INTRODUCTION

As wellness, case, and disease management programs proliferate, health plans must measure their outcomes and potential for return on investment. Here are a few strategies to help with that evaluation.

Recent years have witnessed a proliferation of wellness, case, and disease management programs in managed care. Despite their prevalence, few are subjected to the rigorous research designs necessary to determine whether financial goals are achieved. As a result, the effectiveness of these programs in reducing costs remains to be determined (Linden 2006; Mattke 2007).

The lack of stringent evaluation of these programs is often attributed to difficulty in identifying a suitable control group (Linden 2006; Linden 2003). Program evaluation in managed care has therefore tended to rely on observational pre-post studies that are subject to selection bias and regression to the mean (Linden 2006).

Program evaluators often resort to using a pre-post design with a relatively small percentage of enrolled members, comparing utilization for a specified pre-enrollment period with a similar post-enrollment period, or they compare the results of members enrolled in the program with eligible members who were unable to be enrolled. Each of these designs has critical weaknesses.

The pre-post design is highly subject to regression toward the mean. Members are identified for intervention as a function of being a utilization outlier. Regression toward the mean suggests that outliers at one point in time are likely to be closer to the mean at re-measurement, even in the absence of intervention. As a result, reductions in cost and utilization cannot be fully attributed to the intervention.

Comparing enrolled members to unenrolled members is subject to selection bias. Members who cannot be enrolled in phone- or mail-based interventions may be fundamentally different from enrolled members. Unenrolled members are unreachable or refuse to accept services. Being unreachable may suggest greater instability in housing or telephone service, or greater financial and psychosocial stressors. Or refusal to accept a free intervention may reflect a low level of motivation to improve health. If enrolled members have fewer financial and psychosocial issues, or are inherently more motivated to improve their health, these differences might explain group outcomes, rather than the intervention.

Given the importance of managing health care costs, it is critical that administrators appropriately evaluate outcomes of managed care programs aimed at reducing costs.

Measurement strategies are available that avoid regression toward the mean and selection bias, yet they are frequently overlooked. The purpose of this paper is to provide a framework for program evaluators to select the strongest, most feasible design for program evaluation and to provide general strategies to improve program evaluation.

OUTCOME DESIGNS

Several designs can be implemented to objectively measure outcomes. The strongest is the randomized controlled design. When randomized control is not possible, evaluators may choose to employ a historical, intent-to-treat design or, when that is not possible, a matched comparison design. Each design has limitations; selecting the appropriate one depends on the context of the evaluation. The figure on page 55 illustrates a flow chart for selecting a design. The table on page 56 provides an example of each design.

Randomized controlled design
Description. A randomized con-
trolled trial evaluates whether a cause-effect relationship exists between an intervention and an outcome (Sibbald 1998). These trials involve random allocation of patients or subjects to groups that are treated identically in all ways except the intervention (Sibbald 1998).

**Limitations.** This design is difficult to use in the managed care setting, as it is typically not feasible to withhold an intervention from one group of members.

**When to use.** Randomized controlled trials should be used in accordance with each organization’s policies and internal review board recommendations.

Although often not feasible, it may be possible that an intervention can be rolled out in stages, as opposed to making it available to an entire population at once.

If the outcome period is sufficiently small, an organization may be able to randomly identify a sample in which to administer the intervention while comparing outcomes with a random, wait-list control sample.

This requires a short outcome period and an intervention that does not require all plan members to receive services simultaneously.

**Strategies.** Within the managed care environment it is often difficult to utilize the randomized controlled design because withholding a potentially useful intervention from a member is unappealing.

Utilizing a wait-list control group and clearly presenting to a member assigned to this condition that he or she will receive the intervention within a reasonable amount of time and answering any questions a mem-

---

**FIGURE**

Outcome design flow chart

- Is random design feasible?
  - Yes: Use random design
  - No:
    - Will intervention be received by significant % of eligible members?
      - No: Implement interventions to increase %
      - Yes: Use historical, intent-to-treat design
    - Determine key variables for matching
    - Ensure unmatched variable(s) don’t significantly affect outcomes variables(s)
    - Use matched control design
number has may enhance the feasibility of this design.

Shortening the measurement outcome period may be necessary to reduce wait time for members.

**Historical, intent-to-treat design**

**Description.** A historical, intent-to-treat design involves the creation of a retrospectively identified cohort that would have been identified for an intervention had the program existed during a period of time before program implementation. Once these retrospective members are identified, all clinical, utilization, and financial data are obtained for the given outcome period. This group becomes a baseline group to be compared to the intervention group — members eligible for the intervention as of the program’s inception. Using the same identification criteria for the intervention and baseline group prevents regression toward the mean from disproportionately affecting one group, as this phenomenon will affect both groups similarly.

The “intent-to-treat” component involves the inclusion of each eligible member in the intervention group, regardless of whether the member was enrolled in the program. Because it is not possible to know which members in the baseline comparison group would have been enrolled had the program existed at the time, all eligible members during the actual intervention year must be included to avoid a selection bias.

**Limitations.** Changes associated with time, such as policy changes, new legislation, and inflation may have a greater influence on one of the groups than on another or others, as each exists at a different point in time. Another limitation is that to achieve statistically significant differences between the groups, a considerable proportion of eligible members must receive the intervention, since all members intended to receive the intervention are included in the analysis. Unfortunately, many programs enroll only a small percentage of eligible members.

**When to use.** The historical, intent-to-treat design may be used when a randomized trial is not feasible, a historical cohort can be identified, the impact of time is minimal or can be controlled for, and enrollment is high.

**Strategies.** For this strategy to be effective, a number of elements must be observed or considered.

*Use the same criteria.* Allow no difference between the criteria used to identify the historical baseline comparison group and the criteria used to determine the intervention group.

*Confounding variables.* Identify and control for confounding variables associated with time. Identify and, to the extent possible, control for factors that disproportionately affect one group. But at times there may be good reason to ignore this rule. One can adjust for inflation, for example. Since inflation tends to increase costs in the intervention group, which exists later in time, an evaluator may decide not to adjust for inflation. This would make it more difficult to achieve statistical significance and would yield a more conservative estimate of program benefits.

**Enhance program enrollment.** Determine the percentage of eligible members required to receive the intervention before the entire intent-to-treat group will have a chance to demonstrate significant improvement. Once this is estimated, program administrators should do everything possible to enhance program enrollment beyond that level. Fortunately, strategies are available to increase program enrollment, such as using search engines to find correct contact information, interactive voice

<table>
<thead>
<tr>
<th>TABLE</th>
<th>Examples of different study designs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design</td>
<td>Reference</td>
</tr>
<tr>
<td>Randomized controlled design</td>
<td>Rosenzweig et al. 2010</td>
</tr>
<tr>
<td>Historical intent-to-treat design</td>
<td>Kolbasovsky 2009</td>
</tr>
<tr>
<td>Matched comparison design</td>
<td>Dubois et al. 2010</td>
</tr>
</tbody>
</table>
recordings, contacting members in the evenings and while admitted to the hospital, collaborating with practitioners, involving family members, using opt-out approaches, and motivational interviewing.

**Matched comparison design**

**Description.** The matched comparison design involves identifying a control group of members matched for one or more variables thought to influence outcomes.

**Limitations.** An important limitation to this design is its susceptibility to a motivation selection bias. Members receiving the intervention were motivated enough to accept the intervention and had a stable enough living situation to be reachable at that time. This may not be true for the matched group. Unless a proxy for likelihood of participating in a program is available among members in the matched group, controlling for motivation to participate is difficult.

**When to use.** This design is best used when the previous designs are not possible; when members can be matched on variables that are likely to affect outcomes; when stability of living situation is similar between the groups; and when motivation is not important to achieving desired outcomes or a proxy for motivation is available.

**Strategies.** Several strategies can be employed to enhance the effectiveness of this design.

*Match on risk as well as demographic variables.* Demographic variables such as age, gender, and line of business are often used in matching. However, if a program outcome variable is related to utilization or costs, it is important to match on a variable that represents risk of future utilization, such as predictive model risk scores, comorbidity indexes, or measures of previous utilization.

*Use multiple matches per member.* Select matching variables carefully. After two or three variables, it often becomes difficult to find matches.

**MEASURING OUTCOMES**

All of these designs require certain strategies for measuring outcomes. Strategies can vary between designs, but should include the following elements.

**Design a strategy.** Design a strategy for measuring outcomes before program implementation.

The design selected should be decided after considering both the strengths and limitations. The design selected must be feasible yet able to provide valid and reliable results. Failing to consider these issues and designing a measurement strategy before program implementation often leads to poor design and problems that arise too late to correct.

**Include balancing measures.** Balancing measures assess how interventions aimed at improving one part of a system may be causing problems in another part (Institute for Healthcare Improvement 2010). For example, an intervention aimed at reducing the number of prescribers of controlled substances may increase ER visits and costs if members who are unable to get desired prescriptions from multiple outpatient physicians turn to the ER for service. Evaluators focusing exclusively on intended outcomes may miss the true impact of the intervention.

**Include program costs.** Many evaluators report cost savings associated with program interventions while failing to consider program costs. These may include staff salaries/benefits, operating expenses, and increased costs associated with balancing measures.

**CONCLUSION**

Managed care administrators can evaluate programs more effectively by using the most scientifically rigorous research design feasible. By identifying and implementing an outcome measurement strategy before program implementation, administrators can evaluate program outcomes appropriately and use these results to allocate resources effectively and to make informed decisions about reducing health care costs and improving quality.

**REFERENCES**


