

Risk Adjustment: Towards Achieving Meaningful Comparison of Health Outcomes in the Real World

Yew Yoong Ding,¹*FRCP, FAMS, MPH*

Abstract

Health outcomes evaluation seeks to compare a new treatment or novel programme with the current standard of care, or to identify variation of outcomes across different healthcare providers. In the real world, it is not always possible to conduct randomised controlled trials to address the issue of comparator groups being different with respect to baseline risk factors for the outcomes. Therefore, risk adjustment is required to address patient factors that may lead to biases in estimates of treatment effects. It is essential when conducting outcomes evaluation of more than trivial significance. Risk adjustment begins by asking 4 questions: what outcome, what time frame, what population, and what purpose. Next, design issues are considered. This involves choosing the data source, planning data collection, defining the sample required, and selecting the variables carefully. Finally, analytical issues are considered. Regression modelling is central to every analytic strategy. Other methods that may augment regression include restriction, stratification, propensity scores, instrumental variables, and difference-in-differences. The construction of risk adjustment models is an iterative process requiring both art and science. Derived models should be validated. Limitations of risk adjustment include reliance on data availability and quality, imperfect method, ineffectiveness when comparators are very different, and sensitivity to different methods used. Thoughtful application of risk adjustment can improve the validity of comparisons between different treatments, programmes and providers. The extent of risk adjustment should be guided by its purpose. Finally, its methodology should be made explicit, so that informed readers can judge the robustness of results obtained.

Ann Acad Med Singapore 2009;38:552-8

Key words: Health Services Research, Outcome assessment, Regression analysis, Risk adjustment

Introduction

Consider these hypothetical situations. Patients with heart failure managed within a disease management programme are found to have reduced hospitalisations and improved quality of life scores, compared with patients managed outside this programme. Can we conclude that this programme is an effective intervention, and should therefore be expanded to include all patients with heart failure? The length of stay for adults admitted for pneumonia is shorter (5 days) in hospital A, than in hospital B (7 days). Can we confidently say that hospital A is more efficient than hospital B, for the management of pneumonia? With increased availability of outcomes data, questions like these confront physicians, hospitals and the healthcare system in their quest to understand and improve quality of care. The astute are aware that such conclusions may be

overly simplistic and even misleading, unless appropriate measures were taken to ensure that treatment groups were similar with respect to patient factors that could influence outcomes. Without doing so, we cannot be confident that the results are valid.

Typically, health outcomes evaluation seeks to assess a new treatment or novel programme, and compare it with the current standard of care. In addition, there is a growing imperative to identify variation of outcomes across different providers. However, in the real and complex world of healthcare, comparator groups are almost always different with respect to important baseline risk factors for the outcomes. The challenge is to render these comparisons fair, meaningful, and ultimately, useful. How can we best do this?

A time-honoured solution is to conduct randomised

¹ Department of Geriatric Medicine, Tan Tock Seng Hospital, Singapore
Address for Correspondence: Clin Assoc Prof Ding Yew Yoong, Tan Tock Seng Hospital, 11 Jalan Tan Tock Seng, Singapore 308433.
Email: yew_yoong_ding@ttsh.com.sg

controlled trials to address selection bias and confounding. While their results may represent the highest level of evidence for single studies, these trials may not always be feasible in the real world due to administrative, logistic, and ethical considerations. Instead, we often need to rely on observational data collected prospectively, or obtained retrospectively from routinely collected information sources. Therefore, it is incumbent upon investigators to address patient factors that may lead to biases in estimates of treatment effects. This is where risk adjustment finds its privileged position in health outcomes evaluation and research.

This review is based in part on principles taught by Lisa Iezzoni in her book entitled “Risk Adjustment for Measuring Health Care Outcomes”,¹ and on teaching points shared by Dan Berlowitz of Boston University. Examples of risk adjustment from the literature and personal reflections have been incorporated. The intent is to provide a brief overview of risk adjustment and to offer a framework for performing risk adjustment. The hypothetical example of evaluating a heart failure management programme will be used to illustrate how basic risk adjustment may be performed. For more detailed study, readers are encouraged to examine the referenced texts and articles.

Risk Adjustment: What, Why and When

Risk adjustment is a strategy for reducing the effect of confounding factors in studies where patients are not randomly assigned to different treatments. It accounts for patient factors that could affect outcomes, and that exist prior to the intervention.² Meaningful comparisons within the healthcare system generally require risk adjustment.¹ The purpose is to “make the playing field level”, so that the effectiveness of treatments, programmes, and providers can be fairly evaluated and compared. Risk adjustment is integral to calculating the “algebra of effectiveness”.¹ This concept is represented mathematically by,

Outcomes

$$= f(\text{patient factors, treatment effectiveness, quality of care, random chance})$$

where f denotes “function of” the list of factors in the parentheses.

The underlying idea is when we address patient factors through risk adjustment, we will then be able to isolate and quantify treatment effectiveness and quality of care. This is illustrated in Figure 1.

Risk adjustment is appropriate when there are significant differences across treatment groups with respect to covariate distributions. While it may be relevant to experimental studies where randomisation has not successfully balanced baseline characteristics across the intervention and control



Fig. 1. Algebra of effectiveness (adapted from Iezzoni¹).

groups, its greatest value lies in the analysis of observational studies. Risk adjustment is commonly applied in treatment assessment, programme evaluation, provider profiling, and performance monitoring. It should always be considered when the significance of the evaluation is more than trivial, and when research seeks to compare treatments and programmes. The end-point is the gathering of valid evidence on causal effects of interventions and providers, which in turn informs decision-making in health policy, and guides report cards.

Performing Risk Adjustment

Iezzoni suggests that in devising risk adjustment strategies, answers to 4 major questions should be sought¹:

(i) Risk of what outcome?

The categories of outcomes are: clinical outcomes, such as mortality, morbidity, and quality of life; patient-centred outcomes, such as satisfaction with care; and utilisation outcomes, such as cost of care, length of stay, ambulatory care visits, and hospitalisation episodes. For our heart failure management programme example, the outcomes sought could include mortality, quality of life, patient satisfaction, and hospitalisation days.

(ii) Over what time frame?

Risks need to be framed within specific time windows, such as hospital stay, or 30-days, and 1-year post-event. For our heart failure programme example, a suitable time frame could be 1 year after the index hospitalisation.

(iii) For what population?

Risk factors vary across different populations. Therefore, it is important to define the population of interest. Typical examples include children, women, ICU patients, hospitalised patients, and frail older persons. The population of interest in our example could be adult patients admitted to the acute care hospital with the principal diagnosis of congestive heart failure.

(iv) For what purpose?

The purpose of risk adjustment dictates how well it must perform to be successful. Examples include comparing

efficiency across providers, encouraging providers to treat high risk patients, providing feedback for quality improvement, and public reporting. The purpose in our example could be to determine the effectiveness of the heart failure programme, with a view of deciding on whether to expand it across a group of affiliated hospitals.

The answers to these 4 questions guide the conduct of the 2 aspects of risk adjustment: design and analyses.

(I) Design

Good risk adjustment starts with attention to design issues. Firstly, we need to consider the study data. Possible options for the data source are:

- (i) Administrative data: The advantages are the availability of large sample size, and relatively inexpensive and non-intrusive data collection. However, this data type is usually not clinically detailed.
- (ii) Clinical data: These are obtained from paper and electronic medical records. Although more clinically detailed, they are more expensive and time-consuming.
- (iii) Survey data: These are rich in patient-centred information, especially in the functional and social domains. The collection process is also resource-intensive.

Intuitively, we expect that clinically-rich data should facilitate better risk adjustment. Indeed, coronary bypass surgery report cards based on a validated Massachusetts clinical registry were found to be less problematic than those based on administrative data.³ In addition, data assembled from a combination of sources may achieve better risk adjustment. This was illustrated when models using combined self-reported and administrative data performed better than models using only either data type in the prediction of mortality and hospitalisations among veterans receiving outpatient care in the Veteran Affairs.⁴ Data collection needs to be planned carefully and carried out meticulously to obtain data of good quality, because statistical techniques cannot correct measurement error. For our heart failure programme example, we could choose to assemble information from administrative data (especially utilisation data), medical records data (especially clinical and investigative data), and survey data (interview of patients on their symptoms, function, and perception).

Secondly, the study sample is constructed to obtain the appropriate data to best answer the research or evaluation question, and at the same time avoid excessive data that introduces additional noise and inefficiency in the analyses.

Thirdly, study variables need to be thoughtfully selected. The practice of “more is better” should be avoided. Rather, a measured approach that only selects variables that are relevant to the outcome measure and the intervention is suggested. The spectrum of risk factors listed in Table 1

Table 1. Range of Risk Factors for Possible Independent Variables (Adapted From Iezzoni)¹

Demographic
Age
Gender
Race or ethnicity
Clinical
Acute physiological stability
Principal diagnosis (including severity)
Comorbidities Physical functional status
Cognitive status
Mental health (including mood)
Others – e.g. atypical presentations of illness
Socioeconomic
Education
Finance status
Employment
Marital status
Housing characteristics
Caregiver availability
Health-related behaviours
Smoking
Drinking
Health-related perceptions
Self-reported overall health status
Quality of life

may be represented by independent variables. The importance of careful selection of variables was demonstrated in a study on provider profiling for asthma care in California, where risk adjuster selection wielded more influence on ranking profiles than choice of statistical strategies.⁵ It is a good practice not only to list the selected variables, but also to identify potentially important variables that were omitted and to make this explicit. This is so that alternate proxy variables may be sought. In addition, knowledge that specific variables were omitted can assist in judging the degree of uncertainty surrounding estimates of treatment effects, including the likely direction and size of biases. For our heart failure programme example, study variables should include those relating to severity of heart failure, comorbidity, social support, and socioeconomic status, in addition to the usual demographic variables.

Issues relating to comorbidities merit added discussion. Comorbidities are conditions that are causally unrelated to the principal diagnosis (PD), and are best measured by indices⁶ or condition sets.⁷ On the other hand, complications

are conditions that are causally related to the PD (ie, disease complications), or its treatment (ie, treatment complications). It is important to determine whether these complications occurred before or after treatment for the PD was started, because when occurring after, they should be viewed as outcomes to be measured, rather than risk adjusters to be controlled for. Otherwise, risk adjustment may yield perverse results that bias the estimates of treatment effect.⁸ In practice, it may be difficult to differentiate between comorbidities and disease complications on one hand, and treatment complications on the other hand, particularly when working with administrative data. Clinical data is usually required. Even then, it may not always be possible to decide whether relationships are causal or not. Nevertheless, the guiding principle is that risk adjusters should include comorbidities and complications that are present before treatment, but not complications that occur after treatment. Ultimately, good clinical knowledge and sound judgement are required to make appropriate decisions on this matter.

(II) Analyses:

The range of risk adjustment methods may be categorised as standard or advanced.

The standard methods are:²

- (i) **Restriction:** Certain subjects (eg very old) are excluded from the analyses. The idea is to trim the sample by dropping subjects with extreme values on variables that have uneven distribution across treatment groups. The purpose is to render remaining subjects more comparable, at least with respect to those specific variables. For our example, we could restrict the study population to those patients hospitalised for the first time because of heart failure, and who were aged 50 to 80 years of age. This is because almost all heart failure patients who were young were managed by the programme, whereas those who were very old were not. The result is significantly different distributions of age across the programme and non-programme groups, which may not be adequately controlled for using regression alone.
- (ii) **Stratification:** Subjects are divided into a small number of groups on the basis of a factor that is a confounder for the treatment-outcome relationship. An example is stratification by age groups where mortality is the outcome of interest. Outcomes are then analysed separately within these strata, thereby controlling for confounding by a single variable. For our example, we could stratify patients according to 10-year age bands.
- (iii) **Multivariable regression:** In most studies, there are multiple confounders to control. Stratification is not feasible as it can best handle only a few variables at the

same time. Instead, simultaneous statistical adjustment for multiple factors is achieved by regression modelling. Where the outcome (dependent) variable is continuous or dichotomous, linear or logistic regression is used respectively. Quite often, outcomes are in counts (events) or duration (time-to-event). In these instances, poisson regression, and survival (eg Cox regression) analysis are used respectively. Addition of hierarchical modelling can produce more appropriate results where subjects are clustered.⁹ A good example of risk adjustment using regression modelling is the prediction of 30-day and 1-year mortality after acute myocardial infarction using Ontario administrative data.¹⁰ Even with the availability of more advanced methods, regression modelling remains central to most risk adjustment strategies and should always be performed, either alone or in combination with other methods that serve to augment the capability of regression for this task. For our example, we could evaluate the effectiveness of the heart failure management programme by using Cox regression modelling of mortality over the first year, while adjusting for multiple baseline patient factors. Construction of regression models is discussed later.

The advanced methods are:

- (i) **Propensity scores:** More recently, advanced statistical and econometric methods have been used in health services research. They expand the toolbox of methods available to deal with the fundamental problem of selection bias, which is also known as the “evaluation problem” in the econometric literature.¹¹ Propensity scores is one of these methods. It represents the probability of receiving a treatment or participating in a programme, conditional on a set of covariates that reflect baseline characteristics. It is a scalar summary of all observable confounders and is represented by a score (0 to 1). Using their individual scores, subjects are matched, stratified into a few levels (most often, quintiles), or have an additional variable to be used in regression modelling. Outcomes are then analysed with subjects as matched pairs, in separate strata, or with an additional variable. This method can only address measured factors that may lead to selection bias (selection on observables), but cannot remove hidden biases (selection on unobservables). One of its advantages is the ability to explicitly determine the degree to which we have balanced measured characteristics across treatment groups.¹² It is therefore able to warn investigators of inadequately overlapping covariate distributions that would render risk adjustment untrustworthy. On account of this advantage, it has been suggested that propensity scores

should be used first to determine whether it is feasible to proceed with further analyses.¹³ Its application in risk adjustment was demonstrated in studying the treatment effect of invasive cardiac management on survival after acute myocardial infarction, where the propensity score was the probability of receiving cardiac catheterisation.¹⁴ In recent years, propensity scores have been increasingly used in observational studies, and should be viewed as an additional risk adjustment tool.¹⁵

- (ii) Instrumental variables: To address unmeasured or unknown confounders, an “instrument” that is associated with probability of receipt of the intervention, but not associated directly with the outcome of interest, is identified. Typically, this “instrument” is used in a two-stage least squares (2SLS) regression. Although instrumental variables have been used in econometric research for decades, this method has only recently been applied to analysis of healthcare data. In a landmark observational study using instrumental variables that examined whether more intensive management of elderly patients with acute myocardial infarction (AMI) reduced mortality in the United States, the “instrument” used was the differential distance to hospitals offering cardiac catheterisation.¹⁶ More recently, the same issue was studied further using a similar instrument among patients with AMI in 3 Canadian states.¹⁷ Yet another study used regional cardiac catheterisation rates as the “instrument”.¹⁴
- (iii) Differences-in-differences (DID): Unmeasured or unknown confounding can also be addressed by using repeated measures on the same subject in quasi-experimental studies or “natural” experiments. The causal effect is estimated using the average change of the dependent variable in the group receiving the treatment of interest over the period of time, minus the corresponding average change in the group not receiving the treatment over the same period.¹⁸ This is often combined with regression modelling. DID was used in a study that examined the effect of transition to a high-deductible health insurance plan, on emergency department use and hospitalisations in Massachusetts.¹⁹

The choice of risk adjustment method for a particular study depends on the research question to be answered, purpose of the risk adjustment, data availability, and expertise of the analyst. More often than not, a combination of methods is appropriate.

The construction of risk adjustment models is as much an art as it is a science. Comprehensive discussions and recommendations are found in current statistical and

econometric texts. A simplified approach for risk adjustment is offered here. It starts by specifying the outcome (dependent) measure, and paying attention to its time frame (at what point in time), and structure (dichotomous vs. continuous). Next, the independent variable of interest is identified. In health services research, this is usually a treatment, programme, or provider. Other independent variables are selected on the basis of their a priori relationships with both the outcome variable and independent variable of interest. These include all relevant dimensions of risk based on clinical knowledge. In addition, potential interactions between independent variables are considered, and interaction variables are introduced where relevant. It is prudent to avoid including variables without any plausible link to the outcomes studied. The structure of each independent variable needs to be specified in its appropriate form on the basis of clinical knowledge and the results of bivariable analyses. Multivariable regression modelling is an iterative process, and the goal is to obtain a clinically credible model with good statistical properties. Once the model is decided upon, it is important to conduct specific tests on its performance on the dataset. For linear regression models, the R^2 is computed to estimate the proportion of variability explained by the model. For logistic regression models, calibration is examined by determining how well average of predicted values match average of actual values. In addition, discrimination is also estimated using the c-statistic (area under curve). The model is repeatedly modified until the investigator is confident that its performance is optimised. Finally, the model is validated on another sample,² or using methods such as data-splitting or bootstrapping. An example of the methodical approach used in derivation and validation of risk adjustment models is found in the reports on predicting pressure ulcer development among nursing home residents^{20,21}

Limitations of Risk Adjustment

Like any other strategy, risk adjustment has limitations that should be borne in mind. They are:

- (i) Reliance on data availability and quality: To a large extent, the validity of risk adjustment depends on the availability and quality of data. No degree of sophistication in statistical analyses can adequately compensate for poor or missing data. Therefore, obtaining good data that is as complete and accurately measured as possible, is key to good risk adjustment.
- (ii) Imperfect method: At best, risk adjustment achieves an approximation of the true relationship between treatments and outcomes. There is no such thing as perfect risk adjustment.¹
- (iii) Ineffectiveness when comparators are very different: Risk adjustment cannot compare “apples and oranges”.

It is ineffective when comparators are very different with respect to covariate distributions. One way of assessing this is to use propensity scores to determine the “support” (or overlap) between comparator groups. Unless there is reasonable overlap in case-mix distributions, risk-adjusted outcomes should not be used to directly compare the performance of one provider with another.²²

- (iv) Sensitivity to methods used: Risk-adjusted outcomes may vary with the methods used. Different methods of adjusting for severity influenced the results on survival of patients with congestive heart failure in report cards.²³ Reporting the results of different methods as a form of sensitivity analysis can help readers make better judgements on the validity of conclusions made.

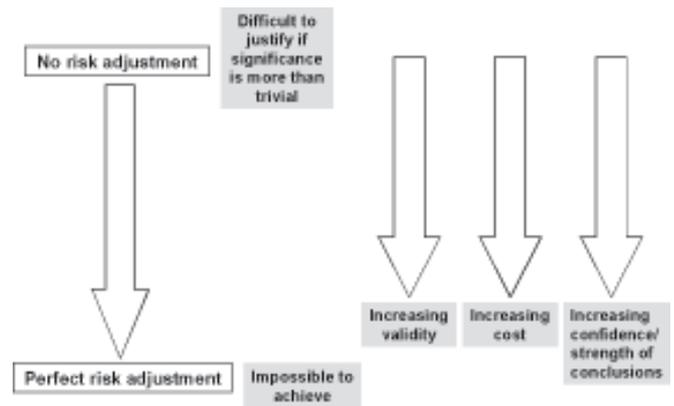


Fig. 2. Extent of risk adjustment: implications.

Conclusion

Despite its limitations, risk adjustment is essential for serious comparison of outcomes in evaluation and research. A suggested framework for risk adjustment is described in Table 2. Good and comprehensive data should be strived for. It is important to assess the degree of covariate differences between comparators, and to be cautious when they are great. The degree to which we can be confident of the causal relationship between treatment and outcome is dependant on the quality of risk adjustment performed. Risk adjustment comes with a price, and the tradeoffs are illustrated in Figure 2. Ultimately, wise decisions on the extent of risk adjustment appropriate for different purposes and situations are required to be made.

Finally, it is not adequate to simply state that outcomes are “risk adjusted”. Rather, risk adjustment methodology needs to be made explicit, so that informed readers can decide on the robustness of results, particularly where health policy and report cards are at stake.

Table 2. Suggested Framework for Performing Risk Adjustment (Adapted From Iezzoni¹ and Berlowitz²)

Steps	Activities
4 major questions	1. Risk of what outcome?
	2. Over what time frame?
	3. For what population?
	4. For what purpose?
Study design	1. Choose data source
	2. Define sample
	3. Select variables
Study analyses	1. Choose methods <ul style="list-style-type: none"> • restriction • stratification • multivariable regression • others: e.g. propensity scores, instrumental variables, differences-in-difference
	2. Construct risk adjustment model <ul style="list-style-type: none"> • derivation • validation

REFERENCES

1. Iezzoni L, editor. Risk Adjustment for Measuring Health Care Outcomes. 3rd ed. Chicago: Health Administration Press, 2003.
2. Berlowitz DR. Measuring healthcare quality [course materials]. USA: Boston University, 2006.
3. Shahian DM, Silverstein T, Lovett AF, Wolf RE, Normand S-L T. Comparison of clinical and administrative data sources for hospital coronary artery bypass graft surgery report cards. *Circulation* 2007;115:1518-27.
4. Fan VS, Maciejewski ML, Liu C-F, McDonell MB, Fihn SD. Comparison of risk adjustment measures based on self-report, administrative data, and pharmacy records to predict clinical outcomes. *Health Serv Outcomes Res Method* 2006;6:21-6.
5. Huang I-C, Dominici F, Frangakis C, Diette GB, Damberg CL, Wu AW. Is risk-adjuster selection more important than statistical approach for provider profiling? Asthma as an example. *Med Decis Making* 2005;25: 20-34.
6. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chron Dis* 1987;40:373-83.
7. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Med Care* 1998;36:8-27.
8. Smith MA, Nitz NM, Stuart SK. Severity and comorbidity. In: Kane RL, editor. *Understanding Health Care Outcomes Research*. Sudbury, Massachusetts: Jones and Bartlett Publishers, 2006.
9. Shahain DM, Torchiana DF, Shemin RJ, Rawn JD, Normand S-LT. Massachusetts cardiac surgery report card: implications of statistical methodology. *Ann Thorac Surg* 2005;80:2106-13.
10. Tu JV, Austin PC, Walld R, Roos L, Agras J, McDonald KM. Development and validation of the Ontario acute myocardial infarction mortality prediction rules. *J Am Coll Cardiol* 2001;37:992-97.
11. Jones A. Introduction: the evaluation problem and linear regression. In: *Applied Econometrics for Health Economists*. 2nd ed. Oxford: Radcliffe Publishing, 2007.

12. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Stat Med* 2006;25:2084-106.
 13. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997;127:757-63.
 14. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA* 2007;297:278-85.
 15. D'Agostino Jr RB. Propensity scores in cardiovascular research. *Circulation* 2007;115:2340-3.
 16. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA* 1994;272:859-66.
 17. Pilote L, Beck CA, Eisenberg MJ, Humphries K, Joseph L, Penrod JR, et al. Comparing invasive and noninvasive management strategies for acute myocardial infarction using administrative databases. *Am Heart J* 2008;155:42-8.
 18. Stock JH, Watson MW. Experiments and quasi-experiments. In: *Introduction to Econometrics*. 2nd ed. Boston: Pearson, 2007.
 19. Wharam JF, Landon BE, Galbraith AA, Kleinman KP, Soumerai SB, Ross-Degnan D. Emergency department use and subsequent hospitalizations among members of a high-deductible health plan. *JAMA* 2007;297:1093-102.
 20. Berlowitz DR, Bradeis GH, Morris JN, Ash AS, Anderson JJ, Kader B, et al. Deriving a risk-adjustment model for pressure ulcer development using the minimum data set. *J Am Geriatr Soc* 2001;49:866-71.
 21. Berlowitz DR, Bradeis GH, Anderson JJ, Ash AS, Kader B, Morris JN, et al. Evaluation of a risk-adjustment model for pressure ulcer development using the minimum data set. *J Am Geriatr Soc* 2001;49:872-6.
 22. Shahain DM, Normand S-L T. Comparison of "risk-adjusted" hospital outcomes. *Circulation* 2008;117:1955-63.
 23. Poses RM, McClish DK, Smith WR, Huber EC, Clemo LW, Schmitt BP, et al. Results of report cards for patients with congestive heart failure depend on the method used to adjust for severity. *Ann Intern Med* 2000;133:10-20.
-