A User's Guide to the Disease Management Literature: Recommendations for Reporting and Assessing Program Outcomes

Ariel Linden, DrPH; and Nancy Roberts, MPH

Recently there has been tremendous growth in the number of lay-press articles and peer-reviewed journal articles reporting extraordinary improvements in health status and financial outcomes due to disease management (DM) interventions. However, closer scrutiny of these reports reveals serious flaws in research design and/or analysis, leaving many to question the veracity of the claims. In recent years, there have been numerous contributions to the literature on how to assess the quality of medical research papers. However, these guidelines focus primarily on randomized controlled trials, with little attention given to the observational study designs typically used in DM outcome studies. As such, general guides to evaluating the medical literature are inadequate in their utility to assist authors and readers of DM outcomes research. The purpose of this paper is to provide authors with a clear and comprehensive guide to the reporting of DM outcomes, as well as to educate readers of the DM literature (both lay and peer reviewed) in how to assess the quality of the findings presented.

(Am J Manag Care. 2005;11:113-120)

Until recently, disease management (DM) has largely been able to avoid scrutiny of its methods for assessing effectiveness in attaining positive health and financial outcomes. Unfortunately, this has led to the reporting of incredible achievements in the lay and industry press that have left many questioning the veracity of these claims.[1-5]

Similarly, DM outcome studies in peer-reviewed literature have reported extraordinary results, at times as a consequence of poor study design. Some basic issues include the use of a pre-post study without a control group, or the misguided application of a more robust design; not addressing biases that may threaten the validity of the results; inadequate description of research methods or characteristics of the population; and inappropriate use or lack of statistical analysis.[6-14] Attempts have been made recently to address these shortcomings publicly.[15-18] However, a more methodical approach to both designing and reviewing DM evaluation and research studies is needed.

In recent years there have been numerous contributions to the literature on how to assess the quality of medical research papers, with perhaps the most notable among them being a series of articles appearing in the *Journal of the American Medical Association*.[19-50] However, these guidelines focus primarily on randomized controlled trials (RCTs), with little attention given to the observational study designs typically used in DM outcome studies. As such, general guides to evaluating the medical literature are inadequate in their utility to assist authors and readers of DM outcomes research. Therefore, the purpose of this paper is to provide authors with a clear and comprehensive guide to the reporting of DM outcomes, as well as to educate readers of the DM literature (both lay and peer reviewed) in how to assess the quality of the findings presented.

ASSESSING METHODOLOGICAL QUALITY

Fundamentally, the objective of any thoughtful critique is to ascertain whether the reported results represent an unbiased estimate of a treatment effect, or whether they were influenced by factors other than the intervention. To make this determination, one must consider 2 major elements in any evaluation or research study; the *study design* and the *analysis* performed on the data. The **Figure** presents a framework for assessing the quality of the study design and analysis used in RCTs and observational studies. As shown, many methodological issues overlap, while others are specific to the given design category. The items are ordered temporally to coincide with each phase of the study.

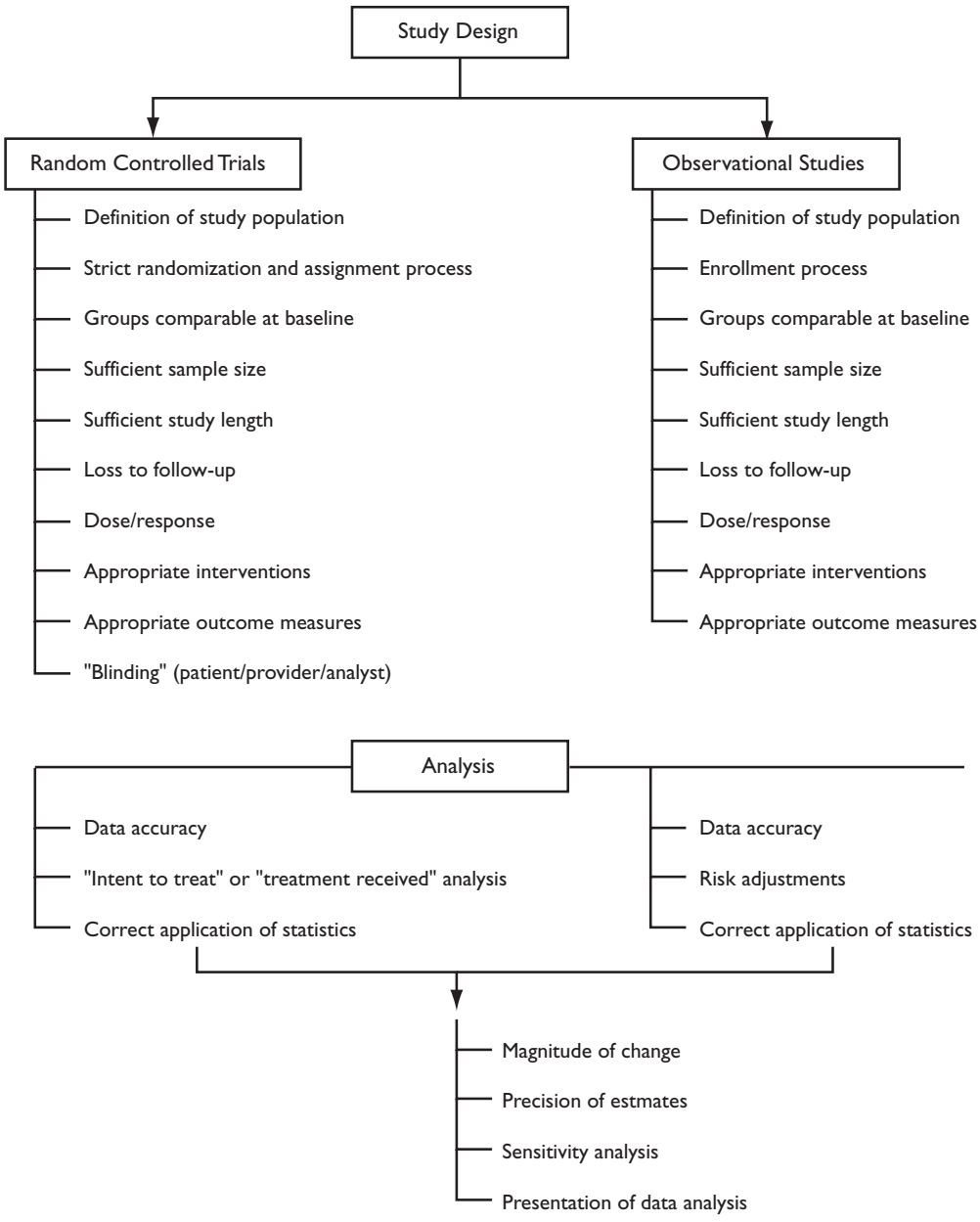
Study Design

The 2 predominant categories of study design relevant to DM research are *experimental* (better known as the RCT) and *quasi-experimental* (generally referred to

From Linden Consulting Group, Portland, Ore, and Oregon Health Science University, School of Medicine, Department of Preventive Health/Preventive Medicine, Portland, Ore (AL); and the Providence Health System, Portland, Ore (NP).

Address correspondence to: Ariel Linden, DrPH, President, Linden Consulting Group, 6208 NE Chestnut St, Hillsboro, OR 97124. E-mail: ariellinden@yahoo.com.

Figure. Framework for Assessing the Quality of the Study Design and Analysis of Randomized Controlled Trials and Observational Studies



as an observational study design).[51-53] The most basic difference between these 2 categories lies in how subjects are assigned to the study. As the name implies, in the RCT, individuals are randomly assigned either to a treatment or a control group, thereby giving each person an equal probability to be chosen for the intervention. Conversely, in an observational study design, eligible individuals are not randomly assigned to the treatment or the control group. In DM, patients and/or their physicians are commonly allowed to decide who will participate in program interventions. This type of assignment process results in a nonrandom distribution of individuals to the intervention and nonintervention groups.

The value of a randomized assignment process is that all variability is distributed equally between the two groups.[54] Variability comes in 2 forms: observed and unobserved. Observed covariates are characteristics that can be measured by the analyst via sources such as claims, medical records, member files, or survey reports; and unobserved covariates are all other characteristics not captured or recorded. Although observed covariates are used for ensuring that subjects in the 2 groups are similar on baseline characteristics (eg, age, sex, disease status), it is left to the process of randomization to ensure that unobserved characteristics are similar in both groups as well. Observational study designs are susceptible to bias precisely because they cannot control for unobserved covariates, and therefore cannot provide unbiased estimates of treatment effect.

The launching point for any study regardless of design category (RCT or observational) is a definition of the

study population. It is important that the individuals eligible for inclusion in the study be representative of the population to which the findings will be applied. For example, many studies exclude women, the elderly, or individuals with multiple illnesses.[55] This obviously limits the ability to generalize results across patients outside of this study population.[56] In DM, program participants typically are not representative of the general population with the disease. By design, program administrators target those patients who are either the sickest or at the highest risk of utilizing services. Therefore, it is important for the researcher and the reader of the DM outcomes literature to recognize the limitations of generalizability of the study findings. A good definition of the study population would include a description of the inclusion/exclusion criteria, and clinical/demographic characteristics of both the treatment and control groups.

The second attribute of study design to consider is the process by which individuals engage in either the treatment or nontreatment group. Strict adherence to the assignment process is absolutely crucial in an RCT. As stated earlier, the basic tenet behind randomization is that it distributes unobserved variation evenly between groups. Imagine if the assignment process allowed a patient's physician to determine study participation. Bias would be introduced if that physician relied on personal judgment to determine whether the patient should or should not be included. Studies in which the process of random assignment was inadequately described or not described at all have been shown to exaggerate the size of the observed treatment effect.[57,58] In observational studies (DM studies in particular), assignment is usually determined through self-selection. Individuals eligible for the study or program intervention are invited to participate. The factors determining why a given individual chooses to participate while another individual does not are at the crux of the issue that differentiates RCTs from observational studies. It has been well demonstrated[59,60] that myriad factors (eg, belief systems, enabling factors, perceived need) help explain why and how individuals access healthcare and perform health-related behaviors. Collecting as much information as possible on those eligible individuals who choose to participate as well as on those who decline participation may assist the researcher in identifying those differential characteristics. Similarly, unusual features of one group or another must be described for readers.

Assessing comparability between the study group and the control group on baseline characteristics is the next element of study design to consider. Baseline comparability of groups is an essential step in determining a causal link between study or program intervention and outcome.[54] Most DM programs are currently being evaluated using a pre-post design with no control group. The most basic limitation of this design is that without a control group for which comparisons of outcomes can be made, several sources of bias and/or competing extraneous confounding factors offer plausible alternative explanations for any change from baseline.[61] Advocates of this approach argue that most threats to validity are nullified by using the entire population in the analysis.[62] However, unless some basic factors are controlled for, such as case mix and turnover rate, bias still remains a significant concern. Even with these controlling variables in place, the pre-post method can be confounded with environmental changes unrelated to the DM program interventions.

Given these concerns, it is absolutely necessary to develop a control group with which comparisons can be made. Considering that DM programs or their payers are not likely to withhold potentially beneficial interventions from eligible individuals by assigning them to the control group, statistical methods can be used to match participants to historic controls.[63] That said, some studies have included control groups in their evaluations.[14] In both RCTs and observational studies, comparability can only be assessed on observed characteristics. Therefore, it is extremely important that the research study include either a table or detailed description of demographic and clinical attributes of the treatment and control groups. If cohorts differ on important observed baseline features, causal inferences about the program impact will be limited.

Determining whether an adequate number of individuals were included in the study is the next design feature to review. Four interrelated parameters have an effect on the conclusions that are attained from a typical statistical test[64]:

- *Sample size,* or the number of observations, subjects, or cases under study.
- *Significance level,* or alpha. This is the probability that the observed result is due to chance alone.
- *Power,* or the probability that a difference will be observed when it actually occurs.
- *Effect size.* This is the magnitude of change between 2 groups or within 1 group, before and after the intervention.[64]

Using this logic, the sample size of the study must be sufficiently large to reduce the effect size necessary to demonstrate statistically significant findings. To put it simply, studies that use large samples require a smaller effect size to show statistical significance. Thus, from an evaluation perspective, DM programs should strive to

enroll as many participants as possible and identify an equal or larger number of controls.[64] Similarly, analyses on subgroups can be carried out only if their sample size is sufficiently large, irrespective of the overall study population size. Study reports should identify the sample size for treatment and control groups as well as for any subgroups analyzed. Significance levels for all findings should be clearly reported.

Study duration is the next important factor in DM program evaluations and is interrelated with 2 other elements that impact the validity of the findings: dose/response and loss to follow-up (attrition). It is generally agreed that it takes at least 6 months after DM program commencement until behavioral changes begin to take effect (dose/response). Therefore, significant changes in healthcare utilization or monetary outcomes may not be realized within the first year. Studies reporting immense decreases in utilization and costs in a short-duration study (less than 1 year) must be viewed with suspicion (especially if the study does not include a control group, or if the cohorts are not comparable at baseline). The most likely bias in this scenario is regression to the mean.[61]

Attrition from a DM program via disenrollment inarguably impacts results negatively. Participants who do not achieve the maximum benefit from the intervention (eg, improved self-management of their disease, improved knowledge of how to access appropriate health services)[65] may continue to exhibit behaviors that run contrary to the program objective. Therefore, it is imperative that studies include a description of the population that did not complete the prescribed length/amount of treatment. Two methods that can be used to adjust for attrition are survival analysis[66] and time-series regression.[67]

The next important, yet often overlooked, aspect of DM program evaluations is the intervention itself. It is mostly assumed that the treatment is robust, and that any change noted in the outcomes are causally linked to that treatment. However, rarely is the intervention described in enough detail to allow readers to decide for themselves if there is sufficient evidence to draw this conclusion.[65] Moreover, specific outcome measures directly related to that intervention should be included. For example, if psychosocial models are used to change health-related behaviors, then analyses should be performed, and reported, to assess the relative change in those behaviors. Without such information, the reader is left to question the causal impact of those interventions.

A treatment effect may or may not be evidenced, depending on the choice of outcomes. Most often in DM program evaluations medical cost is chosen as the primary end point. However, cost is an ill-advised outcome variable because it is influenced by changes in the unit cost of services, members' financial share of the medical expense, introduction of new technologies, and so forth—variables outside of the DM program intervention.[61] It is for this reason that disease-specific utilization measures should be used as indicators of program success.[68,69] Although rising costs may be due to many uncontrolled-for variables, a decrease in utilization is more likely 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 final study design element for consideration in RCTs only is blinding the patient, provider, and analyst to group assignment. Blinding alone eliminates the introduction of several biases that may invalidate the study findings. Failure to use blinding in RCTs has been shown to overstate treatment effects.[58] This issue is not relevant to observational studies since researchers have little control over program participation.

Analysis

The second major area related to the quality of evaluation or research findings is the rigor and applicability of the data analysis. Accuracy of data sources is the first point of concern. Most DM programs rely on large administrative databases (medical claims and membership files) for retrieving information on diagnostic measures to identify suitable participants, baseline characteristics, quality indicators, and utilization and cost values. These data sources are notoriously inaccurate.[70] The influence of data inaccuracy on outcomes can be decisive. For example, in one study comparing the ability to predict mortality after coronary artery bypass surgery, the predictive ability based on data derived from medical records was significantly better than that based on administrative data.[71] Therefore, a description of how validation of data accuracy was accomplished must be presented in studies that rely on administrative data for any aspect of the research endeavor.

Next, the group on which the analysis was performed should be clearly identified. In RCTs, it is common to assess outcomes of all participants assigned to a given cohort, as opposed to evaluating outcomes only of those who received the treatment. The former is called the intent-to-treat (ITT) analysis, and the latter is referred to as a treatment-received (TR) analysis. The ITT analysis preserves the value of randomization (by equally distributing observed and unobserved covariates between the cohorts); however, causal inferences can be made

only about the effects of being assigned to a given treatment, not receipt of that treatment. This method is useful on a policy level, where forecasts of outcomes can be made assuming the program will be implemented on a large-scale basis.[53] In DM programs, individuals self-select to participate in the program and thereby limit the analysis to the TR method. Predictive risk-adjusted models should be used to improve the process by which suitable participants are identified, while establishing a means to provide a more accurate description of eligible individuals. If the tool has high sensitivity (accurately identifying people who meet the eligibility criteria) and specificity (accurately identifying people who do not meet the eligibility criteria), the researcher may feel more confident in using the ITT method for comparing outcomes between cohorts.

The final step in reviewing the soundness of a study's data analysis is consideration of the application of statistics. A comprehensive discussion about the types of statistical models and analyses required to evaluate program effectiveness is unfortunately beyond the scope of this paper. However, 2 books, *An Introduction to Medical Statistics* and *Medical Statistics: A Commonsense Approach,* provide a good introduction to medical statistics for the interested reader with a basic understanding of the field.[72,73] In DM outcome studies, multiple regression analysis almost always is required to estimate the independent effect of covariates on the outcome and to test whether the model provides additional prognostic value. These models also form the basis for most risk-adjustment tools. Two important variables that should be included in any comparative analysis are severity and case-mix adjustment.[74] Especially in pre-post designs, tracking the population's case mix and severity level over the course of the study will assist in determining whether the program had a treatment effect or whether population dynamics influenced outcomes. Several diagnostic groupers can be readily used for this purpose (eg, diagnosis-related groups, ambulatory care groups), as well more simple methods such as counts of comorbid conditions. These variables should be included in the regression model as adjusters in the assessment of a treatment effect.

Actual *P* values and/or 95% confidence intervals should be stated for each outcome variable. While this statement may appear superfluous, many studies either do not include any levels of significance, or they provide inexact measurements. For example, while the general consensus is to report significant *P* values at <.05, several studies report values at *P* <.10. This is potentially misleading to the inattentive reader, who may draw the wrong conclusions based on these values. Similarly, many journals require that researchers report exact *P* values instead of "NS" (nonsignificant). This allows readers to decide for themselves how much stock to put in that actual value, as opposed to a predetermination by the authors on their behalf.

Confidence intervals give an estimated range of values within which the unknown population parameter may lie. Using the mean as an example, we can calculate, based on the sample data, an estimated range of values within which we believe (with a given level of confidence) that the population mean may exist. The width of the confidence interval generally gives us some insight as to the accuracy of the estimate. A wide interval may indicate large variability in the dataset, or may be a result of having a very small number of study participants. In cases where parametric statistics cannot provide confidence intervals, bootstrapping is a viable and suggested option.[75] When the outcome variable is dichotomous (eg, yes/no, 0/1), the proportion of individuals with the outcome should be provided, along with the associated odds ratios. An additional measure that can be used to assess the effect of introducing the intervention is called the "number needed to treat" (NNT)[76] The NNT provides an estimate of the number of patients that must be treated to prevent 1 adverse outcome. While not widely used, this may be a very suitable measure for assessing DM program effectiveness.

Conceptually, the basic premise of a sensitivity analysis is that subjects in observational studies differ from those in RCTs in their recruitment to the treatment group. Although all individuals in a RCT have a 50/50 chance of being assigned to the treatment group, observational studies are limited by self-selection bias. Sensitivity analysis therefore provides an estimate for how far this bias must diverge from the 50/50 split of an RCT to raise concerns about the validity of the study findings (A. Linden, J. Adams, N. Roberts, unpublished data, 2004). Observational studies that fail to include a sensitivity analysis inhibit the reader's ability to judge the strength of the evidence that support a treatment effect.

The presentation of data analyses performed is essential to any research, whether it be an RCT or observational study. Two basic tables should be commonplace in any paper. These are (1) a display of baseline characteristics of the groups under comparison and (2) outputs from statistical analyses, including model parameters and estimates.

Table 1 presents a modified table from an article by Linden et al[63] in which participants in a congestive-heart-failure program were compared with the entire unmanaged congestive-heart-failure population and with a control group matched on propensity score. Included in the table are the major elements discussed in this

Table 1. Pre– and Post–First-Year Program Characteristics of a CHF Intervention Group Compared With the Unmanaged CHF Population and a Matched Control Group*

Variable	Mean ± SE Intervention Group (n = 94)	Mean ± SE CHF Population (n = 4606)	*P* (*T* = *t*)†	Mean ± SE Matched Controls (n = 94)	*P* (*T* = *t*)‡
Age, y	77.4 ± 0.96	76.6 ± 0.19	.539	78.2 ± 0.98	.556
Female, %	0.51 ± 0.05	0.56 ± 0.01	.336	0.51 ± 0.05	1.000
Percent residing in Portland	0.17 ± 0.04	0.69 ± 0.01	<.0001	0.17 ± 0.04	1.000
Risk score (low, medium, high)	0.54 ± 0.05	0.40 ± 0.007	<.0001	0.60 ± 0.05	.379
Pre: CHF admission rate	1.13 ± 0.15	0.50 ± 0.02	<.0001	1.09 ± 0.15	.841
Pre: CHF ED visit rate	0.70 ± 0.11	0.40 ± 0.01	.003	0.67 ± 0.10	.832
Pre: program costs	$18 287 ± $2053	$8974 ± 257	<.0001	$17 001 ± $2449	.688
Post: admission rate	0.59 ± 0.10	0.87 ± 0.02	.0008	1.17 ± 0.18	.005
Post: ED visit rate	0.57 ± 0.08	0.58 ± 0.02	.1874	0.77 ± 0.10	.048
Post: program costs	$11 874 ± $1408	$16 036 ± $370	.005	$24 085 ± $3843	.003

*ED indicates emergency department; CHF, congestive heart failure. Adapted from reference 63.
†Statistical comparison between intervention group and unmanaged CHF population. *P* values were derived from 2-tailed *t* tests for independent samples.
‡Statistical comparison between intervention group and matched controls. *P* values were derived from 2-tailed *t* tests for independent samples.

paper. Baseline characteristics are presented above the dotted line, and outcome measures are shown below it. Sample sizes are noted, as well as group means and standard errors. *P* values are noted for each pairwise comparison. Although this table is meant for illustrative purposes only, it serves as a basic template for presenting comparison group characteristics in a clear and concise manner.

Table 2 presents results from a Cox-regression survival analysis by Linden et al[66] in which age and sex appear to be significant predictors of hospitalization. Each unit increase in a patient's age was expected to increase the risk of hospitalization by 2.6%, while being female reduced the risk of hospitalization by nearly 8%. Also presented are *P* values and 95% confidence intervals. Regardless of statistical model used in the data analysis, tables with a similar structure should be presented to the reader.

CONCLUSION

This paper has provided in some detail a comprehensive guide to the reporting of DM outcomes, including important elements of both study design and data analysis. The information presented herein should be used as an educational tool to enable readers of the DM literature to independently assess the quality of the research findings presented in the lay press and the peer-reviewed literature. This guide also should be used by DM researchers in developing DM evaluation plans and reporting findings. Raising the standards by which DM program outcomes are evaluated should result in improved quality of peer-reviewed and lay publications on the subject, and the healthcare community's confidence in the veracity of these reports.

Table 2. Presentation of Results of a Cox-Regression Analysis*

Variable	χ^2 Model	*df*	*P* Value Model	Regression Coefficient (β)	SE (β)	*P* Value Variables	Exp (β)	95% CI for Exp (β)
Age	118.72	2	<.0001	0.25	0.002	<.0001	1.026	1.021, 1.030
Sex				–0.088	0.039	<.023	0.916	0.849, 0.988

*CI indicates confidence interval. Data are reprinted with permission from reference 66.

REFERENCES

1. Nurse-driven CHF program cuts hospitalization by 87%. *Healthc Demand Dis Manage.* 1997;5:78-80.

2. Humana CHF program cuts costs, admissions. *Healthc Benchmarks.* 1998;5:173-175.

3. Web-based educational effort for CHF patients boost outcomes while cutting costs. *Dis Manage Advisor.* 2001;7:92-96.

4. Costs drop 37% in Colorado Medicaid asthma DM pilot. *Dis Manage News.* 2004;9(8):1, 4. 5.

5. Asthma program targets patient and physician compliance; wins first DM excellence award. *Healthc Demand Dis Manag.* 1998;4(12 suppl 1-4):181-184.

6. Rothman R, Malone R, Bryant B, Horlen C, Pignone M. Pharmacist-led, primary care-based disease management improves hemoglobin A_{1c} in high-risk patients with diabetes. *Am J Med Qual.* March-April 2003;18:51-58.

7. Ibrahim IA, Beich J, Sidorov J, Gabbay R, Yu L. Measuring outcomes of type 2 diabetes disease management program in an HMO setting. *South Med J.* 2002;95:78-87.

8. Erdman DM, Cook CB, Greenlund KJ, et al. The impact of outpatient diabetes management on serum lipids in urban African-Americans with type 2 diabetes. *Diabetes Care.* 2002;25:9-15.

9. Whellan DJ, Gaulden L, Gattis WA, et al. The benefit of implementing a heart failure disease management program. *Arch Intern Med.* 2001;161:2223-2228.

10. Hershberger RE, Ni H, Nauman DJ, et al. Prospective evaluation of an outpatient heart failure management program. *J Card Fail.* March 2001;7:64-74.

11. Baillargeon JP, Lepage S, Larrivee L, Roy MA, Landry S, Maheux P. Intensive surveillance and treatment of dyslipidemia in the postinfarct patient: evaluation of a nurse-oriented management approach. *Can J Cardiol.* 2001;17:169-175.

12. Sidorov J, Gabbay R, Harris R, et al. Disease management for diabetes mellitus: impact on hemoglobin A_{1c}. *Am J Manag Care.* 2000;6:1217-1226.

13. Jowers JR, Schwartz AL, Tinkelman DG, et al. Disease management program improves asthma outcomes. *Am J Manag Care.* 2000;6:585-592.

14. Villagra A, Ahmed T. Effectiveness of a disease management program for patients with diabetes. *Health Aff.* 2004;23:255-266.

15. Johansen LW, Bjorndal A, Flottorp S, Grotting T, Oxman AD. Evaluation of health information in newspapers and brochures. What can we believe? *Tidsskr Nor Laegeforen.* January 20, 1996;116(2):260-264.

16. Cardium claims big savings, DM observers have doubts. *Dis Manage News.* 2002;8(3):1, 5, 6.

17. Critics question CO Medicaid's asthma claims. *Dis Manage News.* 2004;9(9):2, 3, 6.

18. Wilson T, Linden A. Measuring diabetes management [letter]. *Health Aff.* 2004;23:7, 8.

19. Oxman A, Sackett DL, Guyatt GH. Users' guides to the medical literature, I: how to get started. *JAMA.* 1993;270:2093-2095.

20. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature, II: how to use an article about therapy or prevention. A. Are the results of the study valid? *JAMA.* 1993;270:2598-2601.

21. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature, II: how to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? *JAMA.* 1994;271:59-63.

22. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature, III: how to use an article about a diagnostic test. A. Are the results of the study valid? *JAMA.* 1994;271:389-391.

23. Jaeschke R, Gordon H, Guyatt G, Sackett DL. Users' guides to the medical literature, III: how to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? *JAMA.* 1994;271:703-707.

24. Levine M, Walter S, Lee H, Haines T, Holbrook A, Moyer V. Users' guides to the medical literature, IV: how to use an article about harm. *JAMA.* 1994;271:1615-1619.

25. Laupacis A, Wells G, Richardson S, Tugwell P. Users' guides to the medical literature, V: how to use an article about prognosis. *JAMA.* 1994;272:234-237.

26. Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature, VI: how to use an overview. Evidence-Based Medicine Working Group. *JAMA.* 1994;272:1367-1371.

27. Richardson WS, Detsky AS. Users' guides to the medical literature, VII: how to use a clinical decision analysis. A. Are the results of the study valid? *JAMA.* 1995;273:1292-1295.

28. Richardson WS, Detsky AS. Users' guides to the medical literature, VII: how to use a clinical decision analysis. B. What are the results and will they help me in caring for my patients? *JAMA.* 1995;273:1610-1613.

29. Hayward RSA, Wilson MC, Tunis SR, Bass EB, Guyatt G. Users' guides to the medical literature, VIII: how to use clinical practice guidelines. A. Are the recommendations valid? *JAMA.* 1995;274:570-574.

30. Wilson MC, Hayward RSA, Tunis SR, Bass EB, Guyatt G. Users' guides to the medical literature, VIII: how to use clinical practice guidelines. B. What are the recommendations and will they help you in caring for your patients? *JAMA.* 1995;274:1630-1632.

31. Guyatt GH, Sackett DL, Sinclair JC, et al. Users' guides to the medical literature, IX: a method for grading health care recommendations. *JAMA.* 1995;274:1800-1804.

32. Naylor CD, Guyatt GH. Users' guides to the medical literature, X: how to use an article reporting variations in the outcomes of health services. Evidence-Based Medicine Working Group. *JAMA.* 1996;275:554-558.

33. Naylor CD, Guyatt GH. Users' guides to the medical literature, XI: how to use an article about a clinical utilization review. Evidence-Based Medicine Working Group. *JAMA.* 1996;275:1435-1439.

34. Guyatt GH, Naylor CD, Juniper E, et al. Users' guides to the medical literature, XII: how to use articles about health-related quality of life. Evidence-Based Medicine Working Group. *JAMA.* 1997;277:1232-1237.

35. Drummond MF, Richardson WS, O'Brien BJ, Levine M, Heyland D. Users' guides to the medical literature, XIII: how to use an article on economic analysis of clinical practice. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA.* 1997;277:1552-1557.

36. O'Brien BJ, Heyland D, Richardson WS, Levine M, Drummond MF. Users' guides to the medical literature, XIII: how to use an article on economic analysis of clinical practice. B. What are the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group [published erratum appears in *JAMA.* 1997;278:1064]. *JAMA.* 1997;277:1802-1806.

37. Dans AL, Dans LF, Guyatt GH, Richardson S. Users' guides to the medical literature, XIV: how to decide on the applicability of clinical trial results to your patient. Evidence-Based Medicine Working Group. *JAMA.* 1998;279:545-549.

38. Richardson WS, Wilson MC, Guyatt GH, Cook DJ, Nishikawa J. Users' guides to the medical literature, XV: how to use an article about disease probability for differential diagnosis. *JAMA.* 1999;281:1214-1219.

39. Guyatt GH, Sinclair J, Cook DJ, Glasziou P. Users' guides to the medical literature, XVI: how to use a treatment recommendation. *JAMA.* 1999;281:1836-1843.

40. Barratt A, Irwig L, Glasziou P, et al. Users guide to medical literature, XVII: how to use guidelines and recommendations about screening. *JAMA.* 1999;281:2029-2034.

41. Randolph AG, Haynes RB, Wyatt JC, Cook DJ, Guyatt GH. Users' guide to medical literature, XVIII: how to use an article evaluating the clinical impact of a computer-based clinical decision support system. *JAMA.* 1999;282:67-74.

42. Bucher HC, Guyatt GH, Cook DJ, Holbrook A, McAlister FA. Users' guides to the medical literature, XIX: applying clinical trial results. A. How to use an article measuring the effect of an intervention on surrogate end points. *JAMA.* 1999;282:771-778.

43. McAlister FA, Laupacis A, Wells GA, Sackett DL. Users' guides to the medical literature, XIX: applying clinical trial results. B. Guidelines for determining whether a drug is exerting (more than) a class effect. *JAMA.* 1999;282:1371-1377.

44. McAlister FA, Straus SE, Guyatt GH, Haynes RB. Users' guides to the medical literature, XX: integrating research evidence with the care of the individual patient. *JAMA.* 2000;283:2829-2836.

45. Hunt DL, Jaeschke R, McKibbon KA. Users' guides to the medical literature, XXI: using electronic health information resources in evidence-based practice. Evidence-Based Medicine Working Group. *JAMA.* 2000;283:1875-1879.

46. McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature, XXII: how to use articles about clinical decision rules. *JAMA.* 2000;284:79-84.

47. Giacomini MK, Cook DJ. Users' guides to the medical literature, XXIII: qualitative research in health care. A. Are the results of the study valid? *JAMA.* 2000;284:357-362.

48. Giacomini MK, Cook DJ. Users' guides to the medical literature, XXIII: qualitative research in health care. B. What are the results and how do they help me care for my patients? *JAMA.* 2000;284:478-482.

49. Richardson WS, Wilson MC, Williams JW, Moyer VA, Naylor CD. Users' guides to the medical literature, XXIV: how to use an article on the clinical manifestations of disease *JAMA.* 2000;284:869-875.

50. Guyatt GH, Haynes RB, Jaeschke RZ, et al. Users' guides to the medical literature, XXV: evidence-based medicine: principles for applying the users' guides to patient care. *JAMA.* 2000;284:1290-1296.

51. Campbell DT, Stanley JC. E*xperimental and Quasi-experimental Designs for Research.* Chicago, Ill: Rand McNally; 1966.

52. Cook TD, Campbell DT. *Quasi-experimentation: Design and Analysis Issues for Field Settings.* Chicago, Ill: Rand McNally College Publishing; 1979.

53. Shadish SR, Cook TD, Campbell DT. *Experimental and Quasi-experimental Designs for Generalized Causal Inference.* Boston, Mass: Houghton Mifflin; 2002.

54. Wilson T, MacDowell M. Framework for assessing causality in disease management programs: principles. *Dis Manag.* 2003;6:143-158.

55. Healy B. The Yentl syndrome. *N Engl J Med.* 1991;325:274-277.

56. Linden A, Adams J, Roberts N. The generalizability of disease management program results: getting from here to there. *Manag Care Interface.* July 2004:38-45.

57. Schulz KF, Chalmers I, Grimes DA, Altman D. Assessing the quality of randomization from reports of controlled trials published in the obstetrics and gynecology journals. *JAMA.* 1994;272:125-128.

58. Schulz KF, Chalmers I, Hayes RJ, Altman D. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA.* 1995;273:408-412.

59. Andersen RM. *Behavioral Model of Families: Use of Health Services.* Chicago, Ill: Center for Health Administration Studies, University of Chicago; 1968. Research Series No. 25.

60. Aday L, Andersen RM. Equity in access to medical care: realized and potential. *Med Care*. 1981;19(12 suppl):4-27.

61. Linden A, Adams J, Roberts N. An assessment of the total population approach for evaluating disease management program effectiveness. *Dis Manage*. 2003;6:93-102.

62. American Healthways and the John Hopkins Consensus Conference. Consensus report: standard outcome metrics and evaluation methodology for disease management programs. *Dis Manage*. 2003;6:121-138.

63. Linden A, Adams J, Roberts N. Using propensity scores to construct comparable control groups for disease management program evaluation. *Dis Manag Health Outcomes*. In press.

64. Linden A, Adams J, Roberts N. Using an empirical method for establishing clinical outcome targets in disease management programs. *Dis Manage*. 2004;7:93-101.

65. Linden A, Roberts N. Disease management interventions: what's in the black box? *Dis Manage*. 2004;7(4):275-291.

66. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: an introduction to survival analysis. *Dis Manage*. 2004;7:180-190.

67. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness adjusting for enrollment (tenure) and seasonality. *Res Healthc Financ Manage*. 2004;9(1):57-68.

68. Linden A, Adams J, Roberts N. *Evaluation Methods in Disease Management: Determining Program Effectiveness*. Washington, DC: Disease Management Association of America; October 2003. Position paper.

69. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: an introduction to time series analysis. *Dis Manage*. 2003;6:243-255.

70. Jollis JG, Ancukiewicz M, Delong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. *Ann Intern Med*. 1993;119:844-850.

71. Hannan EL, Kilburn H Jr, Lindsey ML, Lewis R. Clinical versus administrative data bases for CABG surgery. Does it matter? *Med Care*. 1992;30:892-907.

72. Bland M. *An Introduction to Medical Statistics*. 2nd ed. Oxford, UK: Oxford Medical Publications; 1995.

73. Campbell MJ, Machin D. *Medical Statistics: A Commonsense Approach*. 2nd ed. Chichester, UK: John Wiley and Sons; 1994.

74. Linden A. *A Risk-Adjusted Method for Bed-Day Reporting at CareAmerica Health Plans* [dissertation]. University of California; Los Angeles; 1997.

75. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: an introduction to the bootstrap technique. *Dis Manage Health Outcomes*. In press.

76. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ*. 1995;310:452-454.