standard error and confidence interval calculations, the analyst may be fooled by specious details.

Methods: A method called the “bootstrap” is introduced as a suitable technique for allowing DM program evaluators to use a broader array of quantities of interest and to extend inferences to the population based on results achieved in the program. The bootstrap uses the power of modern computers to generate many random samples from a given dataset, allowing the use of repeated samples’ statistic (e.g. mean, proportion, median). Using a congestive heart failure program as an example, the bootstrap technique is used to extend a DM program evaluation beyond questions addressed using classical statistical inference; (1) how much of a median cost decrease can be expected as a result of the program? (2) did the program impact the highest and lowest

costing members equally, and (3) how much of a decrease in the proportion of patients experiencing a hospitalization can be expected as a result of the program?

Results: The potential advantages of the bootstrap technique in DM program evaluation were clearly illustrated using this small CHF program example. A more robust understanding of program impact is possible when more tools and methods are available to the evaluator. This is particularly the case in DM, which is inherently biased in case-mix (e.g. strive to enroll sickest first), often has skewed distributions or outliers, and may suffer from small sample sizes.

Conclusions: The bootstrap technique creates distributions that allow for a more accurate method of drawing statistical inferences of a population. Moreover, since classical statistical inference techniques were designed specifically for parametric statistics (i.e. assuming a normal distribution) the bootstrap can be used for measures that have no convenient statistical formulae. Additionally, confidence intervals can be defined around this statistic, making it a viable option for evaluating DM program effectiveness.

Key Words: Disease management, parametric statistics, non-parametric statistics, bootstrap, standard error, confidence interval.

INTRODUCTION

Although disease management (DM) has been in existence for over a decade, there is still much uncertainty as to its effectiveness in improving health status and reducing costs. Part of the struggle to gain legitimacy is the ambiguity in how to best evaluate DM program effectiveness. The most commonly used method for evaluating financial outcomes in DM is a standard pre-post model^[1]. Using this approach, the population's average costs (typically expressed as per-member-per-month [pmpm]) attained in the program year is compared to average costs in the baseline year. After adjustments have been made for vendor fees, etc. a return of investment (ROI) is determined. This method's lack of a control group makes it vulnerable to a myriad of biases and subject to the problem of regression to the mean. Given that DM programs will most always be subject to selection bias, and that typical evaluation designs will be observational as opposed to experimental, techniques that can help control for threats to internal validity while at the same time allow generalization of program outcomes to the principal population should be used where possible. This paper continues a series by these authors of such alternative techniques for DM evaluation^[2-8].

One way to gain confidence in a DM program is to open the "black box" of the PMPM calculation and explore the underlying data. Identifying subsets of the data that have differing results from the aggregate of the whole program can lend insight into where, when, and how the program achieves its results. For example, are the savings across the board or are they concentrated in more expensive cases? Is the mode of the savings different in different subsets of the data? Does one group save on hospitalizations while another saves on ER visits? Perhaps most importantly, do the

variations in savings in subsets of the managed population make logical sense given the principles and features of the DM program?

This sensible exploratory data analysis^[9] is commonly done by good analysts. What may limit the analyst's progress is the limited set of classical tools available for the smaller sample sizes typically encountered in DM. Ideally the analyst would be free to identify the quantities that they find the most informative rather than select from the limited menu of classical statistics (e.g. the mean). Particularly important given the small sample sizes often found in DM these quantities of interest should have readily available standard error and confidence interval calculation to protect the analyst from being fooled by specious details.

Fortunately a method called the "bootstrap" is a very suitable technique for allowing DM program evaluators to use a broader array of quantities of interest and to extend inferences to the population based on results achieved in the program. The bootstrap is a data-based simulation method for statistical inference that was introduced by Bradley Efron in 1979^[10] and regularly improved upon and summarized in a book in 1993.^[11] Since then several additional excellent books have been written on the bootstrap procedure.^[12-14] This paper will introduce the reader to the bootstrap technique and provide examples of how it can be used to determine DM program effectiveness and develop estimates for the population.

CLASSICAL STATISTICAL INFERENCE

In order to provide a context for the use of the bootstrap, a brief history lesson is helpful. One of the early uses of statistical inference, (e.g. generalizing to the population

based on results gleaned from a sample), was by William Sealey Gossett (1876-1937) in the early 20th century.^[15] Better known under the pseudonym “Student” (as an employee of Guinness Brewery Co., he was not allowed to publish under his own name or affiliation), Gossett identified that one can estimate the mean and standard error of a normally distributed population when the sample size is relatively large and the standard deviation is unknown. This gave rise to the Student t-distribution, also called the bell curve or normal curve (FIG 1). The limitations of methods of this type are that approximations (a) rely on the assumption that the data are normally distributed, (b) achieve better accuracy in large samples than in small ones, and (c), were originally developed for a small set of distributions and a limited class of sample statistics and are therefore not applicable to all situations.

The limitations of this “classical” approach are a factor when developing evaluation designs in DM. DM program data are often not normally distributed (thereby making the mean highly influenced by outliers) and may also have relatively small sample sizes. While the mean is an important measure in most DM program evaluations, there are two other measures of central tendency, the *median* and *mode*, which may also provide valuable evaluative information yet are rarely considered. Understanding the variability in the data is also important both in designing intervention approaches and in describing program effects. Classical statistical methods can be complicated and assumption dependent when estimating parameters such as standard error (SE) or confidence intervals (CI) for medians or modes.

The ability to consider several measures of central tendency and their distribution parameters allows for the development of a more complete, and likely, accurate

evaluation of DM program outcomes. Incorporation of the bootstrap technique allows the evaluator to overcome some of the limitations of the “classical” approach and add median, mode and their distribution parameters to his arsenal of evaluative tools.

BASIC STATISTICAL METRICS

Since the basic statistics used for both the “classical” (i.e. Student t-distribution) and bootstrap statistical approaches are similar, both are briefly reviewed. The equations for these statistics can be found in Appendix A.

Mean:

The mean (also referred to as the average or point estimate), is the most universally used statistic in almost every setting. It is inarguably the most familiar measure of central tendency, and it is readily used in conjunction with other statistical measurements. That said, the mean is easily biased by outliers in the data set, which when left unchecked, may provide misdirection in the interpretation of the results.

Standard Deviation:

While the mean offers information about the central tendency of the data, it does not provide information as to the data’s dispersion in the dataset. The most commonly used measure of variation is the standard deviation. A large standard deviation indicates that the data are dispersed far from the mean, or that the data contain outliers.

Mode:

Mode is simply the value that appears most frequently in a dataset. There may be more than one mode in a set of observations, as is commonly found in dichotomous

data. A unique mode may not exist; this is true if all the observations occur with the same frequency.

Median:

The *median* is the middle value of a dataset. In other words, 50% of the values lie above the median and the other 50% lie below it. Although the sample mean is by far the most commonly used measure of the central location in healthcare data, the sample median is a more robust measure as it is not as affected by outliers. Similarly, the median is much better suited for very skewed data sets than the mean. The reason that census data is reported in terms of the median and not as the mean, is because the median, being the halfway point, is better at reflecting the common experience than the mean.

Standard Error:

The standard error (SE) is an estimate of the variation of the sampling distribution of a given statistic (i.e. mean, median, proportion, etc.). Using the mean as an example, the SE estimates the standard deviation of the sample mean based on the population mean. The SE is an important statistic because it is used for both significance testing as well as in the construction of confidence intervals. The SE typically decreases as sample size increases.

Confidence Intervals:

A confidence interval (CI) gives an estimated range of values within which the unknown population parameter may lie. Using the mean as an example, based on the sample data, an estimated range of values can be calculated within which (with a given level of confidence) that the population mean may exist. Confidence intervals are

typically calculated so that the “level of confidence” is 95%, but other levels can be produced for the unknown parameter. The CI is given as the mean with the lower and upper confidence intervals. The width of the CI generally gives some insight as to the accuracy of the estimate. A wide interval may indicate large variability in the dataset or may be a result of having a very small N.

Confidence intervals (CI) are more useful than just the mean because they provide a sense of how far that estimate might truly extend. For example, using the standard DM model an estimate that a DM program will reduce pmpm costs by 10% in the first year (in other words, reducing the average cost by 10%) can be calculated. By adding a 95% confidence interval, that estimate can be qualified by saying (with 95% confidence) that the pmpm cost will be reduced by 10%, give or take 3%. Or in other words, the pmpm costs have a 95% chance that the true will be between 7% and 13%.

Given the tremendous variability in outcomes achieved in a typical DM program population (due to the small number of enrolled members, tremendous variability in costs and utilization, variable severity levels, etc.) it makes more sense to look at the program results as a function of both the mean and its confidence intervals rather than strictly looking at just the mean difference (e.g. changes in pmpm costs). According to the Disease Management Purchasing Consortium LLC ^[16] only about 5 contracts in a hundred currently use this method in their evaluation.

PRINCIPLES OF THE BOOTSTRAP

Compared to the classical method of statistical inference, the theory and practice behind the bootstrap technique is quite straightforward. In simple terms, the bootstrap

uses the power of modern computers to generate many random samples from the given dataset, allowing the use of repeated samples' statistic (e.g. mean, proportion, median, etc.) to generate variation in population estimates. Implicit is the assumption that this dataset represents the characteristics of the real population as much as possible. Given the speed of today's microprocessors, 10,000 simulations can be run within a matter of a few minutes. The term bootstrap is thought to have come from the phrase "to pull yourself up by your own bootstraps," ^[17] meaning that the analyst should rely on his or her own data to derive the statistics required for drawing statistical inference instead of relying on mathematical assumptions for estimating population parameters. Like other nonparametric techniques (e.g. rank-sum statistics) the bootstrap avoids making assumptions about the data. The breakthrough of the bootstrap is that it is a nearly universal technique; it can produce standard errors for almost any quantity of interest. In this way it is a continuation of the traditions of the jackknife and replicate based survey sampling methods.

The bootstrap technique entails drawing a defined number of random samples from the original dataset (which in and of itself is a sample from the population). Since the number of data points within the original dataset is limited, sampling is done with replacement. In other words, once a data point is randomly chosen and assigned to the new sample, it is replaced into the original dataset, so that it has a chance of being reselected for that sample and for all subsequent samples. Table I illustrates how this action is actually performed, using two dice as the random number generator. It is pretty clear that spots cannot be removed from the dice, so dice are an excellent example of sampling with replacement.

Figure 2 illustrates what happens when two dice are rolled 1000 times. As the figures shows, a normal distribution develops around the mean (a value of 7). Interestingly enough, in the game of craps, if either the “come-out” roll or the subsequent roll is a 7, the game is over. The number 7 was chosen as the losing number because at some point in history it was determined to have the highest likelihood of being rolled repeatedly.

The standard error of the distribution of the dice totals is easily derived as well. In the bootstrap simulation, the standard deviation of the distribution of values (e.g. the SD of the 1000 simulations), is in fact the SE. In the case of the two dice, the SD, and hence the SE, was 2.4.

Similarly, confidence intervals can be easily extracted from the sampling distribution. If the values are sorted from low to high, the values representing the 2.5th percentile and the 97.5th percentile represent the lower and upper 95% confidence intervals (2.5% on each tail, respectively). Using this procedure, the mean and 95% confidence intervals are determined to be: 7 (2, 12). Thus one could expect, with 95% confidence, any roll of two dice simultaneously will elicit a value between 2 and 12. Note that there are more sophisticated versions of the bootstrap that achieve even more precise intervals with the available data. These refinements are discussed in Efron and Tibsharoni,^[11] however, they are not essential to understand and use the method. Although the bootstrap does not require normal assumptions like some classical methods it does require a sample that represents the population well enough to support inference. How large a sample is needed depends on how skewed the population distribution is and how challenging it is to estimate the quantity of interest. Obviously,

estimating the standard error for the 90th percentile of a skewed distribution with 10 data points is not a good idea. Since there are no universal rules for the limitations of the bootstrap, analysts should use their judgment when dealing with very small sample sizes (e.g. less than 20), skewed distributions (e.g. data sets with a few very large outliers), or quantities that depend on a subset of the dataset (e.g. extreme percentiles.) All analyses and histograms reported in this paper were generated using Resampling Stats^[18] for Excel.

EXAMPLES FROM DISEASE MANAGEMENT

In the following examples data from Linden et al^[4] is used. In that study, the 1st year outcomes of a congestive heart failure (CHF) DM program were evaluated using propensity scoring to match controls to program participants.

Figure 3 shows the results in total costs between the DM program participants and their matched controls. The propensity score is used as a method for matching cases and controls to baseline characteristics so that both groups can be considered comparable. As illustrated, there was no significant difference between the groups in baseline mean costs. However, a highly significant difference ($p = 0.003$) was realized at the end of the program year, with the DM program group experiencing an average drop of \$6413 in costs while at the same time the control group exhibited an average increase of \$7084.

Using the bootstrap several additional questions using the data can be answered, as well as inferences made to the population from where they were drawn.

The bootstrap will be used in the following three examples. This technique is especially suited to these data because each group had small sample sizes ($N = 94$).

Question 1: How much of a median cost decrease can expect to experience as a result of the program?

The median costs are used in this example for two reasons, (1) because classical statistical methods do not supply measurement parameters such as SE or CI for medians, and (2) because the median may be a more appropriate metric due to the high variability and extreme outliers observed in these data. Note that the bootstrap could be used for trimmed means or log transformations or other robust measures of interest including any outlier rejection rule that you could write as a computer program.

To answer this question, perform the following steps; (1) compute the difference in pre and post costs of each of the 94 program participants (pre costs – post costs = difference score), (2) use the bootstrap to create 1000 samples of 94 randomly selected participant's difference scores, (3) find the median, SE and CI for the difference score.

Table II illustrates steps 1 and 2, providing data for only 10 of the 94 participants (this was done solely due to space limitations, not because of procedural requirements). Similarly, 5 samples are presented out of the total 1000 that were randomly drawn from the differences field. Since sampling with replacement was used, each of the 94 participants has equal opportunity to be chosen multiple times for each sample. For example, Table II shows that participant number 4 (difference score \$221) is present three times in samples 1 and 5, once in sample 4 and is not found in samples 2 and 3. Upon drawing the 1000 samples, the median of the distribution can be determined using

equation 2 (see Appendix A), where x_1 through x_N are all the individual medians of the bootstrap samples, and $N = 1000$. Similarly, the SE can be computed as the SD of that distribution. Finally, the 2.5th and the 97.5th percentiles can be calculated to determine the 95% confidence intervals. Our results using the bootstrap elicited the following statistics: Median difference = \$3100, SE = \$1607, Lower 95% CI = \$249, Upper 95% CI = \$6897. These results indicate that the median and SE are significantly lower than the mean and SE calculated using the classical method with mean = \$6413, and SE = \$2403. This discrepancy should be expected with a data set containing many extreme outliers. Similarly, financial executives should feel more comfortable with these more conservative numbers, unbiased by the outliers. The bootstrap outcomes can be stated as follows; “we are 95% confident that CHF patients enrolled in a DM program for one year will reduce their median costs from between \$249 and \$6897.” A histogram of the distribution of the bootstrap samples is shown in Figure 4.

The results of this analysis illustrate the importance of considering confidence intervals in the evaluation of DM program outcomes. In the scenario above, the distribution of individual results was quite wide, ranging from over \$25,000 to cases where this difference was -\$17,955. Likewise, as shown in Figure 4, the 1000 samples of median difference scores ranged from \$992 to \$11,248. This variability in results drives the wide confidence interval of the median difference \$249 to \$6,897 (however, the CI for the *mean* difference was \$1,860 to \$11,195 - much larger than that of the median). This example provides support for considering the use of the median and confidence intervals when the data are markedly skewed or have extreme outliers.

A different and more appropriate way of constructing this analysis would be to compare the difference in pre and post costs of DM program cohort to the difference in pre and post costs of the control group. This method is referred to as the difference-in-differences estimator (DID). As the data in Figure 3 and Table III suggest, costs decreased in the DM program and actually rose in the control cohort (whether calculated as the mean [FIG 3] or median [Table III]). The DID method provides a more accurate account of the effect on the entire population because, as illustrated in Table III, an adjustment is made for the increasing trend effect that occurred in the untreated group during the same time period. Bootstrapping the DID 1000 times gave the following results: Median DID = \$4,956, SE = \$1,540, Lower 95% CI = \$2,141, Upper 95% CI = \$7,697. These outcomes can be stated as follows; “we are 95% confident that the median difference in pre and post costs of DM program cohort as compared to the control group will be between \$2,141 and \$7,697.” A conclusion can be drawn from these results that the DM program was able to impact both the costs and rising trend of CHF costs.

Question 2: Did the program impact the highest and lowest costing members equally?

Considering that DM program interventions specifically target those behaviors that are high cost (e.g. hospitalization, ED visit) it can be assumed that high cost outliers would be reduced as a result of the program intervention. Similarly, success in the lower costing members may be indicated by a dampening of an upward trend in costs or utilization over time. This example assesses the program impact at the highest and lowest initial cost quartiles.

The 94 members of each group were assigned to quartile rankings according to their pre-program costs. Thus, the 24 lowest initial costing members in each group comprised the 25th percentile, and the 24 highest initial costing members in each group encompassed the 75th percentile. The difference between pre- and post program costs was determined for each member and the mean difference across each quartile and cohort was calculated. The mean DID was calculated by subtracting the control group's mean difference from the program participant's mean difference, for the 25th and 75th percentiles respectively, resulting in two values to be bootstrapped. The bootstrap is ideal in this situation because of the small number of samples (24 for each quartile) and the large variability between individual values. The results are shown in Table IV.

As illustrated, the mean difference in both the 25th and 75th quartile was positive (\$15,025 and \$15,115 for the 25th and 75th quartile, respectively). A positive value indicates that the cases experienced a greater reduction in costs than did the controls (because the calculation was based on cases – controls). More specifically, in the 25th quartile (those members with the lowest baseline year initial costs) program participants had, on average, a \$15,025 greater reduction in costs than controls (with 95% CI between \$3,103 and \$30,236). We see similar results in the 75th quartile. However, since the lower CI crosses 0, the difference is not statistically significant.

These results can be interpreted as follows: it appears that in the lowest quartile (i.e. those members who had the lowest costs in the baseline year), DM program participants were able to demonstrate smaller increases in costs overall than the control group (by about \$15,000 on average). In the 75th percentile (those members who had the highest healthcare costs in the baseline period), program participants achieved

lower costs than their controls (by about \$15,000). However, looking at the 95% confidence intervals shows that the groups in the 75th percentile did not differ significantly because of the variability around the mean.

Question 3: How much of a decrease in the proportion of patients experiencing a hospitalization can be expected as a result of the program?

As discussed earlier, classical statistics were developed for continuous variables, where the mean and SD are the main parameters under study. As such, the preceding example could have achieved similar results had it been computed using the equations found in Appendix A. This example will describe a statistic that achieves better results when using the bootstrap method as opposed to the classic calculations – proportions.

For this example the hospitalization data from Linden et al.^[4] was re-characterized so that a CHF patient was assigned a score of 0 if he or she had no hospitalizations, and was assigned a score of 1 if he or she had any hospitalizations (regardless of how many). Table V provides the results of the analysis. As shown, there was a 29% decrease in the proportion of the DM program participants who experienced a hospitalization in the program year from the baseline period. Similarly, there was a 6% decrease in the proportion of controls that experienced a hospitalization during the program year. The difference-in-difference estimator was used to adjust for the divergence in scores between the two groups.

The difference-in-difference statistic was bootstrapped 1000 times using sample sizes of N=100. The following results were achieved: a 24% mean reduction in the proportion of patients experiencing a hospitalization, with confidence intervals of 11%

and 37% (lower and upper 95% CI, respectively). Figure 5 presents the histogram of the bootstrap samples DID scores. These results can be restated as follows; “we are 95% confident that the DM program can lead to a reduction of between 11% and 37% in the proportion of CHF patients experiencing a hospitalization.”

DISCUSSION

This paper has demonstrated the utility of the bootstrap technique in drawing statistical inferences about the population using three scenarios very relevant to DM program evaluation. There are several reasons why DM program evaluators should consider using bootstrapping in lieu of more standard methods of inference.

Firstly, DM programs are inherently biased in their case-mix. They typically strive to enroll the sickest members first resulting in an enrolled cohort that does not accurately represent the population from whence they were drawn. Moreover, this participant cohort is usually quite small. These two factors alone can lead to tremendous variability in the outcome metrics when measured at the aggregate level. Because it is unknown if these data in fact follow a normal distribution, using classical metrics may provide erroneous estimates for inference. However, this variability may in fact be of assistance when using the bootstrap technique. Sampling from a cohort with extreme variability allows the bootstrap to develop more heterogeneous samples, which may more accurately reflect the uncertainty in the true population’s parameter. Conversely, a homogeneous cohort (little variability) may not provide confidence intervals that are wide enough to incorporate the true population’s estimate.

Secondly, the distribution of scores or values in the standard error and confidence intervals is much more revealing than just an assessment of the mean. It would not be surprising if one of the effects that a DM program has on outcomes is of reducing the variability around the mean. Moreover, using the median instead of the mean may be the preferred method for analyzing this type of data, since it is not susceptible to the impact of outliers or skew. Nonetheless, by educating doctors to follow evidence-based practice guidelines, and educating patients on how and when to use health services, one might expect to see a reduction in outlier behavior. This impact only becomes evident upon examination of the SE and CI. Adding other measures of central tendency (such as median) and information on data distribution to the standard analysis of the mean may provide a more complete picture of program effects.

Thirdly, the ease of implementation and simplicity of the bootstrap procedure allow this technique to be applied in a wide variety of additional situations that arise during program analysis, such as with the use of; categorical data, regression and correlations, analysis of variance (ANOVA), probability estimates, etc. It allows the analyst to derive the chosen statistical parameters programmatically, as opposed to mathematically. This method also has the potential to reduce the threat of multiple comparisons bias that may be introduced when many statistical calculations are performed repeatedly. Moreover, it is superior to standard statistical tests of significance in that it provides information on the distribution of scores as opposed to parametric distributions and is generally more accurate.^[19,20]

CONCLUSIONS

This paper presented a data-driven simulation technique that produces thousands of random samples from the dataset creating distributions that allow for a more accurate method of drawing statistical inferences of a population. Moreover, since classical statistical inference techniques were designed specifically for parametric statistics (i.e. assuming a normal distribution) the bootstrap can be used for measures that have no convenient statistical formulae. The median is one such measure that is worth considering when evaluating health care data. It is more robust than the mean, and is not impacted by outliers. Using the bootstrap, confidence intervals can be defined around this statistic, making it a viable option for evaluating program effectiveness.

Address for correspondence:

Ariel Linden, DrPH, MS

President, Linden Consulting Group

6208 NE Chestnut Street

Hillsboro, OR 97124

(503) 547-8343

APPENDIX A

Mean:

The equation for the mean is as follows:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_N}{N} \quad (1)$$

\bar{x} is used for a sample, and μ is used when all values of the population are included in the calculation. x_1 through x_N are all the individual values in the dataset, and N is the total number or count of those values.

Median:

The equation for the median is as follows:

$$\begin{aligned} \tilde{x}_{odd} &= Y(N+1)/2 \\ \tilde{x}_{even} &= \frac{1}{2}(Y_{N/2} + Y_{1+N/2}) \end{aligned} \quad (2)$$

first, all data must be ordered. Then depending on whether the data set is odd or even-numbered, one of these formulas will be used, where Y is the ordered value, and N is the total number or count of those values.

Standard deviation:

The equation for the standard deviation is as follows:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \quad (3)$$

where \bar{x} is the sample mean, x_i represents a given data point, and n is the total number of data points. Standard deviation is typically notated as s , SD, or σ .

Standard error:

The equation for the SE is as follows:

$$SE = \frac{s}{\sqrt{N}} \quad (4)$$

where s is the standard deviation of the sample, and N is the sample size. The SE typically decreases as N increases.

Confidence intervals:

The simplified equation for the confidence interval is as follows:

$$CI = \bar{x} \pm 1.96 \times SE \quad (5)$$

where \bar{x} is the sample mean, SE is calculated using equation 3, and 1.96 is the z-score corresponding with a confidence level of $\alpha = 0.05$ (1.96 is a z-score that represents 95% of the values within a normal distribution).

REFERENCES

1. American Healthways and the John Hopkins Consensus Conference. Consensus report: standard outcome metrics and evaluation methodology for disease management programs. *Disease Management* 2003;6(3):121-138.
2. Linden A, Adams J, Roberts N. An assessment of the total population approach for evaluating disease management program effectiveness. *Disease Management* 2003;6(2): 93-102
3. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: An introduction to time series analysis. *Disease Management* 2003;6(4):243-255
4. Linden A, Adams J, Roberts N. Evaluation methods in disease management: determining program effectiveness. Position Paper for the Disease Management Association of America (DMAA). October 2003
5. Linden A, Adams J, Roberts N. Using propensity scores to construct comparable control groups for disease management program evaluation. *Disease Management and Health Outcomes* (in press)
6. Linden A, Roberts N. Disease management interventions: What's in the black box? *Disease Management* 2004;7(4):XX-XX (in press)
7. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: An introduction to survival analysis. *Disease Management* 2004;7(3):XX-XX (in press)
8. Linden A, Adams J, Roberts N. Using an empirical method for establishing clinical outcome targets in disease management programs. *Disease Management* 2004;

7(2):93-101.

9. Mosteller F, Tukey J. Data analysis and regression. Reading, MA: Addison-Wesley, 1977
10. Efron B. Bootstrap methods: another look at the jackknife. *Ann Statist.* 1979;7:1-26
11. Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman & Hall, 1993
12. Chernick MR. Bootstrap methods: a practitioner's guide. New York: Wiley, 2000
13. Davison AC, Hinkley DV. Bootstrap methods and their applications. Cambridge: Cambridge University Press, 1997
14. Lunneborg CE. Data analysis by resampling; concepts and applications. Pacific Grove, CA: Brooks-Cole, 2000
15. E. S. Pearson. 'Student', A Statistical Biography of William Sealy Gosset, Edited and Augmented by R. L. Plackett with the Assistance of G. A. Barnard, Oxford: University Press, 1990
16. Disease Management Consortium, LLC. www.dismgmt.com
17. Vogt PW. Dictionary of statistics and methodology: a non-technical guide for the social sciences, 2nd ed. Thousand Oaks, CA: Sage, 1999
18. Blank, S., Seiter, C., Bruce, P., "Resampling Stats add-in for Excel user's guide." Arlington, VA: Resampling Stats, Inc. 2003
19. Ludbrook J, Dudley, H. Why permutation tests are superior to t and F tests in biomedical research. *Amer Statist.* 1998;52(2):127-132
20. Reichardt CS, Gollob HF. Justifying the use and increasing the power of a t test for a randomized experiment with a convenience sample. *Psych Methods* 1999;4:117-128

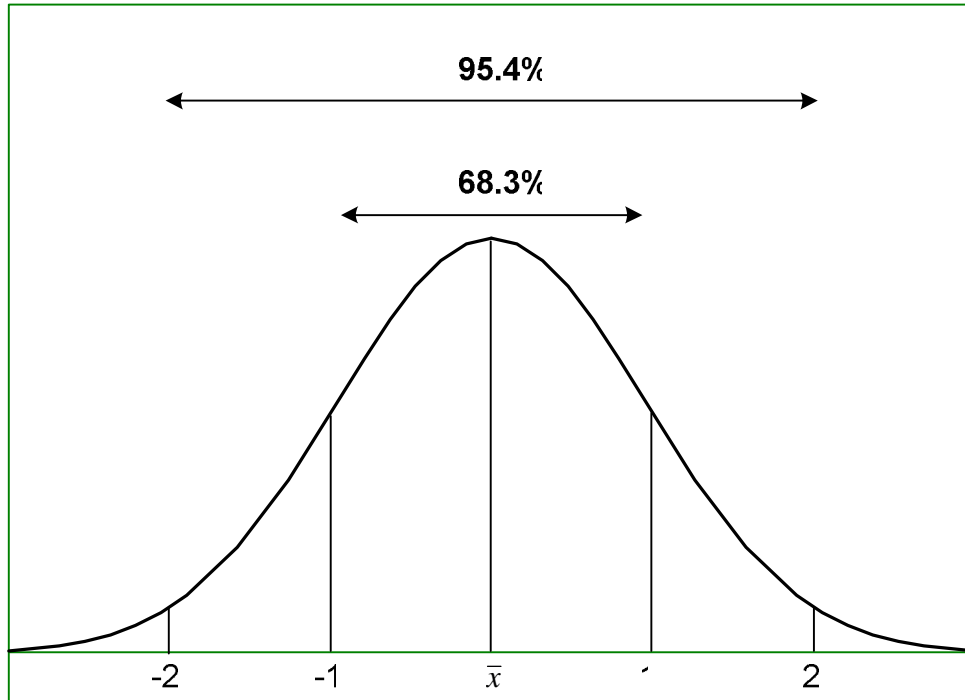


FIG 1. The Student t-distribution. As illustrated, the sample mean (\bar{x}) is expected to lie within 1 standard error away from the true population mean roughly 68% of the time if the sample size is large enough, and 2 standard errors from the true population mean 95% of the time.

Table I. An illustration of how sampling with replacement works by rolling two dice.

	1 st roll	2 nd roll	3 rd roll	4 th roll	5th roll	6th roll	7 th roll	8th roll	9th roll	10 th roll
1 st Dice Value	4	1	3	2	1	5	4	5	6	5
2 nd Dice Value	5	6	5	4	1	1	3	3	1	1
Totals	9	7	8	6	2	6	7	8	7	6

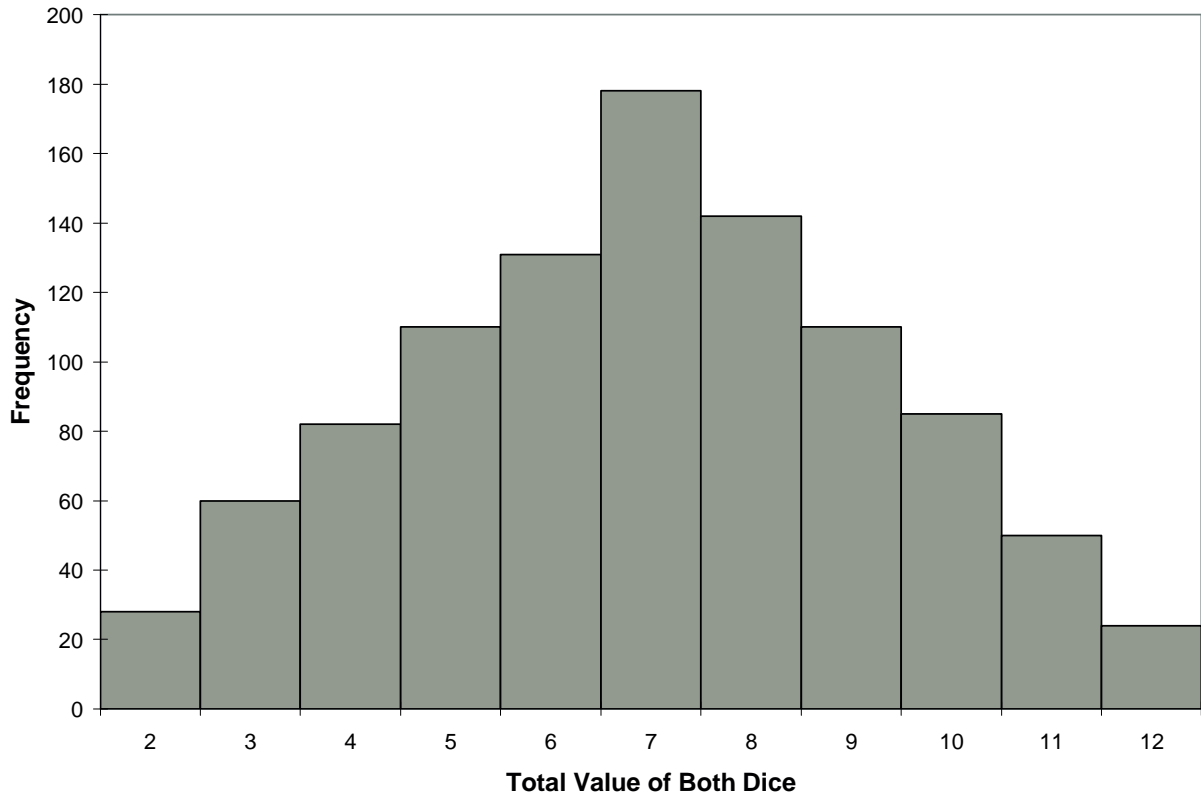


FIG 2. Histogram illustrating the distribution of values for 1000 rolls of two dice.

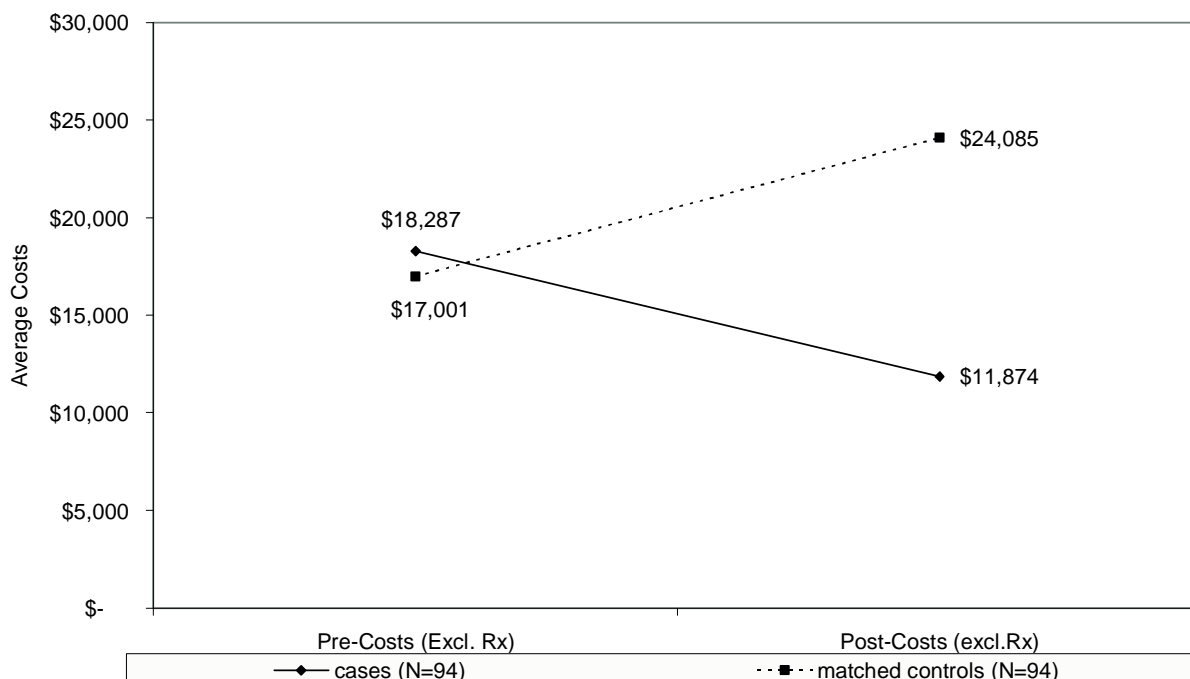


FIG 3. A comparison of total average costs between members enrolled in the CHF program and controls, matched on their propensity score. All subjects were continuously enrolled with the health plan for at least 2 years (1 year prior to program commencement, and the duration of the 1st program year).⁴

Table II. Sampling with replacement on the difference between pre and post costs. Samples 1 through 5 represent five of the total 1000 random samples drawn from the “difference” column. Similarly this table presents data for only 10 of the 94 individuals.

Number	Step 1			Step 2				
	Pre (\$)	Post (\$)	Difference	Sample 1	Sample 2	Sample3	Sample 4	Sample 5
1	\$10,016	\$2,872	\$7,143	\$221	(\$17,955)	(\$17,955)	\$221	\$3,911
2	\$10,088	\$16,504	(\$6,417)	\$25,858	(\$6,417)	\$37,169	\$7,170	\$221
3	\$39,159	\$13,301	\$25,858	\$221	\$7,170	\$25,858	\$7,143	\$221
4	\$7,358	\$7,137	\$221	\$7,170	\$4,989	\$7,170	(\$17,955)	\$728
5	\$12,131	\$4,961	\$7,170	(\$6,417)	\$37,169	\$25,858	(\$6,417)	\$221
6	\$3,347	\$2,619	\$728	\$4,989	\$4,989	\$4,989	\$3,911	\$728
7	\$5,801	\$23,755	(\$17,955)	\$728	\$7,170	\$25,858	\$728	\$25,858
8	\$6,750	\$1,761	\$4,989	(\$17,955)	\$7,143	\$728	(\$17,955)	\$7,143
9	\$43,647	\$6,478	\$37,169	\$728	\$7,143	(\$17,955)	\$25,858	\$3,911
10	\$7,634	\$3,722	\$3,911	\$221	\$4,989	(\$6,417)	(\$17,955)	\$7,170
Totals	\$145,931	\$83,110	\$62,817	\$15,764	\$56,390	\$85,303	(\$15,251)	\$50,112
Means	\$14,593	\$8,311	\$6,281	\$1,576	\$5,639	\$8,530	(\$1,525)	\$5,011
Medians	\$8,825	\$5,720	\$4,450	\$475	\$6,066	\$6,080	\$475	\$2,320

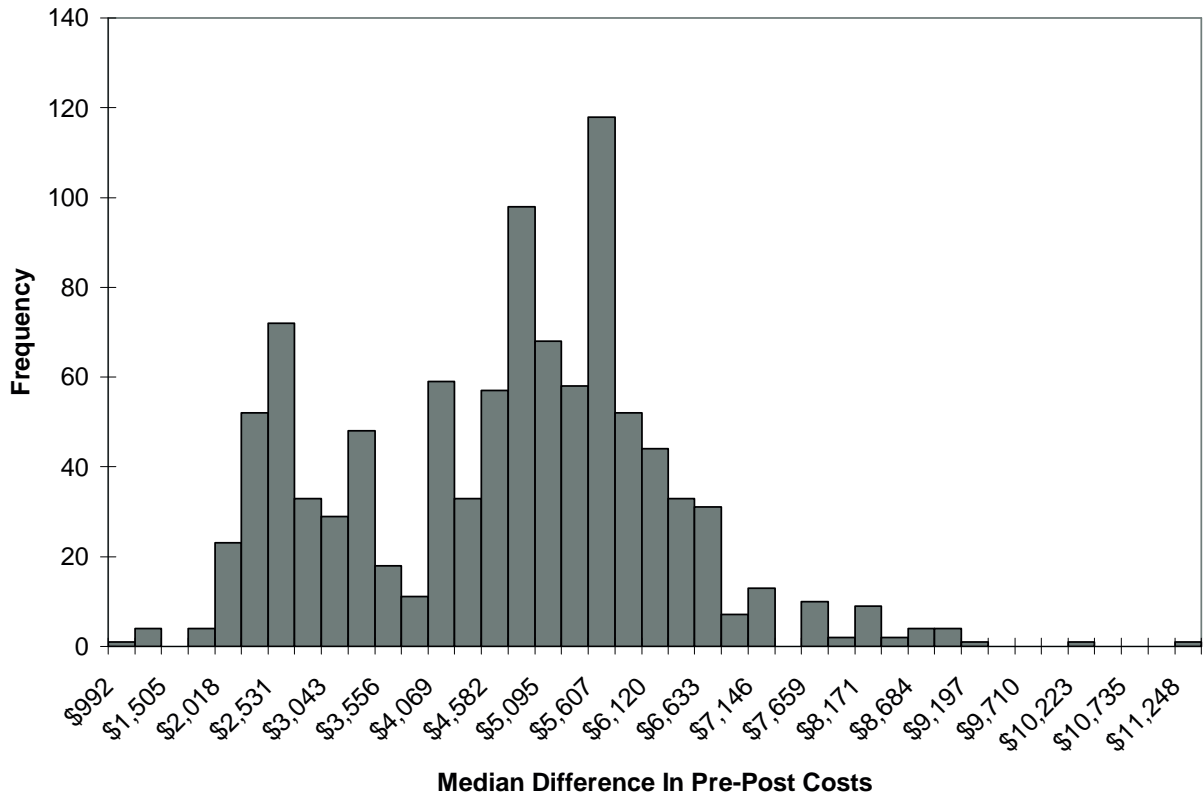


FIG 4. Histogram of 1000 samples of 94 median difference scores (pre costs – post costs) of DM program participants.

Table III. Mean costs of CHF patients during the pre-program and 1st year program periods. The difference-in-difference estimator is based on the difference for the program participants and control group in the difference between the pre- and 1st year patient median costs.

	Baseline Median (\$)	1 st Year Median (\$)	Difference (\$) (Baseline – 1 st Year)
DM Participants (N=94)	12,075	6,583	5,492
Controls (N=94)	8,270	9,714	(1,444)
Difference-in-differences			6,936

Table IV. A bootstrap determination of the 25th and 75th quartile ranking based on initial costs. The difference-in-difference (DID) estimator is based on the difference for the program participants and control group in the difference between pre- and 1st year costs. The 24 values comprising each quartile were bootstrapped 1000 times.

	25 th Percentile (difference-in-differences)	75 th Percentile (difference-in-differences)
Mean	\$15,025	\$15,115
Lower 95% CI	\$3,103	(\$8,894)
Upper 95% CI	\$30,236	\$40,221

TableV. Proportion of CHF patients experiencing at least one hospitalization during the pre-program and 1st year program periods. The difference-in-difference estimator is based on the difference for the program participants and control group in the difference between pre- and 1st year proportion of hospitalizations.

	Baseline Proportion	1 st Year Proportion	Difference (Baseline – 1 st Year)
DM Participants (N=94)	0.59	0.30	0.29
Controls (N=94)	0.51	0.45	0.06
Difference-in-differences			0.23

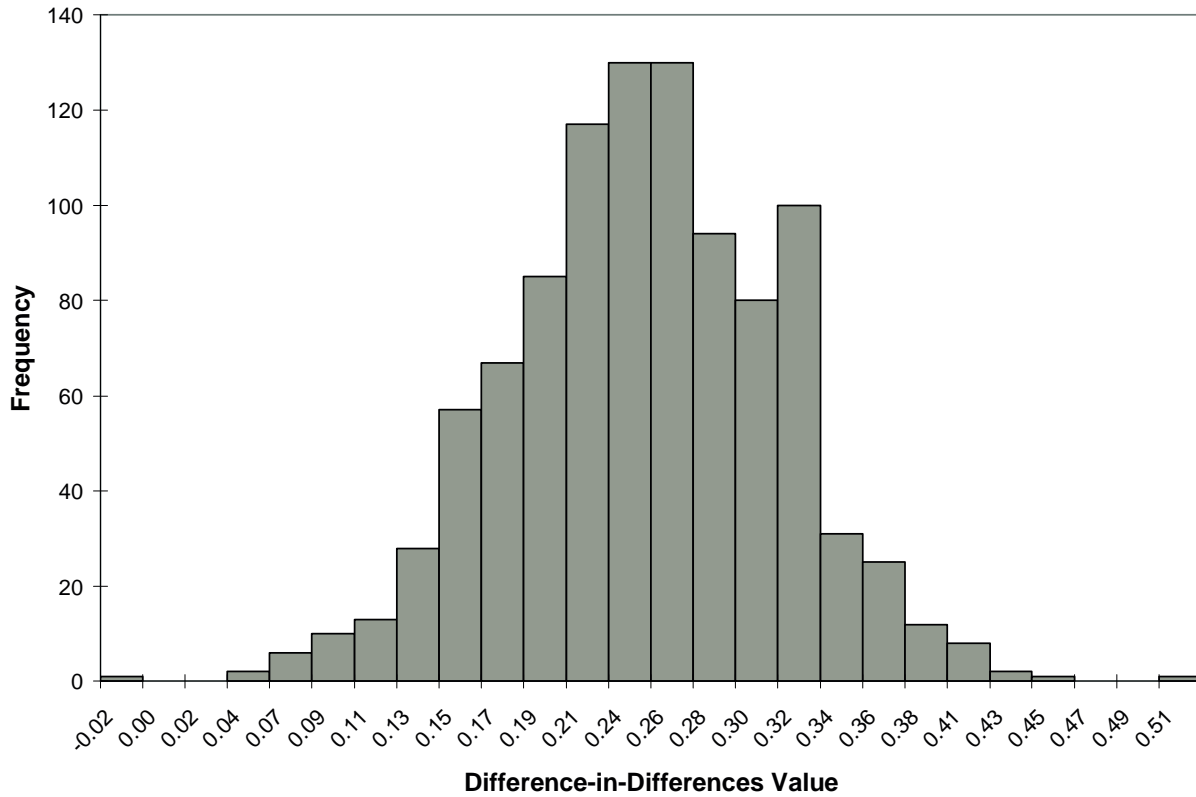


FIG 5. Histogram of 1000 samples of 100 difference-in-difference values (based on the difference for the program participants and control group in the difference between pre- and 1st year proportion of hospitalizations).