DISEASE Management

PREPUBLICATION COPY

Generalizing Disease Management Program Results: How to Get From Here to There

Ariel Linden, DrPH, MS, John L. Adams, PhD, Nancy Roberts, MPH

For a disease management (DM) program, the ability to generalize results from the intervention group to the population, to other populations, or to other diseases is as important as demonstrating internal validity. This article provides an overview of the threats to external validity of DM programs, and offers methods to improve the capability for generalizing results obtained through the program. The external validity of DM programs must be evaluated even before program selection and implementation are begun with a prospective new client. Any fundamental differences in characteristics between individuals in an established DM program and in a new population/environment may limit the ability to generalize.

Although disease management (DM) has existed for more than 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 ambiguity concerning the best way to evaluate DM program effectiveness. Although several methods have recently been proposed,[1–7] the DM industry heretofore has remained steadfast in its support of the standard pre–post model for addressing financial outcomes,[8] despite its vulnerability to myriad biases.[1]

The literature about DM program evaluation has primarily focused on the threats to internal validity (e.g., can one rightfully attribute the program's outcome to its intervention, or was it influenced by confounding factors or biases?). However, little attention has been paid to the issue of external validity, or generalization of results. External validity concerns inferences about the extent to which a causal relationship is sustained over variations in persons, settings, treatments, and outcomes.[9]

The concept of generalization is important to DM programs for the following reasons: Disease management programs, by design, create selection bias. In general, DM programs strive to enroll individuals at greatest risk of incurring high costs or utilization during the program term, thereby creating an unbalanced case mix between enrolled and nonenrolled groups. As a result, outcomes may not be accurately extrapolated to healthier members, or to members excluded from the program.

In addition, disease management programs often use the whole population in the outcomes analysis, but intervene in only a subset of that disease population. This is related to the first point; if the program participants do not adequately represent the population from which they were recruited, accurate inferences cannot be drawn about the larger population on the basis of the program's results.

Disease management program vendors promise, and their clients expect to achieve, similar outcomes. Health plans or other purchasers typically start their search for a DM program with companies that have achieved successful outcomes. Moreover, they expect to realize the same positive results as the vendor's best-case scenario. These assumptions can be validated only if variations in region, practice patterns (including what care is provided, and where and how), and patient demographics (including differences in culture and beliefs)[5] are considered.

The DM industry generalizes results from one DM program to all programs that address the same

Dr. Linden is President, Linden Consulting Group, Portland, Oregon. Dr. Adams is Senior Statistician, RAND Corporation, Santa Monica, California. Ms. Roberts is Regional Director of Integrated Performance/Six Sigma Champion of Managed Care, Providence Health System, Portland, Oregon.

disease. It is often heard, for example, that DM programs for asthma do not provide a good return on investment, or that DM programs for diabetes do not realize short-term savings. These statements are inaccurate unless all programs used the same interventions, were evaluated by the same methodology, and were assessed in terms of relevant threats to external validity.

The DM industry generalizes results from interventions applied to one disease to other diseases. For example, programs that manage multiple diseases may generalize results from a telephonic nursing intervention in chronic obstructive pulmonary disease (COPD) to congestive heart failure (CHF) or even to patients with several comorbid processes.

Since there is little information in the DM literature that addresses external validity, the purposes of this article are to provide an overview of the types of threats to external validity, illustrate how these and other threats to external validity are manifested in DM programs, and provide methods to improve the capacity to generalize the results gained from a program intervention.

THREATS TO EXTERNAL VALIDITY

As described earlier, external validity concerns inferences about the extent to which a causal relationship persists over variations in persons, settings, treatments, and outcomes. According to Shaddish and colleagues,[9] five general categories of reasons have been identified as to why study results cannot be generalized. One important point is that interactions between variables may actually influence several factors simultaneously—that is, a treatment may have different effects on units in different settings, ultimately affecting the outcome. (A unit may be a patient or doctor; a larger entity, such as a medical group, employer, or health plan; or even a region or country.)

Interaction of the Causal Relationship With Units. This is the best-known threat to external validity when the topic of generalizing results is discussed. An example of how differences among individuals may affect generalization is illustrated by a recent study in which survival rates of patients with cancer who had successfully undergone chemotherapy for T2–T4a bladder cancer in a clinical trial were later compared with two other groups: (1) patients who refused to participate in the trial and (2) those who were deemed ineligible for the trial.[10] Even though both of these groups received the same treatment as patients in the clinical trial, the survival rate was significantly better in the clinical trial participants. These findings indicated that there was a selection bias in the trial-enrollment process, which ultimately compromised generalization of the results to patients with cancer outside of the clinical trial.

The ability to generalize among larger units is illustrated by a recent study in which the external validity of a successful, classroom-based drug abuse–prevention program (developed for youths at high risk of drug abuse) at alternative high schools was tested in a general high school.[11] At year 1 of follow-up, the program was effective in reducing alcohol and illicit drug use in the general school population as well, indicating that the program could be generalized.

An example of generalizing across a large domain is a study in which a questionnaire developed in the United States to examine a theoretical model of sexual harassment was given to a large group of Turkish women.[12] Despite the cultural differences between the two countries, the statistical analysis indicated that the model was successfully generalized to the Turkish culture as well.

On an individual level, participants in a disease management program intervention are not typical of the population from which they are drawn.

Parallel examples can be readily found in DM. On an individual level, participants in a DM program intervention are not typical of the population from which they are drawn. They are usually sicker than the universe of patients with that disease and, depending on the illness, they are likely to be older as well (e.g., to have CHF or COPD). As in the cancer study example, DM program participants and persons in the general population may achieve different results with a similar intervention, simply because the characteristics of the two groups are fundamentally different.

The high school study is comparable to a DM program intervention in that a specific instruction is provided to a group of participants who may have different characteristics than other groups. In DM, results may differ across patient populations (e.g., commercial insurance vs. Medicare vs. Medicaid), providers (e.g., medical group practice vs. walk-in clinic vs. solo practitioner), or across location (e.g., urban vs. rural). Similarly, it has been well demonstrated that such factors as belief systems and perceived need are significant drivers of how individuals access health care.[13,14] If not controlled for, these variables may very well result in different outcomes among different people, even though the same intervention is used.

The sexual harassment study exemplifies the need to demonstrate the effectiveness of DM program interventions across cultural and regional domains. The United States comprises many subcultures,

dictated by both regional (e.g., the North vs. the South, the East vs. the West) and cultural divides (Hispanic populations in the Southwest, Caucasian populations in the Midwest, and African-American populations in the South). It is imperative, when implementing a DM program, to determine whether it can be effective across these domains, or if success is limited to a certain segment of the population.

Interaction of Causal Relationships Over Treatment Variations. Variations in treatment relate to the intervention itself (as opposed to the unit, discussed above): Will a given intervention achieve the same results if administered to different persons, at different times, or in different settings?

A recent study investigated whether a patient self-assessment of difficulty across a range of mobility situations, which was originally designed, developed, and validated in patients with retinitis pigmentosa, would be equally valid when used as a measure of independent mobility as perceived by patients with glaucoma.[15] After the results had been subjected to numerous validity and reliability analyses, the assessment was determined to also be valid for patients with glaucoma.

The site where the treatment or intervention is conducted may play a major role in determining if the program's results can be generalized.

Use of an intervention in one set of patients that was originally designed for use in another set is analogous to a standard DM program intervention that is used across all diseases and populations. The behavioral-change model most often used in DM today is the stages-of-change model (SCM), which, although useful, does not apply to all diseases or patients equally.[5] It is imperative that the applicability of a given intervention to the targeted disease state or population be examined before its implementation on a populationwide basis.

Some interventions are time sensitive, meaning that achieving the desired outcome requires that patients be exposed to the stimulus for a given period of time. The SCM is a good example.[16] The theory underlying it is that behavioral change is a continuous process, as opposed to a one-time event, and that individuals are at varying levels of readiness to change along the continuum of five stages. For maximum effectiveness, the DM nurse should identify which stage of change the patient is currently in and then provide the cues or education appropriate to that stage. For this intervention to succeed, members must be engaged for a long enough time to facilitate movement through the stages and ultimately achieve the desired behavioral change. A high turnover rate in the client membership or insufficient duration of a DM program may reduce the likelihood of achieving the expected results.

Interaction of Causal Relationship With Settings. The site where the treatment or intervention is conducted may play a major role in determining if the program's results can be generalized. For example, emergency department (ED) screening for sexually transmitted diseases can be an effective intervention for unrecognized infections. A recent study compared screening rates in an inner-city ED with those in a suburban Baltimore ED.[17] The prevalence of infection did not differ between the two settings, but participation in the screening, distribution of race/ethnicity, health care access, and behavioral risks all were significantly different. The authors correctly concluded that an assessment of population characteristics is necessary to develop appropriate screening methods and increase patient participation.

Another example from the literature in which the setting may have played a role in capacity to generalize results was a study of the characteristics of companies that either accepted or declined participation in a five-year, randomized trial of a multiple risk-factor health promotion intervention.[18] Of the 74 companies that met eligibility criteria, 27 agreed to participate and 47 declined. Although workforce demographics, shift structure, and previous health-promotion offerings were not significantly different in the two groups, participating companies employed fewer workers and had a more favorable financial outlook than did the companies that declined to participate. These baseline disparities could ultimately affect the ability to extend inferences about the outcomes to settings different from the workplaces that participated.

The setting where activities take place may also have a strong influence on DM program outcomes. For example, in end-stage renal disease, DM program nurses typically meet patients at the dialysis center or at an inpatient or outpatient facility. However, not all facilities are amenable to outside nurses engaging patients while they are under the facilities' care and supervision. Moreover, some facilities may even prohibit DM program staff from meeting with patients during their dialysis treatment.

A similar situation may arise in a DM program that relies on physicians to recruit suitable patients. A physician who is strongly opposed to a particular program or the concept of DM in general may actively discourage patients from enrolling in the program or participating in the intervention. Not only does

this prevent patients from receiving potentially valuable health services, it also reduces the ability to generalize the results of the overall program to all physician offices.

Interaction of Causal Relationship With Outcomes. The effectiveness of DM programs varies according to whether the chosen outcome is related to quality, utilization, or cost. For example, satisfaction with the program may be high among participants, whereas their adherence to self-management activities may be low. More typically, an increase in appropriate preventive health screenings or observance of treatment guidelines does not correspond to a reduction in hospitalizations or ED visits. Although this may seem counterintuitive, it may reflect inadequate program duration or simply indicate that a cause–effect relationship does not exist between the intervention and the outcomes.

Using cost as an outcome measure may pose the biggest threat to both the internal and external validity of a DM program's results. Both within a DM program for one client, or across all clients, costs for the same treatment may differ by provider, setting, and region, rendering comparisons difficult to make. Similarly, defining which costs will be included in the outcome measure will affect interpretation of the results (both within a program for one client as well as across clients). For example, if total health care costs per member are included in the measurement, it is entirely possible that disease-specific costs will increase and nonspecific costs will decrease. How the DM program explains that phenomenon may determine whether or not the client will renew the service contract. Therefore, agreement on outcome parameters before program design is necessary to ensure that likely questions about the ability to generalize results will be addressed up front.

Context-Dependent Mediation. To explain how outcomes were achieved, and to ensure that those outcomes can be replicated and generalized, identifying the mediating factors in the process is essential. For example, most DM programs choose ED visit rates as one of their outcome measures, under the assumption that teaching members how to recognize the onset of acute symptoms may lead them to visit their physician before an acute exacerbation occurs. However, a busy physician may instruct the patient to go directly to the ED, thereby nullifying the effect of the program's intervention. Similarly, patients may choose the ED as their primary source of care if they have difficulty accessing their physician's outpatient services when needed.

An alternate scenario is provided by the ED visit rates example. If patients present to the ED despite instructions to the contrary, upon arrival they are triaged to a more suitable level of care (e.g., urgent care or their PCP). In both this and the previous situation, the rate of ED visits changes, although not as a consequence of the DM program design. These examples provide an excellent illustration of how a threat to the internal validity of the program (did the program affect ED visit rates?) can also pose a threat to the generalization of those results to other settings or clients. In this example, physician and ED practices influenced the rate of ED visits. Therefore, identification of mediating factors will greatly improve the ability to estimate whether program results can be generalized.

IMPROVING GENERALIZATION

The discussion thus far has illustrated how extensive the issues can be that relate to generalization. However, it must be remembered that before one can ensure that a causal relationship holds true over variations in persons, settings, treatments, and outcomes, one must first demonstrate that the results are successful in overcoming the biases that may limit the internal validity of the study findings. Following are guidelines and strategies to improve the ability to generalize DM program results across those domains, with the understanding that internal validity has been previously demonstrated.

Sampling Techniques. Random sampling of subjects from the population is considered by many researchers to be the best way to ensure that results can be generalized. Coupling this with further randomization of those subjects to either an intervention or control group would maximize the ability to generalize by distributing explained and unexplained sources of variability evenly between participants and nonparticipants. The reason for this is similar to the explanation given for its use in demonstrating internal validity. That is, if every subject within the population has equal opportunity to participate (either in the intervention program or as a control), the variability (both observed and unobserved) will be spread equally across both groups. Theoretically, this allows the drawing of inferences about the general population.

Several recent studies have questioned the appropriateness of generalizing results obtained in some randomized, controlled trials because certain biases were identified (most notably, selection bias) that make generalizing to different persons, settings, treatments, and outcomes difficult.[19–22] However, if program administrators and evaluators can minimize bias in the study design, the results can be generalized with a high level of confidence.

Stratification. Stratifying the population into meaningful subsets may help identify areas where program results differ. Age is a variable typically used

in stratification, and most diseases affect some age groups more than others (e.g., asthma in children, coronary artery disease [CAD] in adults over age 35, CHF and COPD in the elderly). Some diseases are associated with ethnicity (e.g., sickle-cell trait in African Americans, Tay–Sachs disease in Jews of eastern European ancestry, diabetes in Native Americans). The setting in which care is provided may also produce varied results; for instance, a patient with influenza may present to a doctor's office, an urgent-care facility, an ED, or a hospital outpatient clinic.

These examples illustrate the need to stratify according to characteristics of importance among both the study group and the target population. Any significant difference in characteristics between the two groups should be studied for relevance to the overall outcome. If different effects of the intervention are discovered, the adjustment techniques discussed below may be employed. In some cases, there may be insufficient representation of certain groups in the study population, in which case there is no choice other than to concede that the outcomes cannot be generalized to the target population.

Indirect Estimation. Indirect estimation is a class of methods for generalizing results from one population to another when the effect is different in subpopulations. Consider CAD, which has a different effect on women than on men. When women present with myocardial infarction, they are more likely than men to be misdiagnosed, and they are also more likely to die of their first infarction.[23] Similarly, women are less likely to undergo angioplasty or bypass surgery,[24] or to receive cardiac rehabilitation[25] or therapy with aspirin, beta blockers, or angiotensin-converting enzyme inhibitors.[26] Generalizing a single effect estimate from a study population that is 50% women to a population that is 40% women is likely to overestimate the treatment effect in the target population. However, adjusting for the population differences in this case is straightforward: Separate male and female effect estimates can be weighted with the target population's gender fractions to obtain a more generalizable effect estimate.

This simple adjustment concept can be expanded with the use of stratified sampling and/or regression methods. A regression equation that estimates the effects of patient characteristics on program effects can be used to predict the program effects in the target population. Once again, it is important that the study population used to develop this regression model either has adequate representation of patients and organizational features to support extrapolation to the target population, or can be stratified to support the estimation. In some cases, it may not be possible to achieve the adjustment. For example, a program that potentially has different effects on different racial groups cannot be generalized from a study population with only a single racial group. Similarly, a program whose effects may vary in different delivery models may not be generalized from a study in a staff-model HMO to that of a group-model HMO population.

The Five Principles of Generalized Causal Inference. A discussion of generalization would not be complete without a brief description of the five guiding principles of generalized causal inferences proposed by Shaddish and colleagues.[9] Most researchers routinely make generalizations without following any specific theory. More often than not, these generalizations are based on face validity and common sense. Therefore, these authors set out to identify the prototypical behaviors and assumptions that guide decisions about generalizing and to capture them within these five principles. The following is a very brief description of each, with an example relevant to DM.

1. "Surface similarity" refers to the general assumption that results for one kind of person, setting, treatment, or outcome can be generalized to elements that appear similar in terms of important characteristics. The most basic example in DM is that on the surface, a nursing intervention is expected to elicit a similar response from all patients with the same disease. The same reduction might be expected in hospitalization rates in different settings if the intervention is believed to be linked to that outcome. Participation by providers may also be assumed to be similar for those reimbursed on a fee schedule and those paid under capitation. Payment is not believed to be the driver of physician participation in a health plan's DM program.

2. "Ruling out irrelevancies" pertains to the identification of variables among persons, treatments, settings, and outcomes that are irrelevant to the size or direction of the cause–effect relationship. For example, gender may be ruled out as a potential influence on patients' participation in the DM program intervention. Similarly, physician panel size may not be believed to be a factor in a given physician's support of the DM program. One may also believe that persons with diabetes living in a rural area will be just as compliant with their medication regimen as those living in a large city. Although these presumptions may be based on rational judgment, they may, in fact, turn out to be confounding factors in the outcome. Therefore, data analysis will ultimately be needed to determine the relevance of these factors in making generalizable inferences.

3. "Making discriminations" is the process of discriminating between variations of a factor that are all

thought to have a similar cause–effect relationship with an outcome when, in fact, one or more of those variations may change the direction and magnitude of the causal relationship. For example, a DM program may presume that a telephonic dialogue between a program participant and a nurse will elicit the same results as a computerized telephonic response system. In this scenario, the telephonic intervention is the factor and the two methods (nurse vs. computer) are the variations. Upon further investigation, it may be found that the "human touch" of the nurse elicits better results than the computerized system.

Another example might be different outcomes in males and females with diabetes who receive the same DM intervention. Even though all participants are provided the same self-management instructions, males may be more apt to begin an exercise program than their female counterparts, whereas females may be more likely than their male counterparts to follow the recommended dietary changes. In any such scenario, the challenge is to discriminate between the variations of a given factor that may change the magnitude and direction of the causal relationship and thereby reduce the ability to generalize.

4. "Interpolation–extrapolation" allows inferences to be drawn concerning both values within the range of observations (interpolation) and values below or above the range of observations (extrapolation). For example, a telephonic nursing intervention that targeted the least-compliant members with CHF achieved an increase in the frequency of weight self-monitoring. The intervention was then expanded to all members with CHF, with the expected increase in self-monitoring calculated from the results of the low-compliance group. Thus, to predict results in the low-risk group, the program administrators extrapolated the results from the high-risk group.

Similarly, DM programs typically base their estimates of meeting financial targets on the size of the health plan's general and disease-specific populations. These estimates are either extrapolated or interpolated from their experiences with other clients. As discussed earlier, the effectiveness of many program interventions is a function of the duration of a participant's exposure to that intervention. Estimates of the long-term effects of the program intervention may extend beyond the interval of the program's experience or follow-up. Similarly, experience may inform the estimate of the program's effects on members at any point within the range of the period studied. To improve the accuracy of interpolation, a program should be evaluated for its effect on a broad range of observations of the given factor. To extrapolate with some measure of accuracy, it is recommended that calculated values be relatively close to the upper and lower boundaries of only the actual range of observations. This is because the further one proceeds from the last known observation, the more difficult it will be to estimate the outcome of the next observation.

5. "Causal explanation" might be considered the holy grail of scientific research. In DM, this principle ties all factors together by explaining how the intervention affects participants and achieves the desired result. Once the underlying phenomena are understood, including the mediating factors that may alter or enhance the intervention effect, one can be fully confident that reproducing the intervention in different settings will achieve precisely the same results. However, few studies, if any, have demonstrated a true cause–effect relationship between DM program interventions and outcomes achieved. This is not surprising, since most causal relationships in the social sciences have not been definitively proven, owing to the inability to identify and control for the innumerable sources of variation at every level of every factor. Nonetheless, it is imperative that DM programs invest resources to analyze the causal relationships that exist within the program dimension.

It is imperative that disease management programs invest resources to analyze the causal relationships that exist within the program dimension.

Many and varied techniques can be employed to improve the capability to generalize DM program results. Several methodologic and statistical approaches exist, as well as some processes that are a function of common sense and face validity. The following section will illustrate when and how to best use these techniques during the course of a DM program.

GETTING FROM HERE TO THERE

Pursuant to the difficulties and possible remedies discussed thus far, what remains is practical advice as to how to most effectively implement these concepts throughout the course of a DM program. The most logical place to start is during the initial discussions about a DM program with a prospective client. More often than not, a client will choose a DM vendor on the basis of outcomes achieved for its other clients. The prospective client will expect to realize outcomes similar to the program's best-case scenario. It is at this initial point that the potential to generalize previous results to the prospective population should first be explored. Cross-sectional, stratified comparisons between the DM program's other clients and the current prospect will indicate whether the populations are similar enough to generalize outcomes to this target population. When

statistical comparisons do not provide enough clarity, reliance on the related principles of generalized causal inference may be necessary. One may have to ask if there is enough surface similarity in compelling factors between the two DM populations to permit causal inferences. If differences between the populations are deemed irrelevant to the success of the program, they should be ruled out as limiting factors. Additionally, estimating program success may require extrapolation of past observations to the new target population. Although it remains a business decision, rather than a research-based proposal, if a fundamental difference that limits the ability to generalize is found between the DM program's experience base and that of the new prospect, it should be discussed at this point. Unreasonable expectations should not remain unchallenged.

If physicians do not support a disease management intervention that includes an unproved monitoring device, the device may have to be excluded, which may limit the capability to generalize the program or the results.

The second point at which attention must be directed to potential threats to generalization of outcomes is during implementation. Typically, at this juncture, a DM program's potential for success in this new environment becomes clearer. Patients' and providers' interest in participating in the program may be higher or lower than expected. Changes made in program implementation that differ significantly from the approaches used in other settings may affect the overall ability to generalize. For example, some programs institute an "opt-out" enrollment process in areas where it is difficult to get members to enroll voluntarily. With this method, members are automatically enrolled in the program and must actively request not to participate. As a result, the overall membership demographics may be vastly different from those in settings where the traditional "opt-in" method is used (i.e., members are contacted by the DM program and asked to enroll). The same methods and tools used to assess the generalization potential before program initiation should be used here as well. Cross-sectional, stratified measurements should be reviewed regularly to identify unusual patterns in enrollment or disenrollment. Any anomalies should be investigated to determine if they may pose a threat to external validity.

From the initial enrollment phase until completion of the first evaluation period, situations may arise that will require the DM program to be modified in some ways. These may include convincing recalcitrant members to either enroll in the program or adhere to behavior-modification strategies. Another factor may be whether physicians consider the intervention strategy to be sound. For example, if physicians do not support a DM intervention that includes an unproved monitoring device, the device may have to be excluded, which may limit the capability to generalize the program or the results. Similarly, relationships between the health plan (or other payer of DM services) and providers may change as a result of reimbursement issues, medical management policies, or market forces.

Such circumstances may compromise the DM program's success in certain environments. A time-series analysis may be the best method for ensuring that the program is on track at this point.[2] Even a visual inspection of the plotted monthly observations will indicate whether the factor under study is tracking with the forecast, as well as with the comparison population. Deviations from the expected course may be linked to any type of changes in the dynamics of the program.

During the final phase of the program evaluation, all factors that may have influenced the fundamentals of the program must be taken into account. Ideally, on commencement of the evaluation, all observed threats to validity will have been identified. Statistical tools, some of which were described in the preceding section, should be used to adjust for factors that may limit the ability to infer the extent to which a causal relationship persists over individual, environmental, therapeutic, and outcome variations. For instance, a cross-sectional, stratified population analysis will indicate if the characteristics of the DM program participants resemble those of the comparison population. If so, it is possible that the final results can be generalized. Many statistical tools are available to enhance the potential for generalization, including propensity scoring,[4] survival analysis,[6] time-series analysis,[2] and regression models.

CONCLUSION

The ability to generalize results from one DM intervention group to other populations or other disease states may be as important as demonstrating the program's internal validity. The concepts and strategies for assessing external validity are, more often than not, based on a common-sense approach that can include statistical analyses as well as straightforward, educated assumptions based on observation. The key message, however, is that DM programs must begin the process of evaluating external validity even before the contract is signed and implementation begins. Any fundamental differences found between the DM program's experience across people, settings, treatments,

or outcomes and that of the prospective client may limit the ability to generalize. Any anomalies should be discussed at this point, since unreasonable expectations created at the outset cannot be reversed later, during the program evaluation.

REFERENCES

1. Linden A, Adams J, Roberts N: An assessment of the total population approach for evaluating disease management program effectiveness. *Dis Manag* 2003;6:93-102.
2. Linden A, Adams J, Roberts N: Evaluating disease management program effectiveness: An introduction to time series analysis. *Dis Manag* 2003;6:243-255.
3. Linden A, Adams J, Roberts N: Evaluation methods in disease management: Determining program effectiveness. Disease Management Association of America, October 2003 (www.dmaa.org/PDFs/Evaluation_Methods_In_DM.pdf).
4. Linden A, Adams J, Roberts N: Using propensity scores to construct comparable control groups for disease management program evaluation. *Disease Management and Health Outcomes* (under review).
5. Linden A, Roberts N: Disease management interventions: What's in the black box? *Dis Manag* (in press).
6. Linden A, Adams J, Roberts N: Evaluating disease management program effectiveness: An introduction to survival analysis. *Dis Manag* (in press).
7. 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.
8. American Healthways and the Johns Hopkins Consensus Conference: Consensus report: Standard outcome metrics and evaluation methodology for disease management programs. *Dis Manag* 2003;6:121-138.
9. Shaddish SR, Cook TD, Campbell DT: *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston, Houghton Mifflin, 2002.
10. Fossa SD, Skovlund E: Selection of patients may limit the generalizability of results from cancer trials. *Acta Oncol* 2002;41: 131-137.
11. Dent CW, Sussman S, Stacy AW: Project towards no drug abuse: Generalizability to a general high school sample. *Prev Med* 2001;32: 514-520.
12. Wasti SA, Bergman ME, Glomb TM, et al: Test of the cross-cultural generalizability of a model of sexual harassment. *J Appl Psychol* 2000;85:766-778.
13. Aday L, Andersen RM: Equity in access to medical care: Realized and potential. *Med Care* 1981;19(suppl):4-27.
14. Andersen RM: *Behavioral Model of Families: Use of Health Services.* Research Series No. 25. Chicago, Center for Health Administration Studies, University of Chicago, 1968.
15. Turano KA, Massof RW, Quigley HA: A self-assessment instrument designed for measuring independent mobility in RP patients: Generalizability to glaucoma patients. *Invest Ophthalmol Vis Sci* 2002;43:2874-2881.
16. Prochaska JO, DiClemente CC, Norcross JC: In search of how people change: Applications to addictive behaviors. *Am Psychol* 1992;47:1102-1114.
17. Mehta SD, Rompalo A, Rothman RE, et al: Generalizability of STD screening in urban emergency departments: Comparison of results from inner city and urban sites in Baltimore, Maryland. *Sex Transm Dis* 2003;30:143-148.
18. Biener L, DePue JD, Emmons KM, et al: Recruitment of work sites to a health promotion research trial: Implications for generalizability. *J Occup Med* 1994;36:631-636.
19. Peppercorn JM, Weeks JC, Cook EF, et al: Comparison of outcomes in cancer patients treated within and outside clinical trials: Conceptual framework and structured review. *Lancet* 2004;363:263-270.
20. Moore DA, Goodall RL, Ives NJ, et al: How generalizable are the results from large, randomized, controlled trials of antiretroviral therapy? *HIV Med* 2002;1:149-154.
21. Lloyd-Jones DM, O'Donnell CJ, D'Agostino RB, et al: Applicability of cholesterol-lowering primary prevention trials to a general population: The Framingham heart study. *Arch Intern Med* 2001;161:949-954.
22. Margee LA, Bull SB, Koren G, et al: The generalizability of trial data: A comparison of beta-blocker trial participants with a prospective cohort of women taking beta-blockers in pregnancy. *Eur J Obstet Gynecol Reprod Biol* 2001;94:205-210.
23. Kudenchuk PJ, Maynard C, Martin JS, et al: Comparison of presentation, treatment, and outcome of acute myocardial infarction in men versus women (the Myocardial Infarction Triage and Intervention Registry). *Am J Cardiol* 1996;78:9-14.
24. Bearden D, Allman R, McDonald R, et al, for the Systolic Hypertension in the Elderly Program (SHEP) Cooperative Research Group: Age, race, and gender variation in the utilization of coronary artery bypass surgery and angioplasty in SHEP. *J Am Geriatr Soc* 1994;42:1143-1149.
25. Charney P (ed): *Coronary Artery Disease in Women: What All Physicians Need to Know.* Philadelphia, American College of Physicians, 1999.
26. Herholz H, Goff DC, Ramsey DJ, et al: Women and Mexican Americans receive fewer cardiovascular drugs following myocardial infarction than men and non-Hispanic whites: The Corpus Christi Heart Project, 1988–1990. *J Clin Epidemiol* 1996; 49:279-287.

Address for correspondence: Ariel Linden, DrPH, MS, President, Linden Consulting Group, 6208 NE Chestnut Street, Hillsboro, Oregon 97124. E-mail: ariellinden@yahoo.com.