LEADING ARTICLE

Dis Manage Health Outcomes 2005; 13 (3): 159-167
1173-8790/05/0003-0159/$34.95/0

Evaluating Disease Management Program Effectiveness

An Introduction to the Bootstrap Technique

Ariel Linden,[1] *John L. Adams*[2] and *Nancy Roberts*[3]

1 Linden Consulting Group, Portland, Oregon, USA
2 RAND Corporation, Santa Monica, California, USA
3 Integrated Performance/Six Sigma Champion, Providence Health System, Portland, Oregon, USA

Abstract

Disease management (DM) program evaluations are somewhat limited in scope because of 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 (CI) calculations, the analyst may be fooled by specious details.

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 data set, allowing the use of repeated samples' statistic (e.g. mean, proportion, and median). Using a congestive heart failure (CHF) program as an example, the bootstrap technique is used to extend a DM program evaluation beyond questions addressed using classical statistical inference: (i) how much of a median cost decrease can be expected as a result of the program?; (ii) did the program impact the highest and lowest costing members equally; and (iii) how much of a decrease in the proportion of patients experiencing a hospitalization can be expected as a result of the program?

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.

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, CIs can be defined around this statistic, making it a viable option for evaluating DM program effectiveness.

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 with average costs in the baseline year. After adjustments have been made for vendor fees and other variables, 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 almost 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. A series of such alternative techniques for DM evaluation have been published by the authors of this article.[2-8] This paper is a continuation of the series.

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 emergency room 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 (SE) and confidence interval (CI) calculations 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 Efron[10] in 1979 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.

1. 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 Guiness Brewery Co., he was not allowed to publish under his own name or affiliation), Gossett identified that one can estimate the mean and SE of a normally distributed population when the sample size is relatively large and the standard deviation (SD) is unknown. This gave rise to the Student t-distribution, also called the bell curve or normal curve. The limitations of methods of this type are that such approximations: (i) rely on the assumption that the data are normally distributed; (ii) achieve better accuracy in large samples than in small ones; and (iii) 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. Although 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 SE or CIs 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.

2. 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.

2.1 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.

2.2 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 data set. The most commonly used measure of variation is the SD. A large SD indicates that the data are dispersed far from the mean or that the data contain outliers.

2.3 Mode

Mode is simply the value that appears most frequently in a data set. 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 and this is true if all the observations occur with the same frequency.

2.4 Median

The median is the middle value of a data set. 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 are 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.

2.5 Standard Error

The SE is an estimate of the variation of the sampling distribution of a given statistic (i.e. mean, median, and proportion). Using the mean as an example, the SE estimates the SD 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 CIs. The SE typically decreases as sample size increases.

2.6 Confidence Intervals

A 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 that the population mean may exist (with a given level of confidence). CIs 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 CIs. 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 data set or may be a result of having a very small sample size.

CIs 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% CI (as a theoretical example: 7, 13), 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 reduction in cost will be between 7% and 13%.

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

3. Principles of the Bootstrap

Compared with 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 data set, allowing the use of repeated samples' statistic (e.g. mean, proportion, and median) to generate variation in population estimates. Implicit is the assumption that this data set 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. Similar to 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 SEs 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.

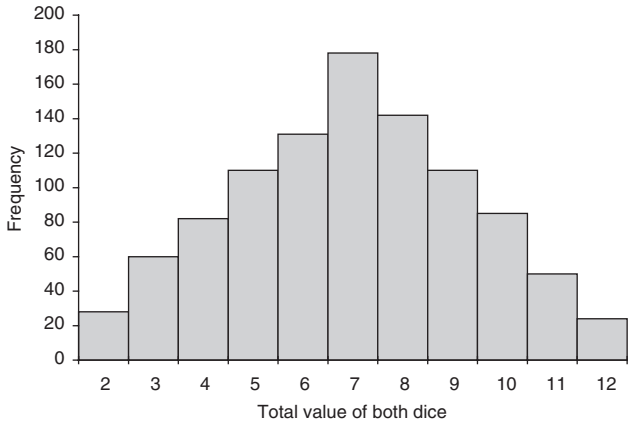

Fig. 1. Histogram of the distribution of dice face values for 1000 rolls of two dice.

The bootstrap technique entails drawing a defined number of random samples from the original data set (which in itself is a sample from the population). Since the number of data points within the original data set 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 data set, so that it has a chance of being reselected for that sample and for all subsequent samples.

Figure 1 illustrates what happens when two dice are rolled 1000 times. As the figure shows, a normal distribution develops around the mean (a value of 7).

The SE of the distribution of the total number when two dice are rolled is also easily derived. In the bootstrap simulation, the SD of the distribution of values (e.g. the SD of the 1000 simulations) is in fact the SE. In the case of 1000 rolls of two dice, the SD, and hence the SE, was 2.4.

Similarly, CIs 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% CIs (2.5% on each tail, respectively). Using this procedure, the mean (and 95% CIs) 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 SE 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. <20), skewed distributions (e.g. data sets with a few very large outliers), or quantities that depend on a subset of the data set (e.g. extreme percentiles.) All analyses and histograms reported in this paper were generated using Resampling Stats[18] for Excel.

4. Examples from Disease Management

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

Figure 2 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 shown at the end of the program year, with the DM program group experiencing an average reduction in costs of $US6413, while at the same time the control group exhibited an average increase of $US7084.

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).

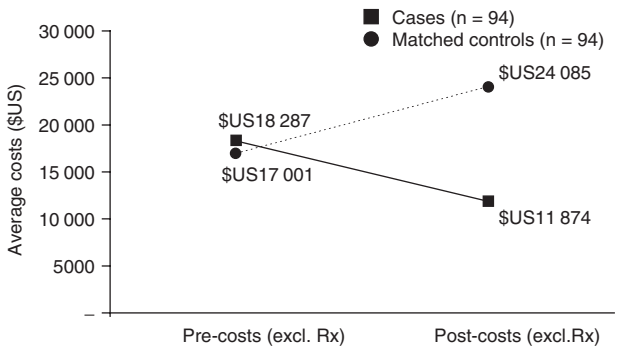

Fig. 2. Costs incurred for disease management (DM) participants and controls during pre- and post-periods of a congestive heart failure DM program (reproduced from Linden et al.,[4] with permission). **Pre** = 1 year prior to the DM program; **Post** = the first year during the DM program; **excl. Rx** = excluding prescription.

Table I. Sampling with replacement for the difference between costs incurred for disease management (DM) participants during pre- and post-periods of a congestive heart failure DM program[a]

Participant[b]	Costs determined using step 1[c] ($US)			Costs determined using step 2[d] ($US)				
	pre	post	difference between pre and post	sample 1[e]	sample 2[e]	sample 3[e]	sample 4[e]	sample 5[e]
1	10 016	2872	7143	221	(17 955)	(17 955)	221	3911
2	10 088	16 504	(6417)	25 858	(6417)	37 169	7170	221
3	39 159	13 301	25 858	221	7170	25 858	7143	221
4	7358	7137	221	7170	4989	7170	(17 955)	728
5	12 131	4961	7170	(6417)	37 169	25 858	(6417)	221
6	3347	2619	728	4989	4989	4989	3911	728
7	5801	23 755	(17 955)	728	7170	25 858	728	25 858
8	6750	1761	4989	(17 955)	7143	728	(17 955)	7143
9	43 647	6478	37 169	728	7143	(17 955)	25 858	3911
10	7634	3722	3911	221	4989	(6417)	(17 955)	7170
Totals	145 931	83 110	62 817	15 764	56 390	85 303	(15 251)	50 112
Means	14 593	8311	6281	1576	5639	8530	(1525)	5011
Medians	8825	5720	4450	475	6066	6080	475	2320

a Data were extracted from Linden et al.[4] Numbers in parentheses represent an increase in costs from pre- to post-program.

b Data for only 10 of the 94 individuals are presented (solely because of space contraints).

c Step 1 is to compute the difference in pre- and post-costs of each of the 94 program participants (pre-costs/post-costs = difference score).

d Step 2 is to use the bootstrap to create 1000 samples of 94 randomly selected participant's difference scores.

e Samples 1–5 represent 5 of the total 1000 random samples drawn from the 'difference between pre and post' column (using sampling with replacement technique of the values in the difference column).

Pre = 1 year prior to the DM program; **Post** = the first year during the DM program.

4.1 How Much of a Median Cost Decrease Can Be Expected as a Result of The Program?

The median costs are used in this example for two reasons: (i) because classical statistical methods do not supply measurement parameters such as SE or CI for medians; and (ii) because the median may be a more appropriate metric because of 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; and
3. find the median, SE and CI for the difference score.

Table I illustrates steps (1) and (2), providing data for only 10 of the 94 participants (this was done solely because of 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 I shows that participant number 4 (difference score $US221) is present three times in samples 1 and 5, once in sample 4, and it is not found in samples 2 and 3. Upon drawing the 1000 samples, the median of the distribution as explained in section 2.4. 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% CIs.

The results using the bootstrap technique on the data from Linden et al.[4] elicited the following statistics: (i) median difference = $US3100; (ii) SE = $US1607; (iii) lower 95% CI = $US249; and (iv) upper 95% CI = $US6897. These results indicate that the median and SE are significantly lower than the mean and SE calculated using the classical method with mean = $US6413, and SE = $US2403. 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

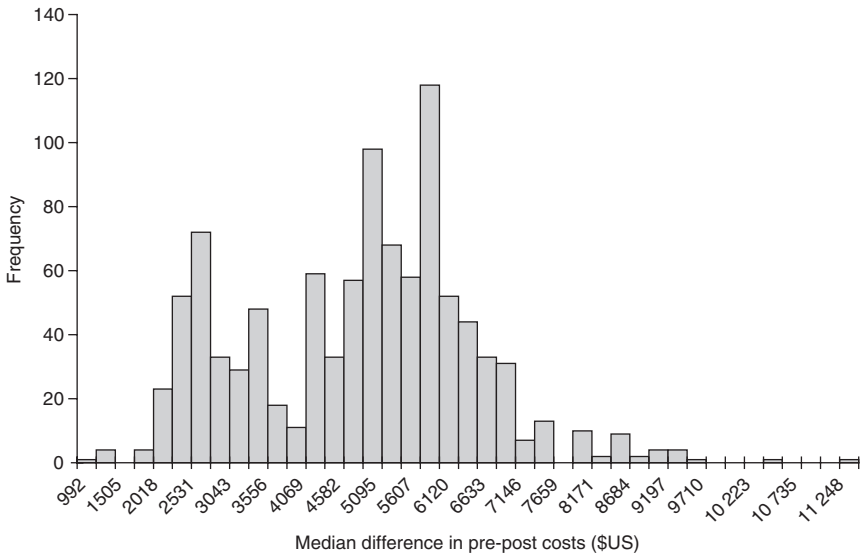

Fig. 3. Histogram of 1000 bootstrap samples of 94 median cost differences between pre- and post-period of a congestive heart failure disease management program. Data were extracted from Linden et al.[4] **Pre** = 1 year prior to the DM program; **Post** = the first year during the DM program.

patients enrolled in a DM program for 1 year will reduce their median costs by between $US249 and $US6897.' A histogram of the distribution of the bootstrap samples is shown in figure 3.

The results of this analysis illustrate the importance of considering CIs in the evaluation of DM program outcomes. In the scenario above, the distribution of individual results was quite wide, ranging from >$US25 000 to cases where this difference was -$US17 955. Likewise, as shown in figure 3, the 1000 samples of median difference in pre-post costs ranged from $US992 to $US11 248. This variability in results drives the wide CI of the median difference, which ranged from $US249 to $US6897 (however, the CI for the mean difference ranged from $US1860 to $US11 195, which was much larger than that of the median). This example provides support for considering the use of the median and CIs 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 with the difference in pre- and post-costs of the control group. This method is referred to as the difference-in-differences (DIDs) estimator. As the data in figure 2 and table II suggest, costs decreased in the DM program and increased in the control cohort (whether calculated as the mean [figure 2] or median [table II]). The DIDs method provides a more accurate account of the effect on the entire population because, as illustrated in table II, an adjustment is made for the increasing trend effect that occurred in the untreated group during the same time period. Bootstrapping the DIDs 1000 times gave the following results: (i) median DIDs = $US4956; (ii) SE = $US1540; (iii) lower 95% CI = $US2141; and (iv) upper 95% CI = $US7697. These outcomes can be stated as follows; 'we are 95% confident that the median difference in pre- and post-costs of CHF patients enrolled in a DM program compared with the control group will be between $US2141 and $US7697.' 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.

4.2 Did The Program Impact The Highest and Lowest Costing Members Equally?

Considering that DM program interventions specifically target those behaviors that are costly (e.g. hospitalizations and emergen-

Table II. Median costs incurred for disease management (DM) participants and controls during pre- and post-periods of a congestive heart failure DM program[a]

Study group	Median costs ($US)		
	pre	post	difference between pre and post
DM participants (n = 94)	12 075	6583	5492
Controls (n = 94)	8270	9714	(1444)
Difference–in-differences[a]			6936

a Data were extracted from Linden et al.[4] The difference-in-differences estimator is based on the difference for the program participants and control group for the difference between the median pre- and post costs.

Pre = 1 year prior to the DM program; **Post =** the first year during the DM program.

Table III. Bootstrap determination of difference in pre- and post-costs for disease management (DM) participants with the difference in pre- and post-costs of controls during a congestive heart failure DM program[a]

Difference between pre- and post-costs using the DIDs estimator[b] ($US)	25th Percentile[c]	75th Percentile[c]
Mean	15 025	15 115
Lower 95% CI	3103	(8894)
Upper 95% CI	30 236	40 221

a Data were extracted from Linden et al.[4] and calculated according to 25th and 75th percentile ranking for pre costs.

b The DIDs estimator is based on the difference in pre- and post-costs of DM program cohort with the difference in pre- and post-costs of the control group.

c The 24 values comprising each quartile (total n = 94) were bootstrapped 1000 times.

DIDs = difference-in-differences; **Pre** = 1 year prior to the DM program; **Post** = the first year during the DM program.

cy department visits), 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 from Linden et al.[4] 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 DIDs was calculated by subtracting the control group's mean difference from the program participant group'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 III.

As illustrated in table III, the mean difference in both the 25th and 75th quartile was positive ($US15 025 and $US15 115 for the 25th and 75th quartile, respectively). A positive value indicates that the cases experienced a greater reduction in costs than the controls (because the calculation was based on cases minus controls). More specifically, in the 25th quartile (those members with the lowest baseline year initial costs), program participants had, on average, a $US15 025 greater reduction in costs than controls (with 95% CI between $US3103 and $US30 236). Similar results are evident in the 75th quartile. However, since the CI crosses 0 (the lower CI is negative and the upper CI is positive), 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 compared with the control group (by an average of approximately $US15 000). In the 75th percentile (those members who had the highest healthcare costs in the baseline period), program participants achieved lower increases in costs than their controls (by approximately $US15 000). However, looking at the 95% CIs shows that the difference in pre- and post-program costs between the groups in the 75th percentile did not differ significantly because of the variability around the mean.

4.3 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 metric statistics explained in section 2. This example, the impact of a DM program on proportions, will describe a statistic that achieves better results when using the bootstrap method as opposed to the classic calculations.

For this example, the hospitalization data from Linden et al.[4] was re-characterized so that a patient with CHF 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 IV 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 who experienced a hospitalization during the program year. The DIDs estimator was used to adjust for the divergence in scores between the two groups.

The DIDs 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 CIs of 11% and 37% (lower and upper 95% CI, respectively). Figure 4 presents the histogram of the bootstrap samples DIDs 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.'

Table IV. Proportion of disease management (DM) participants and controls experiencing at least one hospitalization during pre- and post-periods of a congestive heart failure DM program[a]

Study group	Proportion of hospitalizations		
	pre	post	difference between pre and post
DM participants (n = 94)	0.59	0.30	0.29
Controls (n = 94)	0.51	0.45	0.06
Difference–in-differences[b]			0.23

a Data were extracted from Linden et al.[4]

b The difference-in-differences estimator is based on the difference in pre- and post-hospitalizations of DM program cohort with the difference in pre- and post-hospitalizations of the control group.

Pre = 1 year prior to the DM program; **Post =** the first year during the DM program.

5. 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 whether these data 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 CIs that are wide enough to incorporate the true population's estimate.

Secondly, the distribution of scores or values in the SE and CIs 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 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 allows this technique to be applied in a wide

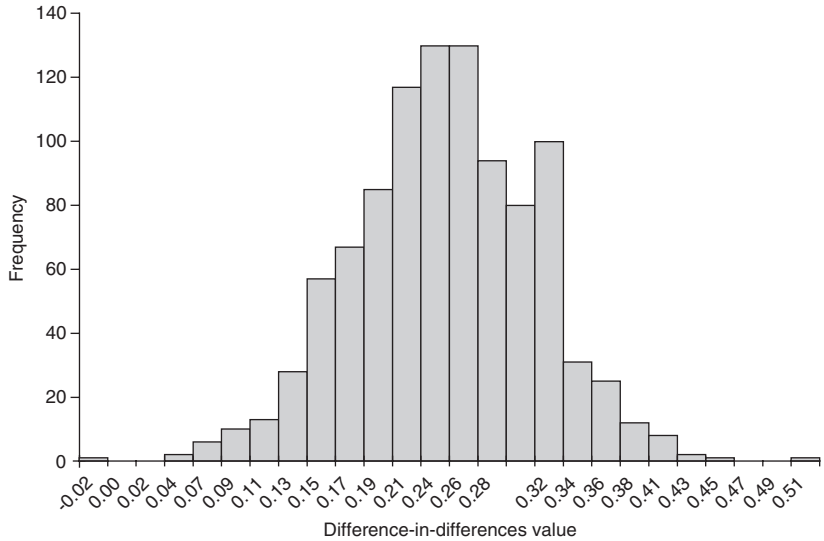

Fig. 4. Histogram of 1000 bootstrap samples of 100 difference-in-differences (DIDs) proportions of hospitalizations between pre- and post-period of a congestive heart failure disease management program for program participants and controls. Data were extracted from Linden et al.[4] The DIDs estimator is based on the difference in pre- and post-hospitalizations of the DM program cohort with the difference in pre- and post-hospitalizations of the control group. **Pre** = 1 year prior to the DM program; **Post** = the first year during the DM program.

variety of additional situations that arise during program analysis, such as with the use of categorical data, regression and correlations, analysis of variance, and probability estimates. 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] That said, the bootstrap technique may be limited in its accuracy when the data represent small sample sizes (e.g. <20) or skewed distributions, contain extreme outliers, or pertain to only a subset of the entire data set.

6. Conclusions

This paper presented a data-driven simulation technique that produces thousands of random samples from the data set 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 healthcare data. It is more robust than the mean and is not impacted by outliers. Using the bootstrap, CIs can be defined around this statistic, making it a viable option for evaluating program effectiveness.

Acknowledgments

No sources of funding were used to assist in the preparation of this study. The authors have no conflicts of interest that are directly relevant to the content of this study.

References

1. American Healthways and the John Hopkins Consensus Conference. Consensus report: standard outcome metrics and evaluation methodology for disease management programs. Dis Manag 2003; 6 (3): 121-38
2. Linden A, Adams J, Roberts N. An assessment of the total population approach for evaluating disease management program effectiveness. Dis Manag 2003; 6 (2): 93-102
3. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: an introduction to time series analysis. Dis Manag 2003; 6 (4): 243-55
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). 2003 Oct
5. Linden A, Adams J, Roberts N. Using propensity scores to construct comparable control groups for disease management program evaluation. Dis Manage Health Outcomes 2005; 13 (2): 107-27
6. Linden A, Roberts N. Disease management interventions: what's in the black box? Dis Manag 2004; 7 (4): 275-91
7. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: an introduction to survival analysis. Dis Manag 2004; 7 (3): 180-90
8. Linden A, Adams J, Roberts N. Using an empirical method for establishing clinical outcome targets in disease management programs. Dis Manag 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. JSTOR 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. Pearson ES. '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, 2004 [online]. Available from URL: www.dismgmt.com [Accessed 2005 Jan 2]
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. Am Stat 1998; 52 (2): 127-32
20. Reichardt CS, Gollob HF. Justifying the use and increasing the power of a t test for a randomized experiment with a convenience sample. Psychol Methods 1999; 4: 117-28

About the Author: Dr Linden is a health services researcher whose focus is on evaluating disease management program effectiveness. In this area alone he has recently had 17 manuscripts published, including a position paper for the Disease Management Association of America. He is President of the Linden Consulting Group and Clinical Associate Professor in the School of Medicine at Oregan Health and Science University, Oregan, USA.
Correspondence and offprints: Dr *Ariel Linden*, Linden Consulting Group, 6208 NE Chestnut Street, Hillsboro, OR 97124, USA.
E-mail: alinden@lindenconsulting.org