Evaluating Disease Management Program Effectiveness: An Introduction to Time-Series Analysis

Article *in* Disease Management · February 2003

DOI: 10.1089/109350703322682559 · Source: PubMed

3 authors:

Ariel Linden
Linden Consulting Group, LLC
110 PUBLICATIONS **1,637** CITATIONS
SEE PROFILE

John L Adams
Kaiser Permanente, Pasadena, United States
189 PUBLICATIONS **13,037** CITATIONS
SEE PROFILE

Nancy Roberts
Providence St. Vincent Medical Center
18 PUBLICATIONS **530** CITATIONS
SEE PROFILE

DISEASE MANAGEMENT
Volume 6, Number 4, 2003

Evaluating Disease Management Program Effectiveness: An Introduction to Time-Series Analysis

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

ABSTRACT

Currently, the most widely used method in the disease management (DM) industry for evaluating program effectiveness is referred to as the "total population approach." This model is a pretest–posttest design, with the most basic limitation being that without a control group, there may be sources of bias and/or competing extraneous confounding factors that offer a plausible rationale explaining the change from baseline. Furthermore, with the current inclination of DM programs to use financial indicators rather than program-specific utilization indicators as the principal measure of program success, additional biases are introduced that may cloud evaluation results. This paper presents a non-technical introduction to time-series analysis (using disease-specific utilization measures) as an alternative, and more appropriate, approach to evaluating DM program effectiveness than the current total population approach. (Disease Management 2003;6:xxx–xxx.)

INTRODUCTION

CURRENTLY, THE MOST WIDELY USED METHOD in the disease management (DM) industry for evaluating program effectiveness is referred to as the "total population approach." Effectiveness, in this context, relates to the ability to demonstrate medical cost savings as a result of the program's intervention on a specific diseased population.

This model is a pretest–posttest design, with the most basic limitation being that without a control group, there may be sources of bias and/or competing extraneous confounding factors that offer plausible alternative explanations for the change from baseline (refer to Linden et al.[1] for a comprehensive review of the limitations of this approach).

There are three fundamental reasons why, given these limitations, the industry continues to use this method for assessing program impact: (1) Using an experimental design in which the experimental and control groups are compared after the experimental group receives the intervention has several practical barriers. Many organizations are hesitant to offer services to one subset of the population while withholding that same "value-added" benefit to others. In fact, for health plans, doing so might violate commercial and/or governmental contracts. In addition, experimental designs may be challenging to accurately develop given the limitations of claims-based algorithms used for identifying suitable patients. Cost may also be a practical deterrent to experimental designs. (2) DM program administrators can usu-

[1]Providence Health Plans and [3]Providence Health System, Portland, Oregon.
[2]RAND Corporation, Santa Monica, California.

ally supply evaluation reports illustrating potential cost savings and improvements in relevant clinical outcomes. Their argument would be that with these data indicating that the program is indeed effective, there is no further need to conduct an experimental research design for evaluation purposes. (3) Finally, given the belief that these interventions are clinically effective, the strategy should be to treat all members with the disease, since each member receiving the intervention has the potential of adding to the medical cost savings promised by the program.

Measuring changes in behavior that take place over time should be an integral component of the DM program evaluation. The early months of a DM program are geared toward enrollment and initial patient assessments. Most DM programs have a phased enrollment process, bringing a percentage of eligible members into the program each month. Full program enrollment is usually not achieved until 3–6 months after program launch. If the intervention is effective at the patient level, it will not be evident until several months or even a year into the program. At that juncture, patients should have been given the educational tools necessary to better self-manage the disease, as well as to ensure that clinical or physiological measures have achieved levels indicating control. In aggregate, patient-level improvements will manifest as population-wide changes in medical utilization variables. At some point in time, intervention effectiveness may flatten or be reduced. Awareness of these temporal influences assists the DM program evaluator to identify, describe, explain, and predict the effects of processes that bring about change as a result of the program intervention.

The methods currently in use are inadequate for evaluating outcomes in uncontrolled settings (i.e., the natural environment) and inefficient for studying associations over time (i.e., multiple measurement years compared with baseline).[2–4] This paper presents a non-technical introduction to time-series analysis as an alternative, and more appropriate, approach to evaluating DM program effectiveness than the current total population approach. This introduction will provide DM program evaluators, both within DM firms and for those managing in-house programs, enough detail to begin using these techniques. For those organizations that purchase DM services, this paper will provide a substantive background with which to discuss alternative evaluation possibilities with their contracted vendors.

PRINCIPLES OF TIME-SERIES ANALYSIS

A time series can be simply defined as a variable that undergoes a repeated periodic observation, or measurement. The variable can be either at the patient level (e.g., EKG readings, respiratory rates, blood pressure), or at some aggregate level (e.g., hospitalization rates, mammography rates). The periodic measurement may be as short as a fraction of a second or as long a century. In general, time-series analyses are used to characterize a pattern of behavior occurring in the natural environment over the measurement period, analyze fluctuations of the variable along the continuum, infer the impact of an intervention introduced during the measurement period, and forecast future direction of the time series variable.[5–7]

An important feature of time series is that of serial dependence. Any variable measured over time is potentially influenced by previous observations (autocorrelation). To take advantage of these relationships time-series models use previous observations as the basis for predicting future behavior. This is the essential difference between time-series analysis and traditional statistical tests for measuring change, such as regression analysis, which rely on variation in independent variables to explain changes in the outcome.

Some time-series methods allow the DM program evaluator to predict future behavior of the observed variable without attempting to measure independent relationships that influence it. This is an extremely important point, since there are countless factors that may govern the behavior of the time-series variable that cannot be identified or accurately measured. This last point indicates why time-series analysis is a preferred design over the currently used total population approach for assessing impact of DM programs.

There are three basic steps to developing a time-series model: (1) graphing the data using a sufficient number of observations to identify any patterns in the series that may assist the evaluator to identify the appropriate model, (2) choosing the correct model and fitting the data, and (3) evaluating the model prospectively by comparing predicted data with the actual data as they become available.

There are several possible patterns that may be identified by visual inspection of graphic displays of healthcare data:

1. *Trends.* This is the long-term movement of a data series that may slope either upward or downward.
2. *Seasonality*. This pattern emerges as spikes at regular intervals in a time series (usually monthly, quarterly, or annually).
3. *Stationarity*. A stationary series reflects data that have a constant variance around a constant mean. Therefore, a stationary series would not have a linear trend or seasonal component, but, instead, would appear as relatively horizontal along the *x*-axis.[6]

In the following sections, two categories of time-series designs will be presented [exponential smoothing and autoregressive integrated moving average (ARIMA)]. They will be described in order from the simplest to the most complex. The process for developing the simplest models can be performed using automated functions found in most statistical packages. As a result, these methods are ideally suited for projects in which many variables must be forecast (i.e., a DM program with many utilization categories). Moreover, these simple models do not require a theoretical understanding of advanced statistics.

The most complex models involve an iterative process requiring an understanding of the underlying statistical phenomena as well as sophisticated functions not always available in basic statistical packages. Consequently, these models take longer to build, which may limit the number of variables to be forecasted, and will require a person with expertise in statistics or business forecasting to develop them. However, these models are thought to be more accurate and versatile and allow for the addition of an independent variable representing an intervention effect.

EXPONENTIAL SMOOTHING

In general, exponential smoothing is an offshoot of the standard moving average technique. A moving average is simply the average of a predetermined number of past observations. As each new value comes available, the oldest observation is removed, and the average is then recalculated, including the newest value. Thus, the moving average always gives equal weight to all past values. Alternatively, exponential smoothing weights observations unequally, with more weight given to the most recent observations and less weight given to older values. Given that time series are serial-dependent, the most recent observations tend to have the largest impact on present and future observations; thus they need to be weighted more than earlier observations. Appendix A provides additional detail on the three exponential smoothing models discussed below.

Simple exponential smoothing (SES)

SES is the smoothing model most often used when the data series appears stationary (does not exhibit any seasonal variation or trend). In simple terms, with this model a given forecast is basically nothing more than the prior forecast with adjustments made for the error in that forecast (an error refers to the difference between the actual and predicted value). This adjustment process is analogous to a thermostat where corrections are made to alter the temperature if it is either too hot or too cold.

The process for manually determining model specification is tedious. Fortunately, most statistical software packages today can perform these functions by running through an algorithm that compares the forecasted values with actual observations, varying a weighting constant, to determine the model that provides the least amount of error (this process is usually referred to as a grid search or optimization). SES is only appropriate if there are no obvious trends or cycles in the series.

Double exponential smoothing (DES)

DES was developed by Holt[8] as an offshoot to the SES model as a means of adjusting for trends in the time series. It is referred to as a "double" because it estimates both the level (as in the SES design) and the trend. As is the case with the SES model, development of this model can be accomplished manually through a grueling iterative process, or can be left to the statistical package to determine via the grid search or optimization technique. DES is appropriate if there is a noticeable trend in the data but no obvious seasonality.

Holt–Winters multiplicative trend and seasonality model

Winters[9] broadened Holt's linear exponential smoothing model[8] by introducing an additional equation to cope with seasonality. The most complicated of the three exponential smoothing models introduced, the Holt–Winters model includes variables to account for randomness, trend, and seasonality components of the time series. As is the case with the previous methods, it is recommended that the evaluator use the software's grid search to find the best-fitting model parameters.

A final note on exponential smoothing: A model using historical data should fit those data quite well (since all the data are available and no forecasting is required). However, predicting DM program impact requires extrapolation to future behavior. Therefore, in order to find the best-fitting model for prospective observations, a model should be developed as illustrated in Figure 1. F1

As shown, the first step is to divide the initial data series into two sets: a historical data set (sometimes referred to as the initialization or training set), and a validation data set (also referred to as a hold-out sample or test set). The historical set should include a substantial number of past values in order to reduce the effect of variability observed at various points in the time series. The validation set should include about 12 data points following the historical set time period. For a DM program, this would be 1 year's worth of monthly observations.

The next step is to develop the model and generate forecasts (the number of forecasts should at least match the number of observa-

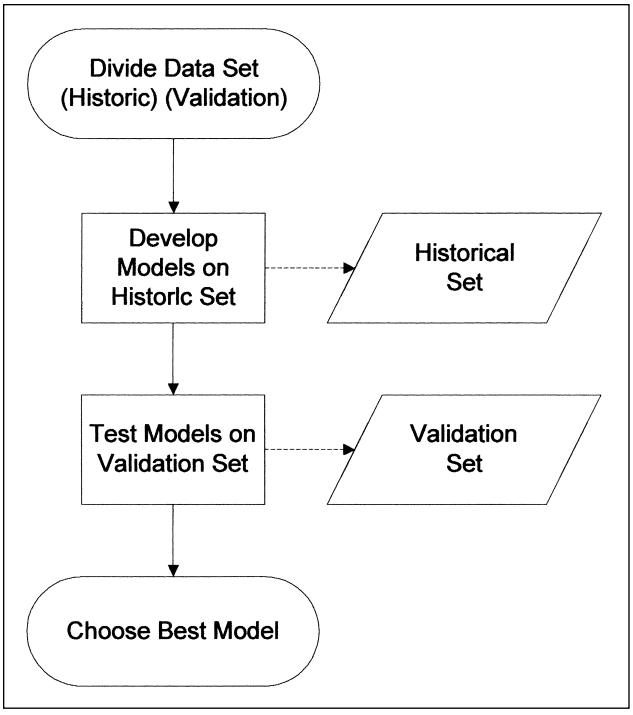

FIG. 1. The general method for developing and choosing time-series models.

tions in the validation set) using the historic data set. These forecasts are then compared with the actual data in the validation set and measured for accuracy. If several time-series methods are compared (i.e., SES, DES, or Holt–Winters), the model that provides the best fit should be chosen (measures of accuracy and identifying the best-fitting model vis-à-vis a comparison of actual versus predicted values will be discussed later in this paper). After model selection, but prior to use in the DM evaluation, both the historical and validation data set should be combined, and the model refit with all of the available data.

ARIMA MODELS

ARIMA models are the most widely used time series methods found in the health services literature.[4,10–16] This class of models was developed by Box and Jenkins,[17] and is intended to describe, mathematically, the changes in a series over time. While this model is quite complicated, the fundamentals of the design can be explained in basic terms. In the Box and Jenkins methodology, modeling of a time-series consists of three empirically driven phases: identification, estimation, and diagnostic testing.[7,17,18] Appendix B provides more detail on elements of ARIMA models discussed below.

Identification

This first phase of the model building process involves: (1) examination of the data to uncover any patterns, (2) if necessary, manipulation of the data to achieve stationarity, and (3) identification of potential models.

As opposed to the exponential smoothing technique, where different models are quickly developed using computer-generated algorithms and the best-fitting model is chosen, ARIMA modeling requires inspection of the data to uncover patterns before any model can be developed. While a visual inspection of a time-series plot may clearly identify a linear trend or a seasonal effect, most healthcare data contain enough variability to lead to unreliable results using this method. However, this task can also be performed empirically using a statistical tool called the autocorrelation function (ACF), which will produce more accurate results.

While the nomenclature alone may strike fear in the hearts of non-statisticians, the underlying principle supporting the ACF is quite straightforward. As discussed earlier in this paper, what differentiates time-series analysis from other statistical methods is that time-series models are built on the premise of relationships between observations (as opposed to independent explanatory variables). An ACF simply indicates how each observation is correlated to prior observations (hence the term autocorrelation). A review of the ACF allows patterns to be detected. For example, if a significant autocorrelation exists between every 12th observation, we can presume that annual seasonality is present in the time series (assuming the data are aggregated monthly). In order to build a successful ARIMA model, the data must be made stationary (i.e., no linear trends or seasonality are present). This is to allow the analyst to identify other underlying relationships present in the data that would otherwise be obscured by these factors. The most widely used technique for making a time series stationary is through a method called *differencing*, in which a new series is created using the *actual* difference between successive values ($X_t - X_{t-1}$). For example, first differences are the change between one observation and the next, and seasonal differences are the change from one year to the next (assuming the seasonal factor arises every 12 months). A series is considered stationary when the ACF shows no statistically significant autocorrelation existing between data points. Differencing of a time series is somewhat analogous to the technique, used in other statistical models, of transforming a dataset as a means of achieving normality so that parametric statistical tests may provide valid results.

Upon completion of these initial steps, the evaluator should have the necessary information to establish a provisional model. While there is a multiplicity of ARIMA designs, a basic model includes at least one of the following two components: (1) an autoregressive parameter that relies exclusively on past observations

of the outcome variable to forecast present or future observations, or (2) a moving average component that relies entirely on past errors (i.e., the difference between past forecasts and actual values) to explain the variation. Moreover, if the data indicate, these two components can be united to create an autoregressive moving average (ARMA) model. Typically though, the data must first be differenced, thereby creating an ARIMA model. The "I" stands for "integration," which refers to the fact that the original series may be recreated from a differenced series by a process of integration (involving a summation in the typical discrete environment).[6] As stated above, differencing of adjacent observations will remove a trend from the data, while differencing observations separated by 12 months will remove seasonality from the data.

Estimation

This past section has presented an overview of the elements of an ARIMA model. Once a provisional model has been identified, the evaluator can rely on the statistical software to generate the statistical estimates, including a test for significance for each parameter in the model. If any parameter is determined not to contribute to the model (i.e., is not significant), it may be eliminated as a means of improving the overall model fit.

Diagnosis

The final step in the ARIMA model-building process is to perform a review of the residuals, or errors (the difference between the forecasted values and actual values) vis-à-vis the ACF. If the model fits adequately, the ACF should show that there are no significant autocorrelations among the residuals. If there remain significant autocorrelations, the analyst must return to the identification procedures to assess whether all the patterns hidden in the data have been uncovered.

MEASURING ACCURACY BETWEEN VARIOUS MODELS

We have presented several different model types from SES to sophisticated ARIMA designs. The question remaining is how does one determine which model is the most accurate? Appendix C provides more detail on the options outlined below.

As mentioned earlier, all models can be made to fit historical data reasonably accurately; therefore it is important to test the "goodness of fit" of these models on test data. There are several different measures of accuracy for time series models: Mean absolute error, mean squared error (MSE), mean percentage error (MPE), and mean absolute percentage error (MAPE) are among the most frequently used tools. They are all founded on the basis of observed differences between the predicted values and the actual values in a series. The time-series model of choice should produce the lowest amount of error (or the lowest value using any of these methods) compared with other models tested.[6]

For DM, the MPE and MAPE are probably the most helpful measures of accuracy among these because they provide an interpretable product for the evaluator. It is easier to comprehend a MAPE or MPE of 4% than an MSE of 132. As such, comparisons of goodness of fit between the various models in this paper will be made using the MAPE and MPE.

Figure 2 illustrates a series representing admission rates for angina in a medium-size health plan in the Northwest United States. The values are presented as per thousand members per year (PTMPY) to adjust for population changes over time. Three time-series models were developed using the 48-month historic data period (January 1998–December 2001), and then validated against the test data in the first 6-month period of 2002. MAPE values were determined for each model on the test data set, and presented in the legend. As indicated, SES proved to have the lowest MAPE at 13.38%, followed closely by DES at 13.92%. For this data series, the ARIMA (1,0,1) model appears to be the least accurate method for forecasting angina admission rates. The reason that SES and DES provide an almost identical MAPE is because the data do not indicate a trend; therefore the equations used for the two models will produce similar results. Had the data uncovered an existing trend, we would expect the DES model to outperform the SES model.

F2

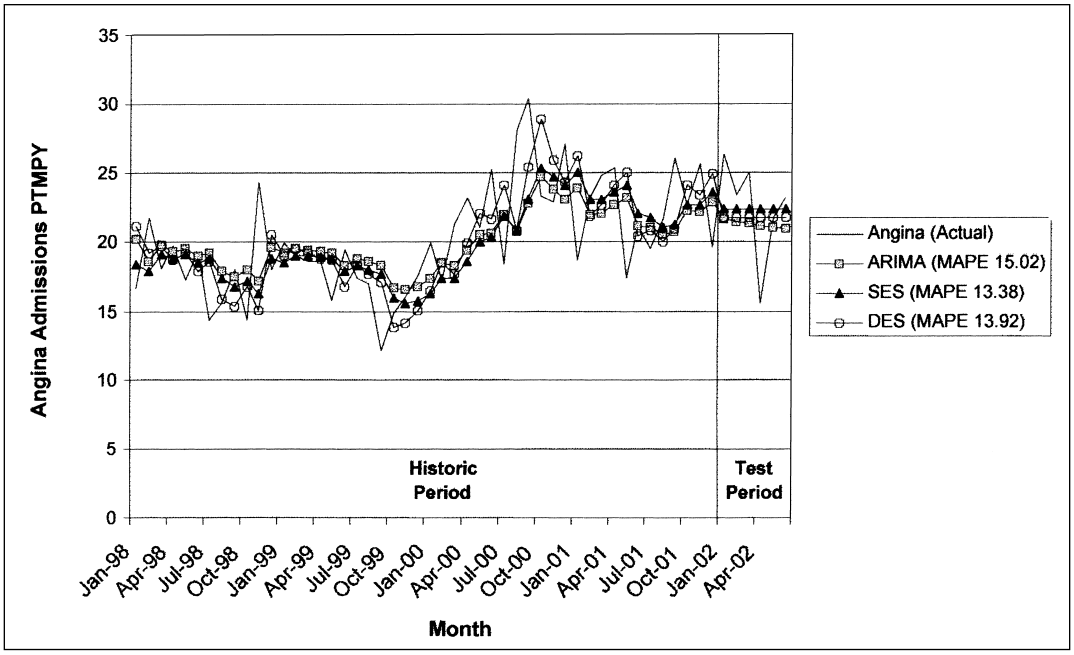

FIG. 2. A comparison of different forecasting methods for predicting angina admissions (PTMPY). The legend provides the MAPE value for each technique, which was determined using the 6-month test period data.

F3 Figure 3 represents admission rates (PTMPY) for acute myocardial infarctions in the health plan membership during the same time period as in Figure 2. For this time series, SES appears once again to elicit the most accurate fit. Interestingly, the ARIMA (0,0,1) was the second most accurate, and DES appeared to be the least accurate. The explanation for this switch between ARIMA and DES is most likely due to the spike in admissions during the winter

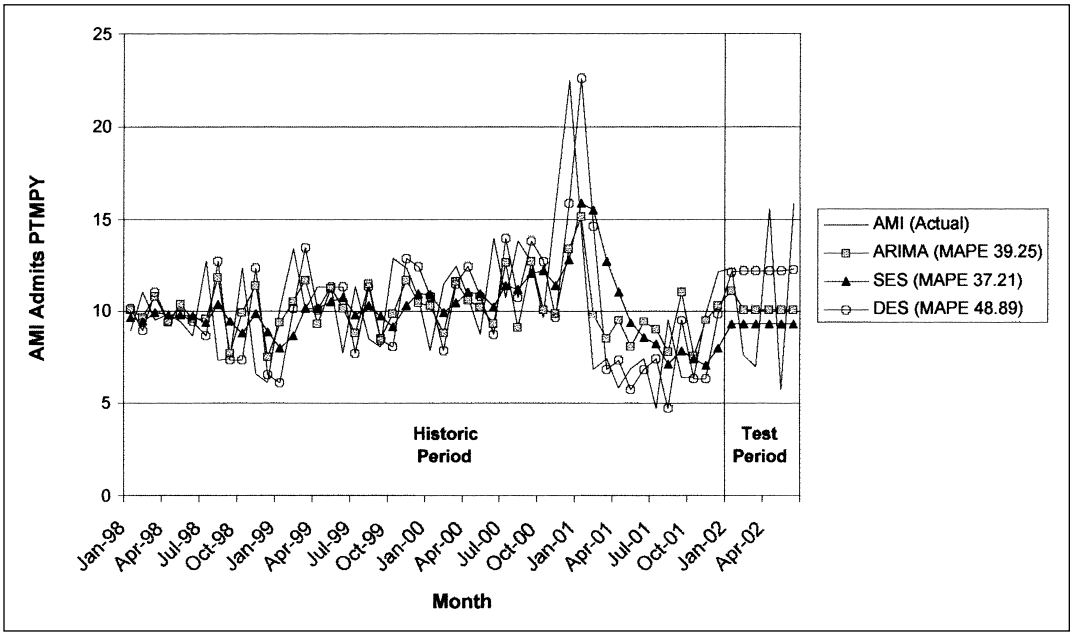

FIG. 3. A comparison of different forecasting methods for predicting acute myocardial infarction (AMI) admissions (PTMPY). The legend provides the MAPE value for each technique, which was determined using the 6-month test period data.

months of 2000. While we would expect this spike to be a seasonal factor, there is no similar peak in any other year. It could be that the DES calculation incorrectly determined that this spike and the subsequent dip were trend-related and developed the model specifications accordingly.

The result of these model comparisons is the identification of the model that best fits the data (e.g., lowest MAPE or MPE). As described above, depending on the type of data or if there is an indication of a trend or seasonality, a given model will fit the data better than others. This should be the model of choice.

APPLYING TIME-SERIES ANALYSIS TO DM PROGRAM EVALUATION

Thus far in the paper, the extensive process for developing and comparing various time-series designs to identify the best-fitting model has been presented. Figures 2 and 3 have illustrated the end result of this course of action. In this section we will provide some ideas for the practical application of this analysis to evaluate a DM program's effectiveness.

The most elementary launching point for this discussion lies in the determination of the appropriate outcome variable to be measured. The current inclination of DM programs and their health plan partners is to use financial indicators as a measure of program success. Generally, annual disease-specific medical costs are compared with pre-program costs, and success is considered achieved if medical costs decline. Unfortunately, as detailed by Linden et al.,[1] there are many factors that impact costs to make the outcome appear different than was actually achieved. For example, changes in the unit cost of services, members' financial share of the medical expense, introduction of new technologies, etc., are just a few of the confounding effects that may cause annual costs to change, irrespective of the DM program intervention.

It is for this reason that we suggest using disease-specific utilization measures as indicators of program success. While rising costs may be due to many uncontrolled-for variables, a decrease in utilization should be considered evidence of a DM program's intervention. By measuring the specific utilization variables that a DM program intends to impact directly, the evaluation should draw the appropriate conclusions from the data analysis.

The process for building the appropriate model has been detailed in previous sections, so only the important highlights will be repeated here: (1) In developing the best-fit model, at least 50 observations should be analyzed. For a DM program, this would require at least 4 years of past data leading up to the month prior to the commencement of the intervention. This will allow the model to accommodate any patterns in the data that may impact the fitting parameters. (2) The forecast period should extend to no longer than the first 12 months of the program to ensure an accurate forecast horizon. (3) Finally, MAPE and MPE should be used as the measures of model accuracy, so that results have a tangible meaning to the program evaluator.

How would program success be determined using a time-series design? The simplest of methods include plotting out the actual monthly observations for the annual intervention period and then comparing them with their forecasts. Finally, an MPE would be calculated to assess the divergence between the actual and predicted series for the measurement period. Typically, DM programs offer some expectation of decreased utilization, expressed as a percentage. If, for example, a DM program contends that they can reduce disease-specific hospitalizations for the year by 30%, we would expect to see the actual series eventually running 30% below the forecast series.

More specifically, DM programs should evaluate their historical datasets to identify expected outcomes at each point along the continuum. For example, the first year of a DM program includes program launch and a massive enrollment process. As such, this period would not be expected to show a significant intervention impact at the patient level. On the other hand, the program in the second year should heavily impact patient-level (and thereby aggregate-level) outcomes. Once these outcome levels are identified, specific targets can be determined for each contract period. Us-

ing the MAPE or MPE, the desired outcome would be expressed as a percentage difference between the actual result and what was predicted, assuming that no intervention was implemented.

Following this rationale, the evaluator can reset the model every year and renew the forecasts for the following year, as illustrated in Figure 4. As shown in this hypothetical data set, (1) observations from the historical data set were used to develop the time-series model, (2) forecasts were then produced for the test period or baseline (which is the 12 months immediately prior to the commencement of the DM program), (3) actual values were compared with predicted using the MAPE (in these data, the best-fitting model was determined to be the SES, based on that model producing the lowest percentage error compared with the other models), (4) the model was then recalculated adding the actual observations from the baseline period to the historical set, (5) forecasts were produced for the first measurement year, (6) at the end of that program year MPE was calculated, and finally (7) the incremental changes calculated from each program year are summed into an estimate of the current year's total change. MPE is used as the measure of DM effectiveness in the first and all subsequent measurement periods in order to provide the direction of the percentage change in the utilization variable. A value less than 0 would indicate that the DM program had a positive effect on reducing utilization, while a value greater than 0 would indicate that the DM program had no effect. Conversely, MAPE uses absolute values, rendering it difficult to assess the direction of change in the variable.

In this example, the MPE showed that admission rates were 3.6% higher than what was predicted for the period, indicating that the DM program was not successful in reducing admissions in this period. Note that in Figure 4, the first measurement period showed actual observations both above and below the forecast line. Since MPE takes into account the direction of the value (e.g., positive and negative numbers), it is possible (and likely) that some observations will cancel others out. In this case, the higher than expected observations from February through June cancelled out the lower than expected values of January and July. Therefore the overall average MPE resulted in a net 3.6% higher than expected admission rate.

The simplest method for recognizing program success is to determine whether the target was met. For example, if the target for a given period is a 5% reduction in actual uti-

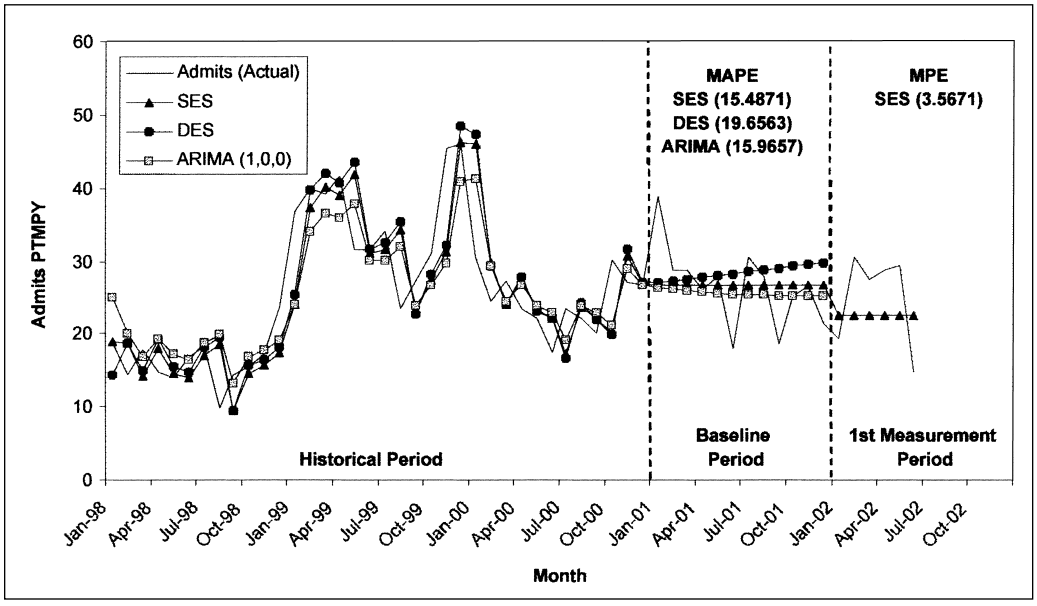

FIG. 4. An illustration of how time-series analysis can be used for evaluating DM program effectiveness. The observations are disease-specific hospital admissions (PTMPY).

lization compared with predicted for the period, and the MPE indicates that the actual utilization was higher than this predicted value (meaning that the MPE value was less than –5.0%), we can say that the program did not meet its goal for the period.

There are other methods for defining program success, such as determining whether a statistically significant difference exists between periods, or between actual and predicted values for a given period. In the business setting, however, statistical significance is not as meaningful a measure of success as meeting (or not meeting) a target value. Nonetheless, there are nonparametric statistical tests such as bootstrapping methods that can be used if this option is considered appropriate for the evaluation.

Conversely, the reconciliation method that is currently being used by the DM industry compares each measurement period with the baseline, rather than evaluating year-over-year changes. In other words, if DM program administrators promise a 10% reduction in medical costs, they are referring to a comparison of costs at any point in the program life with the baseline measurement period. The primary concern with this method is that changes in the level of observations occur at different points along the continuum as a result of the DM program intervention. With the current approach, one cannot anticipate or determine when these changes occur, or adequately prepare for them. While hypothetical, Figure 4 indicates that the intervention has, potentially, a different impact on the outcome variable in each measurement period. Understanding these time effect differences will allow DM program providers and purchasers to more accurately set and evaluate performance targets.

THREATS TO VALIDITY

The basic rationale for using time-series analysis in lieu of the total population approach for evaluating DM program effectiveness is that analyses of utilization variables over many time periods can demonstrate patterns of response to the intervention where the effects of confounding variables, including the variety of interactions between the individual and the environment, make determining causality extremely difficult. As a result, the time-series design controls for most of the factors that the currently used total population approach cannot control for. Campbell and Stanley,[19] Cook and Campbell,[20] and Shadish et al.[21] provide a detailed comparison of the threats to validity of the pretest–posttest design and the time-series model.

QU1

QU1

Maturation (a term describing the natural effect of the aging process or disease progression) is ruled out as a factor because it would have been identified and incorporated into the model-building process using the historical data set. *Regression to the mean* is implausible for explaining changes in a series because over a substantial number of time periods the effects of regression would be nullified, as opposed to a distinct pre- and post-measurement period. This also holds true for controlling the effects of s*election bias, loss to attrition, testing, and instrumentation* (in this context, we refer to the measurement methodology). Finally, the threat to external validity—*generalizability*—is not of particular concern in a DM program evaluation using a time-series design. The program is intended to impact a specific predefined population, and the results are not meant to be extrapolated to other disease states or diseased groups.

While the time-series methods control for most of the factors that may threaten the validity of the results, there are some elements that require extra diligence. The primary threat to internal validity is the effect of *history*—the possibility that certain confounding factors, other than the DM program intervention, impacted the measurement. For example, the introduction of an innovative surgical device that guarantees less pain and quicker recovery times may increase the hospitalization rate for a disease-specific population. Similarly, an inadequate amount of influenza vaccine during the flu season may increase hospitalizations or emergency department visits for high-risk patients who are more likely to contract the flu during those months. These issues, once documented, can provide the necessary explanation for any unexpected spikes in the data set. Given this concern, it is im-

portant to carefully select utilization measures to be as specific as possible to the DM program of interest. This will reduce the likelihood that externalities will confound the evaluation results.

Finally, there are additional design features that can be added to the standard time-series analyses, including the addition of nonequivalent no-treatment control group time series, nonequivalent dependent variables, multiple replications, or switching replications. These designs may help the evaluator build better causal inferences between the effect of the intervention and the outcome variable, while reducing the potential threats to validity of the program results.

Since the detail of these models is beyond the scope of this paper, the reader is referred to Shadish et al.[21] for a comprehensive description of these procedures and their analytic value.

CONCLUSIONS

This paper has described in some detail the application of time-series methods to the evaluation of DM program effectiveness. These techniques are ideal for measuring the impact of such programs that exist in the natural setting, and are subject to the influence of many factors beyond the control of the evaluator. For this approach to be successful it is crucial that each measurement period be compared with the forecasts developed in the prior measurement period. This runs contrary to the current practice of comparing results from each measurement period with the baseline. Similarly, the use of specific utilization variables as a measure of program effectiveness was proposed as an alternative to medical costs, since any positive change that occurs in disease-specific utilization patterns should, rightfully, be attributed to the program intervention. Conversely, the effect of confounding is amplified by the numerous variables impacting cost. There are fewer threats to validity using this methodology, as opposed to the total population approach, where the threats are numerous and cannot be controlled.

APPENDIX A: EXPONENTIAL SMOOTHING

SES

The general SES formula is:

$$F_{t+m} = \alpha Y_t + (1 - \alpha)F_t \tag{1}$$

where F_{t+m} is the forecast value for period $t + m$ based on the observation Y_t, with a weighting constant (most statistical software packages estimate the weighting constant to provide the model with least amount of error) α (in a range from 0 to 1), and weighting the forecast F_t with a weight of $1 - \alpha$. If α is close to 1, a considerable adjustment will be made for the error in that prior forecast. On the other hand, little adjustment will be required if the α is close to 0.

DES

The basic equations for Holt's method[8] are:

$$L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + b_{t-1}) \tag{2}$$

$$b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1} \tag{3}$$

$$F_{t+m} = L_t + b_t m$$

As shown, Eq. 3 is similar to Eqs. 1 and 2 except that the smoothing is done on the trend, and not on the actual data points. The trend, b_t, is multiplied by the number of periods that will be forecasted ahead, m, and then added to L_t, the current level of the data. Therefore, the only significant difference between the DES and the SES is the added equation to control for trend.

Holt–Winters multiplicative trend and seasonality model

The basic formulae are as follows:

$$L_t = \alpha \frac{Y_t}{S_{t-s}} + (1 - \alpha)(L_{t-1} + b_{t-1}) \tag{5}$$

$$b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1} \tag{6}$$

$$S_t = \gamma \frac{Y_t}{L_t} + (1 - \gamma)S_{t-s} \tag{7}$$

$$F_{t+m} = (L_t + b_t m)S_{t-s+m} \tag{8}$$

The only significant difference found in this equation as compared with the previous for-

mulae is the estimate of seasonality (Eq. 7), which is given as an index fluctuating around a value of 1. Nonetheless, Eq. 7 is similar to all other smoothing equations in that a value $\frac{Y_t}{L_t}$ is multiplied by a constant γ and is then added to its previous smoothed estimate that has been multiplied by $1 - \gamma$. An important difference is that the S_t requires more data to estimate well since we only learn more about them once per year. For the Holt–Winter method, α is used to smooth the randomness found in the level of the series, β is used to smooth trend, and γ is used to smooth the seasonality component of the time series.

APPENDIX B: ARIMA MODELS

Identification

The notation for an ARIMA model found in the literature is ARIMA (p,d,q), where p is the number of autoregression parameters, d is the order of differencing required to produce stationarity, and q is the number of moving average parameters. Thus, an ARIMA (1,0,0) indicates an autoregressive model that predicts the current observation based on the preceding observation, ARIMA (0,0,1) indicates a moving average model that predicts the current observation based on the previous prediction and the previous error, and ARIMA (1,1,0) indicates an integrated autoregressive model that is a nonstationary series that was differenced once, after which the preceding observation is used for prediction of the current observation.

If there is found to be a seasonal element to the data during the identification process (i.e., during inspection of the ACF), an ARIMA seasonal model should be developed in the same way as just explained. The ultimate model will be distinguished by the following notation: ARIMA $(p,d,q)(P,D,Q)_X$, where (P,D,Q) refers to the seasonal component, and X indicates the number of periods between seasonal spikes (i.e., $X = 12$ in an annual cycle with monthly data and $X = 7$ in a weekly cycle with daily data).

Note that differencing is only one method to achieve stationarity. In any given forecasting problem there may be fewer mechanical ways to transform the data to stationarity. If the increase in the series is a simple consequence of inflation, stationarity could be achieved by transforming the data to constant dollars. If the nonstationarity is the result of changing fee schedules, changing the modeling from dollars to hospital admission counts or some other combination of utilization measures may achieve stationarity.

APPENDIX C: MEASURING ACCURACY BETWEEN VARIOUS MODELS

The equations for MAPE and MPE are as follows:

$$\text{MAPE} = \frac{\sum_{abs}|(y_t - \hat{y}_t)/y_t)|}{n} \times 100 \qquad (9)$$

$$\text{MPE} = \frac{\sum|(y_t - \hat{y}_t)/y_t|}{n} \times 100 \qquad (10)$$

As shown, the only difference between these two measures is that the MAPE removes the sign, so that positive and negative values cannot cancel each other out.

REFERENCES

1. Linden A, Adams JL, Roberts N. An assessment of the total population approach for evaluating disease management program effectiveness. Dis Management 2003;6:93–102.
2. Faithfull S. Analysis of data over time: a difficult statistical issue. J Adv Nurs 1997;25:853–858.
3. Jirovec MM. Time-series analysis in nursing research: ARIMA modeling. Nurs Res 1986;35:315–319.
4. Tsouros AD, Young RJ. Applications of time-series analysis: a case study on the impact of computer tomography. Stat Med 1986;5:593–606.
5. Chatfield C. The analysis of time series, 5th ed. London: Chapman and Hall, 1996.
6. Makridakis SG, Wheelwright SC, Hyndman RJ. Forecasting: methods and applications, 3rd ed. New York: John Wiley and Sons, 1998.
7. McCleary R, Hay R. Applied time series analysis for the social sciences. Beverly Hills, CA: Sage, 1980.
8. Holt CC. Research memorandum no. 52: forecasting seasonal and trends by exponentially weighted moving averages. Bethesda, MD: Office of Naval Research, 1957.
9. Winters PR. Forecasting sales by exponentially weighted moving averages. Management Sci 1960;6: 324–342.

10. Choi K, Thacker SB. An evaluation of influenza mortality surveillance, 1962–1979. I. Time series forecasts of expected pneumonia and influenza deaths. Am J Epidemiol 1981;13:215–226.
11. Gibson E, Fleming N, Fleming D, et al. Sudden death syndrome rates subsequent to the American Academy of Pediatrics supine sleep position. Med Care 1998;36:938–942.
12. Haines LM, Munoz WP, Van Gelderen CJ. ARIMA modeling of birth data. J Appl Stat 1989;16:55–67.
13. Linden A, Schweitzer SO. Using time series ARIMA modeling for forecasting bed-days in a Medicare HMO [abstract]. AHSRHP Annual Meeting 2001;18:25.
14. Martinez-Schnell B, Zaidi A. Time series analysis of injuries. Stat Med 1989;8:1497–1508.
15. Zechnich AD, Greenlick M, Haxby D, Mullooly J. Elimination of over-the-counter medication coverage in the Oregon Medicaid population: the impact on program costs and drug use. Med Care 1998;36: 1283–1294.
16. van Walraven C, Goel V, Chan B. Effect of population-based interventions on laboratory utilization: a time series analysis. JAMA 1998;280:2028–2033.
17. Box GEP, Jenkins GM. Time series analysis: forecasting and control. San Francisco, CA: Holden Day, 1976.
18. Helfenstein U. Box-Jenkins modeling in medical research. Stat Methods Med Res 1996;5:3–22.
19. Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.
20. Cook TD, Campbell DT. Quasi-experimentation: design and analysis issues for field settings. Chicago: Rand McNally College Publishing, 1979.
21. Shadish SR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin, 2002.

Address reprint requests to:
Ariel Linden, Dr.P.H., M.S.
Director, Clinical Quality Improvement
Providence Health Plans
3601 SW Murray Boulevard, Suite 10
Beaverton, OR 97005

E-mail: ariel.linden@providence.org

LINDEN

QU1
Sentence structure OK for sense?