the EVALUATION@HSRI CENTER

# *Conducting Case-Mix Adjustment for Mental Health Performance Indicators*

Michael Hendryx, Ph.D.

Washington Institute for Mental Illness Research and Training

Washington State University

*May 2004*

*Appendix:*

## *Use of Hierarchical Linear Models in the Case Mix Adjustment of Mental Health Provider Profile Scores*

Brain J. Cuffel, Ph.D.

Vice President, Research and Evaluation

***Human Services Research Institute***
2269 Massachusetts Avenue
Cambridge, MA 02140
www.tecathsri.org

U.S. Department of Health and Human Services
Substance Abuse and Mental Health Services Administration
Center for Mental Health Services
www.samhsa.gov

# Part I. Using Case-mix Adjustment



# Part II. Exercises

# I. Toolkit Overview

## A. Purpose of This Toolkit

The purposes of the toolkit are, first, to provide a description of the definition, rationale, limitations, required tasks, and analytic methods of mental health case-mix adjustment; and second, to provide the reader with computer exercises using a hypothetical database to practice conducting a case-mix adjustment using either SAS or SPSS. (Case-mix adjustment is also commonly known as "risk adjustment." We use the first term to emphasize that the process described here is primarily for benchmarking, comparison, and quality improvement, rather than for actuarial purposes.

## B. Toolkit Content and Intended Audiences

The toolkit is divided into two main parts corresponding to its two purposes. Part 1 contains an introduction to mental health performance indicator case-mix adjustment, designed for anyone with an interest in the topic and a basic appreciation of statistics, research methods and designs. It may be considered a conceptual overview for state mental health authority directors, quality improvement managers, or mental health agency administrators. Part 1 addresses a) the definition and rationale for case-mix adjustment, b) the criteria for selection of performance indicators and case-mix indicators, c) when case-mix adjustment may be unnecessary, and d) the tasks that must be done in order to do case-mix adjustment (study design, systematic data definition and collection methods across performance sites, storing and managing data, choosing methods of data analysis and display of results, and using case-mix adjusted results to improve quality of care).

Following the discussion of these issues, Part 2 of the toolkit provides a fabricated dataset and exercises in SAS and SPSS on a CD to practice conducting a case-mix adjustment analysis. See Table 1 for a list of the CD file names and contents. Part 2 is designed primarily for data analysts, data managers, or other persons who are directly responsible for the management and analysis of performance indicators in public or private mental health services settings and who are familiar with statistical analysis programming using either SAS or SPSS. However, the text description contained in Part 2 may be of value to persons who wish to examine how case-mix adjustment is undertaken without actually conducting the practice exercises. The toolkit is not intended to serve as a detailed guide to interpreting quality assurance or improvement data or conducting quality improvement activities; such guides are available from other sources.

## C. Toolkit Prerequisites

The toolkit is not intended as a statistical primer and assumes that the reader has an understanding of research methods and statistical techniques, including:

- Basic descriptive and inferential statistics
- Sampling and sample size issues
- Design and measurement principles, such as the basics of psychometrics (reliability and validity) and experimental design.

# II. Definition and Rationale for Case-mix Adjustment

## A. What is Case-Mix Adjustment?

Case-mix adjustment is the process of statistically controlling for group differences when comparing non-equivalent groups on outcomes of interest. It is done on a post-hoc basis, after the treatment groups have been formed and the performance measures collected. The groups may be treatment agencies, consumers, providers, programs, regions, or states. Any time these groups are to be compared on performance indicators, case-mix adjustment must be considered. For the sake of illustration, the following descriptions assume that we are interested in comparing the performance of multiple mental health treatment agencies that constitute a larger treatment system, but it should be kept in mind that the comparison may be of other types of groups as well.

## B. Why Do Case-Mix Adjustment?

Mental health authorities and providers in both the public and private sectors are increasingly interested in measuring outcomes of mental health care. Outcome or performance measurement serves a number of purposes: to set outcome expectations, guide quality improvement, monitor intervention effects, assist purchasers in choosing providers, and compare the performance of groups of providers (Eddy 1998). Performance measurement is mandated by some state public mental health systems and managed care organizations.

By using comparative performance indicators, mental health systems can track the effects of changes within their systems and the effectiveness of routine care provision across sites. They can identify sites providing the highest quality care and sites that may need to improve the quality of care they provide. Comparative performance indicators may be summarized in written form and distributed as "report cards"; such report cards may include a variety of specific performance indicators. The report cards must be provided within a framework that acknowledges the issues that arise when comparisons are made among different populations or using different assessment instruments and data collection methodologies. These issues demand the development of case-mix adjustment tools (Hendryx, Beigel and Doucette 2001).

Populations of mental health consumers served by different behavioral health care agencies can be vastly different. Agencies serving individuals with severe and co-morbid impairment cannot equitably be compared using raw outcome scores to agencies serving individuals with less challenging mental health concerns. The outcomes that providers or agencies strive for, and for which they are held accountable, are only partly under their control; many person and environmental variables affect outcomes independently of care (e.g., Hendryx and Teague 2001; Dow et al. 2001; Banks, Pandiani and Bramley 2001). These critical case-mix variables are not evenly distributed across groups. We need to ask how indicators can be compared across different agencies if the agencies treat a different mix of clients or otherwise vary in important ways beyond agency control. Case-mix adjustment attempts to identify the person and environmental variables that influence outcomes, measure those variables, correct for their influence through post-hoc statistical methods, and display the case-mix adjusted results in ways that allow for ease of interpretation and use.

Case-mix adjustment is a partial correction that cannot create perfectly equivalent groups or duplicate the rigor of experimental assignment (Iezzoni 1997). In a true experiment, the researcher assigns people randomly to different treatment groups, controls the administration of the treatment, and measures the outcome or dependent variable. Statistical laws tell us that, with enough people, the average characteristics will be equal in all groups;

the only systematic variation is the treatment. So if the results show that the groups are unequal on the dependent variable, one concludes that the treatment caused the difference.

To apply the analogy of the experiment to behavioral health, the treatment agencies are like the various treatment groups in an "experiment," with performance therefore being like the effects of treatment. We are still interested in whether the treatment causes higher or lower scores on the performance measures, but the critical difference is the absence of random assignment. People are not randomly assigned to the agencies being compared; rather, they self-select to one or another based on geographic proximity, agency specialty, treatment needs, or other factors. Case-mix adjustment is a post-hoc effort to correct for these differences among the groups served by the agencies. In this manner we may view case-mix adjustment as similar to quasi-experimental studies (Cook and Campbell 1979; Shadish, Cook and Campbell 2001). Case-mix variables are thereby analogous to the covariates in a quasi-experimental design or an analysis of covariance or multiple regression analysis.

Case-mix adjustment has an additional function in setting appropriate reimbursement rates in capitation contracts. Adequately and fairly compensating providers on the basis of how much service will be needed, as indicated by case-mix adjustment, removes the incentive for providers to attract only those who are relatively healthy and avoid those with more severe conditions that will require more services (see section III.b.6).

### C. When is Case-Mix Adjustment Unnecessary?

As suggested by some of the foregoing discussion, there may be situations where case-mix adjustment is unnecessary. This situation will occur when the case-mix adjusted results lead to the same conclusions as the unadjusted results regarding group level performance. It may also occur when the gain from doing case-mix adjustment is considered to be small relative to the costs, or when the potential case-mix indicators that are available in a limited dataset do not correlate with the outcome. In the latter case, it is important to recognize that any results to be compared among groups are unadjusted and therefore potentially misleading.

## III. Criteria for Selection of Performance Indicators and Case-mix Indicators

*Common mental health performance indicators include such things as hospitalization or rehospitalization, employment, incarceration, functioning, and symptom severity. When choosing indicators, the following criteria may be used to guide the selection process.*

### A. Characteristics of Good Outcome or Performance Measures:

1 **Outcomes are important to stakeholders.** The outcomes have to be the important ones from the perspective of one or more stakeholder groups such as consumers, advocates, providers, payers, and administrators.

2 **Measures are reliable and valid.** Reliable measures are those that are repeatable or stable over conditions and contain little measurement error. Valid measures are those that are true: for example, a self-reported assessment of symptom severity really measures symptom severity, not some other characteristic of the individual. Reliability and validity are basic measurement requirements, but sometimes in mental health administrative databases the reliability and/or validity of performance indicators is poor or unknown. We should exercise caution in comparative analysis of groups unless we are assured that we are working with measures that are as reliable and valid as possible. However, in developing "real world"

performance indicator databases, it may be necessary to make comparisons even though the quality of the data is suspect; in such cases, we should recognize the limits of the data and treat the results as suggestive rather than definitive, while striving to do what we can to improve data reliability and validity.

The extent to which data are missing or inaccurate affects data reliability and validity. For example, let us suppose that mental health providers are instructed to complete a data form on all clients every 90 days, but providers are not given any incentive to complete the forms accurately – the data do not influence their job or their client relationship in any observable way. As a result, the quality of the data may be poor or providers may not complete the form in full. Let us assume that one of the case-mix indicators is marital status. If a client experienced a change in marital status, for example recently becoming divorced, that may be an important case-mix indicator for poorer outcome. If that change is not captured in the data, however, the case-mix indicator has greater error and reduced reliability and validity. The case-mix adjusted results will fail to account accurately for outcomes associated with marital status, and the agency may appear to be performing worse than it really is.

Another component of reliability and validity is the requirement that measures from different groups be collected in the same way. For example, if some consumer surveys are done by mail, others by telephone, and still others in-person, and the method of survey administration influences people's responses, it will be difficult, if not impossible, to compare the results validly, with or without case-mix adjustment.

3 **Outcomes are within the control of agencies to influence.** The outcome measure must be one that agencies have some realistic opportunity to affect through services. If providers are expected to keep consumers out of the hospital, this is an appropriate outcome measure. If they are not expected, for example, to improve children's grades in school, it is not appropriate to select this as an outcome measure.

4 **Outcomes are also influenced by variables outside provider control.** Case-mix adjustment is appropriate only with the presence of case-mix variables. These are variables that agencies can't control (e.g., age, age at onset of mental illness, marital status) and that influence the outcome measure.

5 **Measures are outside the control of providers to game.** This consideration could be subcategorized under the validity of measures but deserves special comment. If agencies are under pressure to demonstrate effectiveness, and if they also control the performance scores, this creates a perverse incentive for them to intentionally or unintentionally manipulate scores in their favor. If there is an instance where a performance measure collected by a provider must be used (e.g., a Global Assessment of Functioning score is the only available functioning measure), this measure must be treated with caution and should ideally be subject to some type of independent validation or audit, or efforts should be undertaken to supplement this vulnerable measure with others.

Ways to make measures less susceptible to gaming are to:

- Collect data from consumers rather than providers;
- Audit data collection for accuracy, as is done in the National Association of State Mental Health Program Directors Research Institute (NRI)'s Oryx system and in Indiana's public mental health system (DeLiberty, Newman and Ward 2001);
- Base outcome scores on multiple measures and sources (e.g., combining a functioning outcome score from consumer, provider, and family ratings) so that no single data source determines the result.

As with the goal of establishing reliability and validity, it may not always be possible to safeguard against gaming in real applications, but where there is a possibility that gaming may influence results, this must be recognized and the results interpreted cautiously.

## B. Characteristics of Good Case-Mix Indicators

The preceding section described desirable characteristics of good performance and outcome measures in general. Here, we discuss how those criteria apply to case-mix variables specifically. Commonly used case-mix indicators include such things as demographic variables, and clinical variables such as diagnosis, baseline functioning, or baseline symptom severity. Others may include co-occurring substance abuse, a history of poor medication compliance, weak social support, and early age of illness onset. Choice of case-mix indicators may be guided by the following criteria.

1 **Measurable reliably and validly.** See II.A.2. above. The same principles of data reliability and validity that applied to outcomes also hold for case-mix indicators.

2 **Correlated to the outcome.** Case-mix variables should be significantly correlated to the outcome measure in a multivariate context (i.e., correlated to the outcome after statistically controlling for other case-mix variables). A simple correlation between case-mix and outcome, without considering other variables, is not sufficient, as the correlation may no longer be important when other case-mix variables are taken into account. Inclusion of variables that do not correlate to the outcome may contribute to the error or imprecision of the model.

The strength of the necessary correlation is open to some interpretation, however. One may require case-mix variables to relate to the outcome in a multivariate context at a conventional p value, such as $p<.05$. Alternatively, variables may be retained in models when they offer more modest association, using a rule of thumb such as $p<.20$ or t values $> 1.0$. The possible advantage of a more lax inclusion criterion is that it will include more case-mix variables in a multiple regression model, which can then be subject to cross-validation. The exercises presented on the CD in Part 2 use a more lax inclusion criterion and demonstrate a method of cross-validation (Hendryx, Dyck and Srebnik 1999.)

However, if one is not using a regression model but a simpler approach that considers only single case-mix indicators, a simple bivariate correlation is sufficient. As will be discussed later, regression models offer several advantages to other methods of case-mix adjustment, but there may be instances where single case-mix indicators are all that is required.

3 **Outside the control of agencies to influence.** Case-mix variables are those that influence the outcome but are not under the control or influence of agencies (e.g., age, age at onset of illness, marital status). This is not always an obvious decision, and there is some gray area here. For example, prior service use is often a strong predictor of future service use, but if an agency allows consumers to be hospitalized more than necessary, over time their prior hospitalization history looks as though they treat a more severely ill population when this is not necessarily the case.

4 **Variable among provider groups.** This is not a hard and fast requirement but seems likely to be the case in most instances. If there is a particular case-mix variable (e.g., functional impairment at admission

to service) that is associated with the outcome measure (e.g., functioning at follow-up), but agencies are not different on this case-mix variable, its importance as a case-mix variable is diminished, as all agencies face the same effect of case-mix. When conducting a case-mix adjustment analysis, one does not need to test each case-mix variable for unequal distribution among groups. Whether the case-mix variable is ultimately called for is determined by whether the case-mix adjustment analysis actually changes group level performance (see #7 below.)

5 **Established or theoretical relationship to outcome.** Avoid "data dredging" or examining every possible correlation between potential case-mix measures and the outcome, because this procedure will likely result in including case-mix variables that are identified by chance and are not reliable. It is preferable to begin with a limited set of case-mix variables that theory and prior research identify as consistent predictors of the outcome measure and to subject those variables to the other criteria listed here to determine whether or not they are ultimately included. Again, the realities of a given data set will dictate what is available, and some case-mix variables that one may wish to include in an ideal data set may not be present. To the extent that this is the case, the possible limits of the case-mix adjusted results should be acknowledged.

6 **Not gameable.** See II.A.5. above. For example, if providers know that persons with a diagnosis of schizoaffective disorder are at higher case-mix, and their agency will look good if they treat a higher proportion of persons with this disorder, they may be more likely to assign this diagnosis to consumers. The influence of gaming can be reduced by using multiple case-mix variables, at least some of which come from sources not under the control of the providers or agencies, and by conducting audits for data accuracy.

Another danger related to gaming is the possibility that providers or insurers can deny consumers with more severe illnesses entry into treatment. This is called "skimming" or "creaming." Ideally, case-mix adjustment is a safeguard against this practice because those variables that are used to deny treatment will be included as case-mix variables, and the groups that are most willing to accept consumers with more severe illness will receive proper benefit for it in the case-mix adjusted outcome scores. If all groups deny treatment to certain types of clients, than this is an issue that must be addressed outside of a case-mix adjustment context.

7 **Makes a difference in final performance interpretation.** Conducting a case-mix adjustment takes time and resources. Even if an analysis of case-mix results in a significant model (e.g., a significant adjusted $R^2$ in a regression analysis), the case-mix adjustment is not useful unless it tells us something different about agency performance than what we were told by the unadjusted analysis. The case-mix adjusted scores have to be different in some meaningful fashion from the unadjusted scores at the level of the group. The examples later in the toolkit will illustrate this meaningful difference.

8 **Does not disadvantage vulnerable groups.** The classic case of this concern is race of the consumer. For example, if black consumers experience worse outcomes on a particular measure compared to other groups, is it appropriate to include race as a case-mix indicator? If we include it, do we in effect accommodate racial disparities by assuming that blacks will have worse outcomes? On the other hand, if we do not include race, do we disadvantage those treatment groups that treat a higher proportion of black consumers? The trend in case-mix adjustment is *not* to include race as a case-mix predictor, but 1) to

conduct case-mix adjustment analyses separately for different race groups, and/or 2) in a multiple regression context, to test final case-mix adjustment equations by correlating prediction errors to race. These correlations should be nonsignificant, indicating that the equations are not biased with respect to race.

# IV. Preparing for Case-Mix Adjustment

*Prior to the case-mix adjustment analysis itself, a series of tasks are required. This section briefly describes these preliminary steps of selecting case-mix and outcome variables and planning methods of data collection.*

## A. Choosing Performance Measures and Case-Mix Variables

The first step is usually to choose the outcome or performance indicators. These should be selected based on consideration of theory, local program objectives, empirical research evidence from prior studies, and input from stakeholder groups.

Once performance measures are selected, the process of identifying an appropriate set of case-mix adjustment variables may begin. Selection of case-mix variables also should be based on theory and prior research (i.e., what are the important case-mix variables for a given outcome) but should take into consideration as well the available data collection resources.

## B. Preparing for Data Collection

Planning for data collection and analysis should be done early in the process. A common mistake in planning performance measurement systems is to give all the attention to choosing the outcome indicator and little attention to the case-mix indicators or the method of analyzing results. Selecting case-mix indicators and planning how they will be analyzed should be key activities in the planning process.

The planning process should include creation of operational definitions of variables, choice of instruments to measure case-mix and outcome, decisions about how and when samples will be drawn, and determination of sample sizes necessary for sufficient statistical power to obtain results. Steps of data collection should be specified. These might include the details of how a telephone survey will be conducted, including how phone numbers will be obtained, who will do the calling, how often and when calls will be made, protocols for reaching non-respondents and persons without telephones, protocols for conducting the interview, and determination of how answers will be transcribed onto data collection forms. The result of these efforts should include a data dictionary and a written data collection, management, and analysis protocol.

The data collection protocol should provide for data to be collected across and within groups. If different agencies or treatment groups are to be compared, and one agency uses face-to-face interviews to collect information, another uses mail, and a third uses telephones, comparison of results may be difficult, if not impossible. Consumers interviewed in person may be less likely to offer critical comments about their functioning or the services they received, while persons surveyed by mail would be more likely to offer such comments because of the less personal nature of the contact.

Even when surveys are conducted in a consistent format, the methods must be consistent across each group and over time. For example, if interviewers are trained to conduct the interviews by telephone, all interviewers should receive the same training, and as interviewer turn-over occurs over time, new interviewers should receive the same level and quality of training as the original interviewers. If mail surveys are done, each group should use the same procedures – the same cover letter, post-card reminders, survey paper color and font, etc.

The timing of data collection is another important consideration. Outcomes must be measured at points when they could be influenced by care. A consumer satisfaction questionnaire, for example, should be administered only after consumers have had some experience with receiving care. A functional outcomes tool should be administered at a time when it is believed that consumers could have benefited from treatment sufficiently for the tool to detect.

In some cases the variables that correlate most powerfully with outcomes may be "state" as opposed to "trait" variables, i.e. short-term fluctuations characteristic of the consumer's illness. Apparent improvement over time may reflect only the resolution of a short-term exacerbation of symptoms related to a crisis. It may be useful to collect case-mix and outcome scores on multiple occasions, to track consumer progress from early to later in treatment, as outcomes change in response to treatment as well as fluctuating state variables.

## C. Methods of Data Quality Control

Regardless of how one chooses to address quality control, the central issue is to maximize the reliability and validity of data. Data quality may be examined through a number of strategies particular to the measurement approach. Methods of quality control include the use of software technologies, audits, outside vendors, and reliance on the concept of "multiplism."

Software technologies include data collection templates that prevent invalid responses and inappropriate missing data from entering the dataset.

Independent audits of the clinical records to confirm the ratings may be important when clinicians or other providers make ratings such as diagnosis, symptoms or functioning severity. Audits require time and resources and can be costly, but states have incorporated them into their performance measurement systems (DeLiberty, Newman and Ward 2001).

Use of outside vendors is another quality control strategy to consider, but vendors offer both advantages and disadvantages and their employment must be evaluated on a case-by-case basis. Among the possible advantages, vendors normally have no stake in making one agency or another look better or worse and so can be relied on to be unbiased. Vendors may also have experience in the logistics of implementing and analyzing large numbers of surveys, and in generating quick outcome reports. This experience can relieve the mental health authority or the private administrator of the huge task of mounting and maintaining the performance measurement system.

A possible downside to relying on a vendor is the financial cost of contracting for the service. In addition, the outcome reports available from vendors may in some cases lack sophistication, clinical utility, or flexibility, or they may not have built in the appropriate case-mix adjustment technologies. Thirdly, the vendor may not provide access to raw data and proprietary analytic methods. Finally, a vendor may provide measurement tools and outcome reports but still rely on local staff for data collection, leaving this critical task in the hands of untrained or inconsistently trained persons with a personal stake in looking good. These are only hypothetical advantages and disadvantages, and organizations should evaluate each vendor on its merits.

"Multiplism" is a another strategy to protect the integrity of data. Basically, multiplism means that data will be collected from multiple independent sources and combined such that no one data source can dominate the picture. Combining measures may result in a single score mathematically averaged from several independent sources. Functioning, for example, may be measured through combining ratings made by staff, consumers, and a consumer family member. An important caveat to combining measures, however, is that it should be done only

when there is psychometric justification for it, i.e. when the different sources agree on the measurement of the concept in question. Psychometric justification requires that the measures to be combined meet conventional standards of reliability and validity supporting the combination. Where such standards are not satisfied, the individual measures should be maintained separately. In this case, the perceptions of consumers, providers, and family members may be used as three outcome measures, each offering a piece of information. In this way, the measurement provided by the provider, for example, cannot dominate the assessment of functioning.

## D. Data Transfer and Centralized Storage

The procedures for transferring and storing data should be specified. If local agencies are collecting the data, there should be schedules for when it must be transferred to a central repository and specifications for the transfer (e.g., encrypted files sent via internet or email in SAS format; with precise file layout and data dictionary codes). When a vendor is used, methods of transferring data to the vendor and of returning raw and analyzed data back to the agency must be specified.

## E. Automating Database Creation

Automating the process of entering data into the database can be valuable in increasing efficiency and reducing transcribing and data entry errors. For example, scanning survey forms into a database rather than keying in by hand will reduce errors. Software programs, such as those available in Microsoft Excel, Fox Pro, or Access can be developed to ensure consistency in data entry by means of various features. These include controls such as moving automatically from one field to the next on the screen, preventing the data entry person from skipping a variable or automatically skipping where appropriate, and not permitting out-of–range values (e.g., not permitting a value of "5" to be entered for consumer age when all consumers are adults). Mechanisms can also be created for automatic backup file creation.

## F. Data Analysis Program Options: SPSS, SAS, Excel, Access, FoxPro, Other specialized systems

SAS and SPSS are statistical analysis software programs that may be unfamiliar to some mental health providers or administrative staff, but offer flexible and powerful data analysis capacity. Vendors may have proprietary software they use or recommend, but this software may have limited flexibility, be unavailable for widespread use, produce data files that are difficult to transfer to other software packages, and make it difficult to extract or use raw data outside of the vendor's system. Microsoft Excel may be the easiest program to use, but it has limited data analysis capacity. It is possible to organize data in Excel files (or Access or FoxPro) and transfer them to SPSS, SAS, or other statistical packages for specialized analysis. The chapter on hierarchical models written by Brian Cuffel and included in this toolkit provides other examples of available software programs.

## G. Summary of Common Methods of Case-Mix Adjustment: Pre-post Scores, Stratification, Regression.

There are three common approaches to basic case-mix adjustment. The simplest is to analyze pre-post scores, in effect controlling for baseline scores in examining outcomes. Somewhat more sophisticated and more effective is stratification. The preferred method in most cases is regression analysis.

Pre-post testing consists of obtaining a single performance score at baseline (e.g., functioning) and again after a treatment period, to see if significant change has occurred or if levels of improvement are higher in some agen-

cies than in others. This method suffers from a number of problems, including 1) interpretation difficulties that may be encountered with change scores, such as the poor reliability of change scores or the assumption that change is linear; 2) the failure of this approach to control for many potentially important case-mix variables other than baseline score; and 3) the lack of relevant baseline scores for many outcomes of interest.

Stratification identifies case-mix groups of interest (e.g., men and women divided into three diagnostic groups, for a total of 6 case-mix groups), calculates the case-mix score for each group weighted by the representation of that group in the entire population, and calculates the case-mix adjusted score as the average of the weighted case-mix groups. This method is also flawed, because 1) case-mix groups may be demarcated arbitrarily, especially for case-mix measured on the basis of interval scales (e.g., how to determine appropriate "cut points" for functioning scores); 2) the calculation of weighted standard errors is a complex task and can lead to errors in interpreting final results if not done correctly; 3) the complexity of the adjustment becomes cumbersome when there are more than one or two case-mix variables to be crossed; and 4) some cells may have zero counts, which pose further computational difficulties.

A simpler approach to stratification is not to combine the scores into a weighted average, but present results separately for each subgroup, for example, presenting outcome scores separately for persons with low baseline functioning and high baseline functioning. This approach is appealing because of its simplicity, and there may be situations when it is an appropriate method. It may be a good approach if we know that only one or two case-mix variables are important, and we can agree on where to make the cuts to place consumers into one subgroup or another. It also contains the possible advantage that it has more clinical utility than a single case-mix adjusted score, because it can suggest that persons in a particular group are doing better or worse than expected and can potentially guide clinical care improvement efforts. This approach can be problematic, however, because it assumes that we know the correct case-mix variables to pick. If we were instead to select case-mix variables based on their correlation with outcomes, we might as well use regression.

Regression is by far the preferred approach, recommended for example by Iezzoni and her colleagues (1997) in the book, "Risk adjustment for Measuring Health Care Outcomes." Regression can handle multiple predictors easily and address interactions among case-mix variables in at least a limited form. Regression models can be cross-validated and coefficients applied to estimate group outcomes. Potential case-mix variables can be screened and identified using preliminary regression models. Expected group outcomes can be readily compared statistically to observed outcomes, and group performance summarized with respect to being better or worse than expected given case-mix, using a single intuitive outcome score. The examples on the accompanying CD and exercises in Part II will use regression approaches to conduct case-mix adjustment.

Reference was made earlier to the value in some cases of measuring case-mix and outcomes on multiple occasions. Under these circumstances, repeated measures analysis of covariance models can be used, which are similar to multiple regression models in the advantages they offer.

Several more advanced options to conducting case-mix adjustment are also available, including use of classification and regression trees (CART models), hierarchical regression or hierarchical analysis of covariance models, instrumental variable analysis, and propensity scoring. At the end of this chapter these four alternatives are discussed in more detail, and hierarchical modeling is also described in a separate chapter.

## H. Displaying Results

Once case-mix adjusted results are obtained they need to be displayed and communicated to stakeholders in effective ways. Graphic displays of results are often more effective than tables. It is important to provide feedback to the persons who invested time and energy in collecting the data in the first place – consumers who completed surveys, case managers who turned in interviews, administrators who provided release time to complete interviews, data managers who drew the necessary administrative data to use as case-mix variables, etc. The performance of the group must be communicated in a way that is easy for the reader to interpret. An example of displaying case-mix adjusted results using an Excel graph is presented in Part 2 of the toolkit. Different methods of displaying results can be tested and discussed among stakeholders to identify the most effective approaches.

## I. Using Case-Mix Adjusted Results To Improve Care

The greatest value of case-mix adjusted results is probably their use in improving care. Once an organization has an understanding of comparative adjusted performance, it can begin using this information to identify and learn from best performers and to identify other variables that providers and administrators can change to improve care.

If, for example, we discover that Agency X has the best case-mix adjusted hospitalization rate, we can conduct focused quality improvement studies to identify what Agency X may be doing that creates this favorable result. As another example, if we discover that superior case-mix adjusted consumer functioning is correlated with use of a particular treatment modality, we can decide whether the adoption of that treatment modality should be encouraged throughout the system. Part 2 demonstrates one strategy for identifying correlations between case-mix adjusted performance and treatment variables.

Case-mix adjustment is useful because it helps to identify possible areas of problematic or exemplary performance, not because it provides the complete answer in and of itself. A single case-mix adjusted outcome score for an agency, compared to other agencies, provides only a road sign. Case-mix adjustment by its nature "levels the playing field" by evening out the effects of different clinical, functional, or social groups. At the same time, it may be critically important to identify and respond to these differences in results.

The implications of this important observation for case-mix adjustment are two-fold. First, after overall agency performance is quantified using case-mix adjustment, we need to take the next steps to understand the underlying nature of treatment delivery at each agency to appreciate what may have led to that score. Second, case-mix adjustment should not be the only approach to examining performance. In addition, the ways in which an agency responds to important subgroups (e.g., treatment for the homeless, persons with schizophrenia, persons newly discharged from hospitalization, etc.) should be investigated separately by other quality improvement or quality management mechanisms. Case mix adjustment used alone might otherwise obscure important differences in results for these subgroups.

## J. Recalibrating The Models and Routinizing The Process

Models should be reanalyzed periodically, new regression equations found for calculating case-mix adjusted scores, new stratified outcome scores found, etc. This recalibration may be done annually or after major changes occur in variables collected, service programs, or populations served. Trial ("hold harmless") periods can be used to gain experience with the measures, data collection, and interpretation. After a trial period, case-mix adjusted

performance measures can be implemented into contracts between mental health authorities or insurers and providers. The process for doing this should be clearly understood by stakeholders, as should the interpretation of case-mix adjusted results. Given the limitations of real world mental health databases, there may be sufficient concerns that prohibit taking this step. Using the NRI Oryx indicators as a guide, sets of performance indicators may be introduced incrementally, adding one or a few each year to contract terms. As the methods come into routine use, using the results within non-punitive, quality-improvement models may offer opportunities to improve care, especially if superior performance can be tied to financial rewards.

# Part II. Exercises:

# V. Conducting a Regression-Based Case-Mix Adjustment

## A. Introduction to the CD datasets, programs and exercises

The CD that accompanies this tool-kit contains a hypothetical mental health services database. This database is included on the CD in Excel, SAS, and SPSS format. It contains 250 lines of data, each line corresponding to a person treated as an outpatient in one of five mental health treatment agencies. Please see Table 1 (p. 26) for a list of the CD data file names and contents. The CD Word file, 'data dictionary.doc' contains a description of the data including the names, coding, and content of each variable. The database follows a hypothetical group of 250 consumers beginning with a measurement at entry into outpatient treatment and after a 3-month follow up period. The dataset includes both case-mix and outcome measures. The purpose of the following exercises will be to use this database to conduct a series of case-mix adjustment analyses.

The reader is encouraged to read through the exercises, attempt to conduct the indicated steps, and then compare findings to the results provided on the CD.

## B. Exercise 1: Using SAS or SPSS to Conduct a Linear Regression Analysis

*(Note: all of the SAS programs begin with a "libname." If you choose to use a libname in your SAS programs, you will need to specify its name and the correct path to the data set on your computer.)*

**1** **Identify and analyze the performance indicator.** In this example, the performance indicator is Global Assessment of Functioning (GAF) score at 3-month follow-up (variable name: 'gaf2'). Conduct a preliminary univariate analysis of this variable (mean, median, standard deviation, N, kurtosis, range.) This step allows you to "see" the dependent variable better – to know its measures of central tendency and variability, to determine whether missing data may be a problem, and to decide whether a variable transformation should be considered if the distribution is highly skewed or has outliers. In this case, no transformation of the variable was conducted.

**2** **Identify and analyze the case-mix variables.** In this example, the case-mix variables include consumer GAF score at baseline, sex, age, prior hospitalizations, diagnosis, presence of co-occurring substance use disorder, marital status, education, and age at illness onset. These case-mix variables were built into the database and selected here because of previous research evidence that they may be associated with variation in mental health outcome or response to treatment.

Marital status and diagnosis are categorical variables and must be recoded. For purposes of this analysis, create a new variable, called 'married' that equals 1 if marital status = 1, and equals 0 if marital status = 0, 2, or 3.

Group diagnosis into three dichotomous variables coded 0 or 1: schizophrenia, major affective, and other. Schizophrenia includes all of the 295.xx diagnoses and major affective includes all of the 296.xx diagnoses. Use schizophrenia in the model and use all other diagnoses (major affective and others) as the reference category in the regression.

**3** **Identify interaction terms of interest.** Prior research suggests that diagnosis may not necessarily operate as a main effect but may interact with other case-mix variables. Although there are many possible interaction terms to test, for purposes of this exercise we will create interaction terms between schizophrenia and other case-mix variables. Create these interaction terms in the programming language in either SAS or SPSS prior to specifying the analytic procedures.

**4** **Conduct a preliminary univariate analysis of these variables.** In SAS use Proc Means for interval variables and Proc Freq for categorical variables; in SPSS use Descriptives and Frequencies, respectively. The purposes of conducting this initial descriptive analysis are to understand the characteristics of the data in general, to identify the degree of missing data that may be present, to understand the kind categorical variables that may need to be created with the most appropriate definitions, and to examine the data for outliers or skewness that may indicate the need for data recoding or transformation. Although not demonstrated here, it may also be helpful to conduct a simple bivariate correlation matrix among all variables, to identify interesting correlations or to identify possible sources of multicollinearity.

**5** **Run a preliminary regression model.** Use Proc Reg in SAS or Regression in SPSS to identify terms to retain. If the absolute value of the t statistic for any variable is < 1 (-0.99 to .99), eliminate that variable. There are alternative strategies for identifying the set of variables. It would be possible to retain all predictors, run some form of stepwise selection model, or use a different inclusion criterion (e.g., $p<.20$ or $p<.05$), but the strategy selected here is one way to eliminate variables that are clearly unrelated to the outcome variable.

**6** **Run the regression model again.** Run the model again after removing the weak predictors and check your program and output against those provided in the tool-kit. First try the programming on your own, then examine the results and check the programming steps against the program and output files provided on the CD.

**7** **Evaluate model performance.** Use the output to examine significant model terms and overall model strength as measured by $R^2$. In examining these results we can begin to see the difficulty in trying to capture the effects of these variables through something like a stratification analysis. The model suggests six main effects and five interactions that may be important to retain as case-mix variables. Clearly, a weighted stratification approach would be extremely cumbersome to attempt. Giving up on the weighting approach, we could simply select one variable and examin outcomes separately for levels of that variable. Picking baseline GAF as an example, we could split the baseline GAF scores into below average and above average and see if GAF scores at time 2 vary by treatment agency for low baseline scores and for high baseline scores. We would do this while ignoring the effects of other variables, and we would have to

make a decision about where to cut the GAF distribution. For our five treatment agencies, the GAF time 2 scores for persons with below average baseline GAF are: 32.6, 32.6, 30.0, 29.7, and 27.5. Can we conclude from this that the agencies with the scores of 32.6 are providing better services to people with lower baseline GAF scores? Of course not, because we have not accounted for other important influences, and because the time 2 differences may reflect nothing more than differences in baseline scores that our crude cut score approach failed to capture.

## C. Exercise 2: Using SAS or SPSS to Conduct a Logistic Regression Analysis

*Steps 1 through 5 will be repeated as with Exercise 1, except using logistic instead of linear regression.*

1 **Identify and analyze the performance indicator.** In this example, the indicator is whether or not the consumer was hospitalized in the three month treatment period. Find the overall hospitalization rate.

2 **Identify and analyze the case-mix variables.** In this example, the case-mix variables include consumer GAF score at baseline, sex, age, prior hospitalizations, diagnosis, presence of co-occurring substance use disorder, marital status, education, and age at illness onset.

3 **Code variables as in Exercise 1.**

4 **Code interaction terms of interest, as in Exercise 1.**

5 **Conduct a preliminary logistic regression analysis.** Eliminate case-mix variables with chi-square values < 1 (again, this is only one possible method of variable selection). Rerun the regression after eliminating these weak predictors. Use "Logistic Regression" in SPSS and "Proc Logistic" in SAS.

6 **Run and check your program.** Check the results and output against those provided in the tool-kit.

7 **Evaluate model performance.** Use the "rsquare" option in SAS.

## D. Exercise 3a and 3b: Using SAS or SPSS to Cross Validate a Linear Regression Model

Cross-validating a model offers an improvement over running the basic regression model because it results in estimates that take less advantage of chance association in the data. It requires a sample size large enough that half the sample can be selected for model identification and half for validation.

1 **Split cases in the sample in half randomly.** In SPSS use the uniform filter function and in SAS use the 'ranuni' function.

2 **Conduct the regression analysis using the first half of the data (if half=0;).** Use 'gaf2' as the dependent variable as in Exercise 1. For predictors, use the set of variables retained in the final regression analysis for Exercise 1. Steps 1 and 2 constitute Exercise 3a.

3 **For Exercise 3b, apply the regression intercept and coefficient terms to the data in the second half (if half=1;).** Steps 2 and 3 are similar to steps in doing the basic regression, with the refinement that terms identified from the first half model are applied to the second half. To do this, write a statement that creates for each person an estimated GAF2 score, called 'gaf2hat'. Make 'gaf2hat' equal to the regression intercept term plus the term for each predictor*coefficient, as in:

$$\text{gaf2hat} = \text{intercept} + B_1(\text{var}_1) + B_2(\text{var}_2) + \ldots + B_n(\text{var}_n).$$

4 **Run the regression model.** Note that the regression model to predict 'gaf2' has only one predictor: 'gaf2hat'.

5 **Run and check your program.** Check the results and output against those provided in the CD. Your results will not be exactly the same as the ones on the CD because the 'ranuni' or 'uniform' function will pull a different random half of the sample in your program versus this one. But the results, probabilistically, should be comparable.

6 **Evaluate model performance.** Cross-validated model performance can be examined with such indices as adjusted $R^2$, comparison of observed to expected variances, and the correlation between the residuals and policy relevant groups such as persons of different racial groups. Residuals should be uncorrelated with race. Construction of different 'gaf2hat' scores may be tested to identify the one that offers the highest adjusted $R^2$.

Since each person will have an observed and an expected score, the variance of those two distributions can be examined, and the ratio of the observed over the predicted variance is an f value. Models which are relatively more successful in lowering the f value, ideally to where f is non-significant, are preferred. Each predicted score also has a prediction error or a residual; these values can be correlated to race groups or other vulnerable groups, and if the correlations are non-significant, it indicates that the model is not biased with respect to race.

In the case where the model fails these performance tests, modifications should be considered in the form of adding variables, interaction terms, or sample size. Models biased with respect to race should be discarded or run and reported separately by race groups (though the latter action will reduce sample sizes and potentially create other weaknesses in the model).

Exercise 3b provides code for conducting these tests. In SAS, the "data one" step creates the estimated score and, in examining the output, we can see the adjusted r2. The "data two" step repeats the regression model, but creates an output data set called "ghat"; this output dataset includes the predicted values (p=pgaf) and the residuals (r=rgaf). Using this output data set, the predicted variances are compared to the observed variances using proc means: the f ratio of the variances is 1.68, $p<.05$ for this sample. This means that the predicted variances underestimate the observed variances, but this figure should only be compared to alternative models to see which model form can best estimate observed variance.

Use the correlation procedure in SAS or SPSS to correlate the residuals to black, Hispanic, and other race categories. Here, the residuals are uncorrelated to black, other race, and Hispanic categories. If this had not been the case (e.g., if Hispanic status was correlated to the residuals), we could compare the residuals to hispanic status on the full sample (rather than the half sample shown here). If the problem persists, we should run models separately for Hispanics if there is sufficient sample size. We should also attempt to identify additional variables that can improve prediction for Hispanic consumers, or we should introduce a strong caveat into the communication of model results to stakeholders alerting them to the fact that the model is not necessarily accurate in predicting results for Hispanic consumers.

Although race/ethnicity is used for this example, the same technique could be applied to other groups. For example, if there were concerns that the models may not be accurate for persons with a certain diagnosis, or persons with certain levels of functional impairment, etc., the residuals could be correlated to these groups.

# VI. Using Case-Mix Adjustment Regression Results to Calculate Agency or Group Level Performance Scores

## A. Exercise 4: Using SAS or SPSS to Calculate Group Performance

Use the regression model intercept and coefficient terms to predict a score for each person. Use the Exercise 1 coefficients and intercept for this instead of Exercise 3, so that you can take advantage of the full data set and compare your results directly to the CD results. The mean predicted score, or expected score, is compared to the mean observed score, at the level of the treatment agency or group, either through a difference score or a ratio, to determine whether observed group performance is statistically significantly better than, worse than, or not different from, expected group performance. More specifically:

1 **Use the regression model intercept and coefficient terms identified in Section VI, Exercise 1, to create for each person an expected (or case-mix-adjusted) outcome score.** Then, for each person find the difference between the observed score minus the expected score.

2 **Examine the distribution of difference scores at the level of each agency or group.** Aggregate each person's difference score to a mean difference score for the agency, and examine the mean and standard deviation for each agency. A mean difference score of 0 indicates that the agency's observed and expected scores are the same, that the agency performed exactly as expected. Positive scores (observed score is higher than expected) indicate that the agency performed better than expected; negative scores (observed score is lower than expected) indicate that the agency performed worse than expected.

3 **Determine whether the mean difference score for each group is significantly better than, worse than, or not different from zero.** This can be done using "Proc ttest" in SAS or the "T Test" procedure in SPSS. Include three "outcome" variables in this test: the unadjusted time 2 GAF scores, the raw time 2 minus time 1 GAF change scores, and the case-mix adjusted difference scores.

4 **Compare your program and results to the CD.**

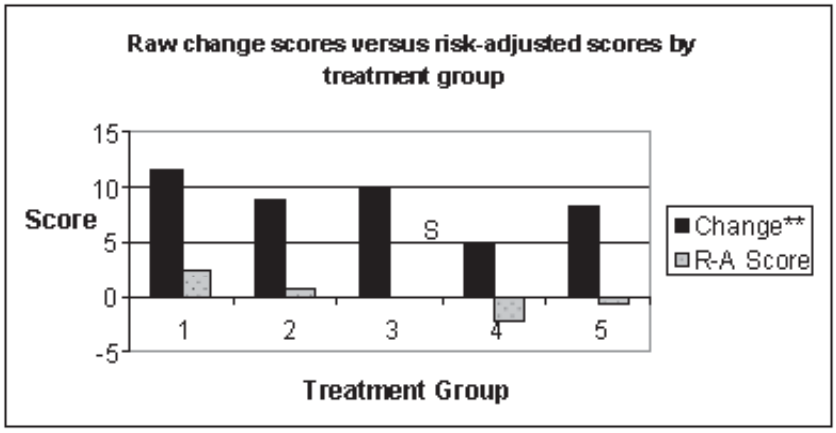
# VII. Reporting Case-Mix Adjusted Results

## A. Graphing, Tabling, Describing, and Interpreting Results

This section presents a method of displaying the case-mix adjusted scores graphically. Vertical bar charts show case-mix adjusted scores for each group, marking groups that score significantly below or above expected performance.

Using the example of GAF scores at time 2, the GAF scores themselves are not much different between groups. If we rely only on examining the change between time 1 and time 2 GAF scores without examining other case-mix variables, all five treatment groups improve significantly. But when we adjust statistically for other variables, the final results suggest that Group 1 scores 2.3 points better than expected, Group 4 scores 2.1 points worse than expected, and the other groups are not different from expected.

Excel or other programs can be used to create graphs to display the case-mix adjusted score for each group. Below is an example, taken from the Excel file on the CD, "reporting results (section VIII).xls"




# VIII. Using Case-Mix Adjusted Results to Identify Quality Improvement Opportunities

## A. Exercise 5: Identifying Quality Improvement Opportunities

As with other exercises, the reader's work can be compared to the programs and outputs provided on the CD.

*1* **Select the case-mix adjusted dependent variable.** Each person's case-mix adjusted score becomes the dependent variable or performance indicator for this analysis.

*2* **Identify variables that can be controlled or influenced by providers or the larger treatment system.** These variables, which may include treatment methods, budgets, staffing, etc., serve as the independent variables. In the practice dataset, two such variables are provided: the formal educational level of the primary care provider (Bachelor, Master, or Doctoral level), and whether or not the consumer was enrolled in a best practices care program.

3 **Conduct a regression analysis.** Conduct an analysis in SAS or SPSS to determine whether the two independent variables are related to better or worse performance.

4 **Compare your results to the Exercise 5 output file.** The results show that better case-mix adjusted GAF improvement is associated with best practices, but not with staff educational level. The findings may be interpreted from a quality improvement perspective. That is, the results may be used to identify agencies or providers who are demonstrating top performance. Those agencies or providers may be given rewards, financial or otherwise. They may also be the subject of qualitative study, such as focus groups, flowcharting, and other quality improvement tools, to identify what they are doing that leads to their success and attempt to spread their successful methods and behaviors to other groups in the system. This examination need not be limited to only those variables measured in the dataset (e.g., the best practices dichotomous variable), but can be a more general exploration of characteristics or processes that appear to promote successful outcomes in the top performing groups.

# IX. An Introduction to Advanced Alternatives to Basic Regression

## A. Decision Trees or CART Models

Classification and regression tree (CART) models, also called recursive partitioning models, are interaction-intensive models. Whereas standard regression approaches rely on identifying main effects, CART builds models iteratively based on testing interactions. Disadvantages of this procedure are that it is computer resource intensive and depends on software less familiar to most users. Its advantage is to be found in the premise that "real world" case-mix factors are also likely to be interactions. Computer programs "R" and "S-plus" provide ways to estimate these models.

## B. Hierarchical Models

Hierarchical regression models (also known as nested models or mixed models), are technically more accurate than standard regression models, but also require more advanced statistical analysis procedures. These procedures are available on SAS and on related software such as SUDAAN, and other software programs. Because consumers are "nested," or treated within agencies, consumer observations are not independent but rather correlated to each other within agency. It is more accurate to take these correlated observations into account when estimating standard errors and identifying models. Although hierarchical models are technically more accurate, the practical gain achieved by using them may be small.

A chapter on hierarchical modeling written by Brian Cuffel, Ph.D., is provided at the end of the tool-kit. This chapter illustrates the advantages of hierarchical approaches that lead to more accurate models.

## C. Propensity Weighting

This approach may be useful if there are two groups, such as two treatment groups, and a large sample. Propensity scoring is done, first, by conducting a logistic regression analysis where group is the dependent variable. The resulting distribution of probabilities is rank ordered from highest to lowest for each case and divided into quintiles. Within each quintile, a matched sample is drawn of an equal number of persons from each group (this is why the initial sample must be very large, in order to get enough persons from whichever group has the

smaller representation in each quintile). Finally, using this subsample of matched cases, a case-mix adjustment analysis is conducted in the typical way. The utility of this approach for comparative mental health case-mix adjustment seems small, since most comparisons of multiple treatment agencies or providers will involve more than two groups. Readers interested in more information about propensity scoring may refer to other published sources (e.g., Rubin 1997).

## D. Instrumental Variable Analysis

This approach also relies on regression models used in a unique way (McClellan and Newhouse, 2000). It can be used when case-mix factors are not observed. The key is to identify an instrumental variable, an observed variable that is strongly associated with group membership but *not* associated with the outcome. (Recall that in standard case-mix adjustment we identify case-mix variables that are correlated to the outcome.) An example might be the number of miles between the person's residence and the location of the treatment centers, which should be correlated with what treatment center the patient goes to, but uncorrelated with treatment outcome. The distance from each residence to each treatment center must be included in the data. The difficulty with this approach is in finding a valid instrumental variable that can be measured and is available in the data. Another possible problem is that instrumental variable analysis rests on the unlikely assumption that unmeasured case-mix variables are equal between groups.

## E. Summary of Alternatives to Regression

Propensity scores and instrumental variable analysis will not be useful under most circumstances for multigroup mental health case-mix adjustment. CART models and hierarchical models hold more promise, but their advantage to more straightforward regression modeling remains to be seen. Computer programs to run CART models and hierarchical models remain specialized and their availability is limited. As these alternatives move into more mainstream practice, however, they may compete with or even supplant traditional regression models.

# References

Banks SM, Pandiani JA, Bramley J. (2001). Approaches to risk-adjusting outcome measures applied to criminal justice involvement after community service. Journal of Behavioral Health Services and Research, 28, 235-246.

Cook TD, Campbell DT. (1979). Quasi-Experimentation: Design and Analysis Issues for Field Settings. Boston: Houghton Mifflin.

DeLiberty RN, Newman FL, Ward EO. (2001) Risk adjustment in the Hoosier Assurance Plan: Impact on Providers. Journal of Behavioral Health Services and Research, 28, 301-318.

Dow MG, Boaz TL, Thornton D. (2001). Risk adjustment of Florida mental health outcomes data: Concepts, methods, and results. Journal of Behavioral Health Services and Research, 28, 258-272.

Eddy DM. (1998). Performance measurement: Problems and solutions. Health Affairs, 17(4), 7-25.

Hendryx M., Dyck D, Srebnik D. (1999). Risk-adjusted outcome models for public mental health outpatient programs. Health Services Research, 34, 171-195.

Hendryx M, Beigel A, Doucette A. (2001). Introduction: Risk-Adjustment issues in mental health services. Journal of Behavioral Health Services and Research, 28, 225-234.

Hendryx M. and Teague G. (2001). Comparing alternative risk-adjustment models. Journal of Behavioral Health Services and Research, 28, 247-257.

Iezzoni L. (1997). The risks of risk-adjustment. JAMA, 278, 1600-1607.

Iezzoni L. (ed.) (1997). Risk Adjustment for Measuring Health Care Outcomes. Health Administration Press: Ann Arbor.

McClellan MB, Newhouse JP. (2000). Overview of the special supplement issue. Health Services Research 35(5), Part II (Instrumental Variable Analysis Applications in Health Services Research), 1061-1069.

Rubin DB. (1997). Estimating causal effects from large datasets using propensity scores. Annals of Internal Medicine, 127, 757-763.

Shadish WR, Cook TD, Campbell DT. (2001). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton-Mifflin.

## TABLE 1. CD FILES

| File Name | Content |
|---|---|
| Data dictionary.doc | Word file data dictionary for practice dataset |
| Reporting results (section VIII).xls | Excel data and graph showing risk-adjusted results |
| **Data Files:** | |
| Testfile.sav | The dataset saved as SPSS file |
| Exercise database.xls | The dataset saved as Excel file |
| Testfile.sas7bdat | The dataset saved as SAS file (version 8) |
| **Programs:** | |
| Exercise 1.sas | SAS program for exercise 1 |
| Exercise 2.sas | SAS program for exercise 2 |
| Exercise 3a.sas | SAS program for exercise 3a |
| Exercise 3b.sas | SAS program for exercise 3b |
| Exercise 4.sas | SAS program for exercise 4 |
| Exercise 5.sas | SAS program for exercise 5 |
| Syntax1.doc | SPSS program for exercise 1 |
| Syntax2.doc | SPSS program for exercise 2 (in Word) |
| Syntax3.doc | SPSS program for exercise 3a (in Word) |
| Syntax3b.doc | SPSS program for exercise 3b (in Word) |
| Syntax4.doc | SPSS program for exercise 4 (in Word) |
| Syntax5.doc | SPSS program for exercise 5 (in Word) |
| **Output files:** | |
| Exercise 1.lst | Output of SAS program for exercise 1 |
| Exercise 2.lst | Output of SAS program for exercise 2 |
| Exercise 3a.lst | Output of SAS program for exercise 3a |
| Exercise 3b.lst | Output of SAS program for exercise 3b |
| Exercise 4.lst | Output of SAS program for exercise 4 |
| Exercise 5.lst | Output of SAS program for exercise 5 |
| Output 1.doc | Output of SPSS program for exercise 1 (in Word) |
| Output 2.doc | Output of SPSS program for exercise 2 (in Word) |
| Output 3a.doc | Output of SPSS program for exercise 3a (in Word) |
| Output 3b.doc | Output of SPSS program for exercise 3b (in Word) |
| Output 4.doc | Output of SPSS program for exercise 4 (in Word) |
| Output 5.doc | Output of SPSS program for exercise 5 (in Word) |

# Appendix: Use of Hierarchical Linear Models for Case Mix Adjustment of Mental Health Provider Profile Scores

Brian J. Cuffel, Ph.D. , Vice President, Research and Evaluation, United Behavioral Health

## I. Introduction

Recent advances in statistical methods for profiling health care providers raise important concerns for mental health systems that engage in this practice. In particular, statistical methods that do not account for the hierarchical nature of profiling data may result in misleading or meaningless profile scores. Despite their applicability, hierarchical linear models (HLMs) are not widely used for profiling mental health provider performance. This paper presents the underpinnings of hierarchical linear models in simple terms and as they are applied to adjusting mental health provider performance measures for unreliability and case-mix. The application of HLMs is illustrated using some data from a large managed behavioral health organization, United Behavioral Health. Software and resources for learning about hierarchical modeling are described in the final section of this manuscript.

## II. Background

The increasing use of provider profiling has spurred the application of sophisticated statistical methods in the areas of coronary artery bypass graft surgery mortality and average visits per primary care patient (Normand et al, 1997; Christensen et al. 1997; DeLong et al. 1997; McClellan & Staiger 1999; Hofer et al. 1999). The rapidly growing literature capitalizes on computational advances in the area of hierarchical regression models sometimes called multilevel models or mixed models. Hierarchical models offer important advantages over statistical methods that are more typically used in profiling (Feinglass et al. 2000; Spoeri & Ullman, 1997). Namely, hierarchical models allow profiles to account for both patient and provider sources of variation in profile scores.

Use of hierarchical models has challenged popular notions about provider-driven variation in health care (Bindman 1999). In one recent study of primary care physicians treating diabetes, the percentage of variance accounted for by providers was so small as to make profile estimates meaningless (Hofer et al. 1999). Hofer and colleagues used hierarchical linear models to profile physicians on average number of outpatient visits per patient and other typical profile measures (Hofer et al. 1999). Hierarchical Linear Models were used to calculate the reliability of profiles using the intraclass correlation coefficient (ICC), a measure of the percentage of total variance in profile scores that is attributable to providers. The ICC for average outpatient visits was .04 indicating that only 4% of the total variation in profile scores was accounted for by differences in primary care physician practices. An ICC of .04 implies that a profile score does not achieve a reliability of .80 until caseload sizes exceed 100 patients. When ICCs are small but non-zero, profile scores may still be reliable if computed over hundreds or thousands of cases, such as when community mental health centers or group practices are the objects of profiling. In addition, analyses presented in this chapter will demonstrate that the ICCs of mental health providers may be substantially higher than those of primary care physicians, suggesting that profile scores computed on as few as 20 to 25 patients may be quite reliable.

Increasing attention to profiling methods is part of a larger movement in health care towards accountability of clinicians and systems of care. Clinicians' worst fears about such changes are that quantitative profiles of their performance will be developed and used covertly without opportunity for public scrutiny and validation. This paper discusses alternative statistical models for provider profiling of mental health data and presents results

demonstrating the influence of these models on inferences drawn about providers. Analyses compare hierarchical linear models to ordinary least squares (OLS) models on three factors critical to provider profiling: 1) variance estimation, 2) reliability of the estimated profile scores, and 3) inferences about individual providers.

Although the models developed and tested in this paper use data on outpatient visits per patient, their more important application will be to understand variability in provider practice and outcome. Results of these analyses are being used to shape provider profiling at United Behavioral Health and are being distributed to promote the open development of appropriate methods for profiling in the managed behavioral health industry. Toward this end, description of the HLM models are accompanied by the code necessary to use the SAS mixed model procedure to estimate each model and the resulting SAS output. SAS is widely used by health care organizations but the mixed model procedure is not commonly used to estimate HLMs because of some unique aspects of its specification (Singer 1999).

# III. Underpinnings of Hierarchical Linear Models

Mental health provider "report cards" or performance measurement systems attempt to compare measures of quality of care or use of services. Statistical models for provider profiling must accomplish two goals if they are to permit valid inferences about how an individual provider compares to other providers. First, a model must provide a means by which chance factors or random variation in the patients seen by a provider can be ruled out as a cause of provider score differences. Most statistical models accomplish this by estimating a patient-level error variance ($\sigma^2$) that allows a test of whether the difference between a provider's score and a benchmark normative score is different from zero.

Second, the statistical model must provide a means by which systematic differences in patient characteristics can be ruled out as a cause of the discrepancy between a given provider's score and the normative provider score, frequently known as *case-mix adjustment*. A common method of case-mix adjustment uses ordinary least squares regression (Feinglass et al. 2000; Powe et al. 1996).

## A. Simple Averaging or Ordinary Least Squares Regression

Simple averaging or ordinary least squares (OLS) regression is the most straightforward and common method of estimating profile scores. The model implied by OLS is:

$$Y_{ij} = \gamma_{00} + \alpha_j + r_{ij} \qquad (1)$$

where

$Y_{ij}$ = profile score for the ith patient and the jth provider
$\gamma_{oo}$ = a grand mean across providers
$\alpha_j$ = a dummy variable indicating the effect of the jth provider on Y
$r_{ij}$ = random variation of the ith patient from the jth provider
with $e_{ij} \sim N(0, \sigma^2)$

When $\alpha_j$ is a vector of dummy coded variables with the provider closest to the mean designated as the reference provider then $\alpha_j$ measures the deviation of the jth provider from the average provider.

OLS assumes provider effects are fixed and are not a source of random variation in Y. As a result, $\sigma^2$ from an OLS model may be a poor estimate of profile error because it assumes that observations are

independent when in fact providers may be a significant source of variation and induce correlation among patient scores. In fact, provider profiling would be unjustified in the absence of a provider effect. Hofer and colleagues demonstrate that $\alpha_j$ may be biased resulting in overestimates of the number of providers identified as atypical particularly when case load sizes are small (< 100) (Hofer et al. 1999).

## B. Hierarchical Linear Models

Hierarchical Linear Models (HLMs) yield unbiased estimates when observations are clustered within higher level groups such as when patients are clustered within providers (Normand et al. 1997; Hofer et al. 1999; Singer 1999). They accomplish this by accounting for the fact that variance in profile scores arises from two random sources: patients and providers. HLMs allow for inclusion of patient level characteristics, referred to as case-mix adjustment variables, and provider characteristics (e.g., rural versus urban provider and years of experience) that can be used to develop more refined normative comparisons. This paper compares two basic HLMs: 1) an HLM with no patient or provider covariates and 2) an HLM with selected patient covariates.

## C. Hierarchical Linear Models Without Covariates

The simplest HLM, sometimes referred to as an unconditional means model, has no patient covariates. The unconditional mean model will be presented in some detail because it illustrates how HLMs estimate variance components. The statistical model underlying HLMs is often written in two levels. In the case of profiling, the first level represents a provider level equation and the second, a patient level equation. The patient level equation expresses the score for the ith patient seen by the jth provider as the sum of two parts:

$$Y_{ij} = \beta_{0j} + r_{ij} \qquad (2a)$$

where

$Y_{ij}$ = profile score for the ith patient and the jth provider

$\beta_{0j}$ = a constant or "intercept" term for the *jth* provider. It represents the *jth* provider's performance on the profile score and is constant across all patients seen by the provider.

$\mathbf{r}_{ij}$ = a random term representing the deviation for the *ith* patient from the *jth* provider. Therefore, a patient's score is conceived of as a deviation from a provider mean.

$$\text{And } \mathbf{r}_{ij} \sim N(0, \sigma^2)$$

The provider level equation then expresses the *jth* provider's "intercept" as the sum of two parts:

$$\beta_{0j} = \gamma_0 + u_{0j} \text{ where } \mu_{oj} \sim N(0,\tau_{00}) \qquad (2b)$$

The provider specific intercept term is the sum of a grand mean ($\gamma_{oo}$) and a random deviation of the *jth* provider from that mean ($\mu_{0j}$).

Substituting (2b) into (2a) yields a single multilevel equation:

$$Y_{ij} = \gamma_{00} + \mu_{0j} + r_{ij} \qquad (3)$$

In HLM parlance, a patient's score is a function of a *fixed component*, which is the grand mean, and two *random components* corresponding to providers ($\mu_{oj}$ with associated variance component $\tau_{00}$ which tells us about variability between providers) and patients ($r_{ij}$ with associated variance component $\sigma^2$ which tell us about average variability of patients within providers). The $\mu_{oj}$ are similar to the $\alpha_j$ in (1) in that they are estimates of provider *j*'s deviation from the average provider. However, the assumption is that the $\mu_{oj}$ are a sample from a normal distribution of providers.

The variance components from a typical HLM model can be used to estimate the percentage of variance attributable to systematic differences in providers:

$$\rho = \frac{\tau_{00}}{\tau_0 + o_2} \qquad (4)$$

Otherwise known as the IntraClass Correlation Coefficient (ICC). The higher the ICC, the stronger the provider practice effects relative to the total variation in profile scores.

## D. Illustrating the Use of HLMs for Profiling Providers

Data used as a running example throughout this paper come from the Rhode Island Health Plan of United Behavioral Health. Data contain all providers with more than 20 outpatients in 1998 yielding a total of 73 psychiatrists, 49 psychologists, and 154 masters level therapists. Average outpatient visits is chosen as the profiling measure and descriptive statistics on this measure are presented in Table 1.

**Table 1. Descriptive Statistics on Outpatient Visits**

| | MD's | PhD's | MA's |
|---|---|---|---|
| Providers | 73 | 49 | 154 |
| Patients | 4980 | 1641 | 5403 |
| Mean Visits | 3.26 | 4.32 | 4.61 |
| SD Visits | 2.77 | 4.03 | 4.00 |
| Minimum visits | 1 | 1 | 1 |
| Maximum Visits | 36 | 28 | 34 |

The HLM model in equation 3 can be estimated with PROC MIXED in SAS as well as other software. The syntax for PROC MIXED is shown below:

```
Proc Mixed noclprint covtest noifprint;
Class provider;
Model Mh_visits = /solution;
Random intercept / sub = provder;
```

Where the variable *provider* is a unique identifier for the provider and *mh_visit* is the number of outpatient visits for each patient. The resulting SAS output for the sample of 73 MDs is:

Use of Mixed Models in Provider Profiling

Unconditional Mean Model

The Mixed Procedure

Model Information

| | |
|---|---|
| Data Set | WORK.TEMP2CMD |
| Dependent Variable | MH_VISIT |
| Covariance Structure | Variance Components |
| Subject Effect | PROVIDER |
| Estimation Method | REML |
| Residual Variance Method | Profile |
| Fixed Effects SE Method | Model-Based |
| Degrees of Freedom Method | Containment |

Dimensions

| | |
|---|---|
| Covariance Parameters | 2 |
| Columns in X | 1 |
| Columns in Z Per Subject | 1 |
| Subjects | 73 |
| Max Obs Per Subject | 420 |
| Observations Used | 4980 |
| Observations Not Used | 0 |
| Total Observations | 4980 |

Covariance Parameter Estimates

| Cov Parm | Subject | Estimate | Standard Error | Z Value | Pr Z |
|---|---|---|---|---|---|
| Intercept | PROVIDER | 1.4609 | 0.2783 | 5.25 | <.0001 |
| Residual | | 6.6987 | 0.1353 | 49.52 | <.0001 |

Fitting Information

| | |
|---|---|
| Res Log Likelihood | -11890.4 |
| Akaike's Information Criterion | -11892.4 |
| Schwarz's Bayesian Criterion | -11894.7 |
| -2 Res Log Likelihood | 23780.8 |

Solution for Fixed Effects

| Effect | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 3.1531 | 0.1503 | 72 | 20.98 | <.0001 |

Looking at the "Solution for Fixed Effects" section of the SAS output, the overall mean of the 73 providers is 3.1531 outpatient visits and is the ($\gamma_{oo}$) in equation 3. Note that this is not the same as the mean of the 4980 patients, which was shown to be 3.26 visits in Table 1. The variance components associated with providers and patients is shown in the "Covariance Parameter Estimates" section of the above SAS output. Provider variability ($\tau_{00}$) is estimated to be 1.4609 and is significantly ($p < .0001$) different from zero suggesting a significant provider variance component. Variability attributable to patients is considerably larger and also statistically significant ($\sigma^2 = 6.69$, $p < .0001$).

From the patient and provider variance estimates, a measure of the amount of within-provider clustering of profile scores can be obtained. Plugging the patient and provider variance estimates into equation 4, gives an ICC of .18. Later in this report, we will explain how to calculate the reliability of a profile score using the ICC that is obtained from an HLM and how to calculate the minimum number of patients required to yield a reliable provider performance score.

## *E. Hierarchical Linear Models with Case Mix Adjustment:*

Although the equations presented above result in robust variance estimates and appropriate statistical tests for individual providers, they do not include patient level covariates that are necessary to accomplish case mix adjustment. For purposes of illustration, consider the addition of two case-mix adjustment covariates to the unconditional mean model: age and diagnosis of depression. Case-mix adjustment in actual practice may contain far more covariates. However, in this simple example, the patient level equation becomes:

$$Y_{ij} = \beta_{0j} + Age\beta_{1j} + Depress\beta_{2j} + r_{ij} \tag{5}$$

The provider level equation is somewhat more complicated with the addition of patient covariates. Like the unconditional means model, a provider specific intercept is estimated along with its corresponding variance. In addition, provider specific relationships are estimated between the covariates and outpatient visits. In this example, age and depression are allowed to have a provider specific relationship with outpatient visits. This may be important if age and depression have a stronger association in some providers than others. The provider level equations include a provider intercept ($\beta_{0j}$) and age ($\beta_{1j}$) and depression ($\beta_{2j}$) slopes:

$$\beta_{0j} = \gamma_{00} + \mu_{0j} \tag{5a}$$

$$\beta_{1j} = \gamma_{10} + \mu_{1j} \tag{5b}$$

$$\beta_{2j} = \gamma_{20} + \mu_{2j} \tag{5c}$$

The usual assumptions of normality and independence regarding $r_{ij}$ and $\sigma^2$ apply. Substituting 5a,5b, and 5c into 5 results in the following multilevel equation:

$$Y_{ij} = [\gamma_{0j} + \gamma_{10}Age + \gamma_{20}Depress] + [\mu_{0j} + \mu_{1j}Age + \mu_{2j}Depress + r_{ij}] \tag{6}$$

This model has three fixed components for the intercept, age slope, and depression slope effects and four random components for the provider specific intercepts ($\mu_{0j}$), slopes($\mu_{1j}$ and $\mu_{2j}$), and for patient specific deviations ($r_{ij}$).

The SAS PROC MIXED syntax for equation 6 is:

```
proc mixed data= temp2cmd noclprint covtest;
Title "Use of Mixed Models in Provider Profiling";
Title2 "Conditional Mean Model - 2, Level-1 Predictors";
class provider;
model mh_visit= age dep_pt/solution ddfm = bw notest;
random intercept age dep_pt/sub = provider type = un;
run;
```

Where age is in years and dep_pt is a 1/0 indicator variable for whether or not the patient was diagnosed with depression in the past year. Age and dep_pt have been centered so that they have a mean of 0 in order to aid in the interpretation of the findings. The output is shown in the attached figure. Following the forward slash of the model statement, "solution" option causes the PROC MIXED output to contain the regression coefficients for all fixed effects specified in the model statement. The "ddfm=bw" option indicates that PROC MIXED should use method of computing the denominator degrees of freedom that is more efficient than the procedures default method. Notice that the random statement includes not only the intercept term ($\beta_{0j}$) but also the age ($\beta_{1j}$) and depression status ($B_{2j}$) terms. On the random statement, type = un (unstructured) tells PROC MIXED to estimate the variances of $\beta_{0j}$, $\beta_{1j}$, and $\beta_{2j}$ and all possible covariance. The variances of the intercept, age, and depression slopes are labeled UN(1,1), UN(2,2), and UN(3,3) respectively in the Covariance Parameter Estimates section of the SAS output. The intercept-age slope covariance is labeled UN(2,1), the intercept-depression slope covariance is labeled UN(3,1), and the age slope – depression slope covariance is labeled UN(3,2).

Use of Mixed Models in Provider Profiling

Conditional Mean Model - 2, Level-1 Predictors

The Mixed Procedure

Model Information

| | |
|---|---|
| Data Set | WORK.TEMP2CMD |
| Dependent Variable | MH_VISIT |
| Covariance Structure | Unstructured |
| Subject Effect | PROVIDER |
| Estimation Method | REML |
| Residual Variance Method | Profile |
| Fixed Effects SE Method | Model-Based |
| Degrees of Freedom Method | Between-Within |

Dimensions

| | |
|---|---|
| Covariance Parameters | 7 |
| Columns in X | 3 |
| Columns in Z Per Subject | 3 |
| Subjects | 73 |
| Max Obs Per Subject | 420 |
| Observations Used | 4980 |
| Observations Not Used | 0 |
| Total Observations | 4980 |

Covariance Parameter Estimates

| Cov Parm | Subject | Estimate | Standard Error | Z Value | Pr Z |
|---|---|---|---|---|---|
| UN(1,1) | PROVIDER | 1.3830 | 0.2740 | 5.05 | <.0001 |
| UN(2,1) | PROVIDER | 0.002122 | 0.006290 | 0.34 | 0.7359 |
| UN(2,2) | PROVIDER | 0.000596 | 0.000300 | 1.99 | 0.0233 |
| UN(3,1) | PROVIDER | -0.05317 | 0.1697 | -0.31 | 0.7540 |
| UN(3,2) | PROVIDER | 0.003580 | 0.004318 | 0.83 | 0.4071 |
| UN(3,3) | PROVIDER | 0.2382 | 0.1515 | 1.57 | 0.0579 |
| Residual | | 6.5456 | 0.1344 | 48.71 | <.0001 |

Fitting Information

| | |
|---|---|
| Res Log Likelihood | -11868.3 |
| Akaike's Information Criterion | -11875.3 |
| Schwarz's Bayesian Criterion | -11883.4 |
| -2 Res Log Likelihood | 23736.7 |

Null Model Likelihood Ratio Test

| DF | Chi-Square | Pr > ChiSq |
|---|---|---|
| 6 | 532.49 | <.0001 |

Solution for Fixed Effects

| Effect | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 3.1556 | 0.1486 | 72 | 21.23 | <.0001 |
| AGE | 0.01584 | 0.004849 | 4905 | 3.27 | 0.0011 |
| DEP_PT | 0.1690 | 0.1083 | 4905 | 1.56 | 0.1186 |

Examining the "Solution for Fixed Effects" section of the output gives the overall provider mean (3.1556) controlling for patient age and depression status. Age is a statistically significant predictor of outpatient visits but depression status is not.

A likelihood ratio test for the significance of the covariates can be created from the output of the two models by subtracting the −2 Log Restricted Log Likelihood Statistic (23736.7) from the model with covariates from that from the model without covariates (23780.8). The covariates appear to have significantly increased model fit ($\chi^2(2) = 44.1$, $p < .01$). However, the variance explained by the covariates is relatively small. Expressing the reduction in the residual from the first model to the second model as a percentage ([6.69 − 6.56]/6.69 * 100) yields 1.9% explained by the covariates age and dep_pt.

Again examining the "Covariance Parameter Estimates," we can see that variability in mean outpatient visits across providers, UN(1.1), remains statistically, even after case mix adjustment, suggesting that provider effects remain. Variability in the relationship between age and outpatient visits is tested by the UN(2,2) parameter and is also significant, suggesting that the age effect is not constant across providers. The same parameter for the depression-outpatient relationship approaches significance (UN(3,3)).

Of course, adequate case-mix adjustment models are considerably more complicated than two covariates. In the full case mix adjustment model tested in this paper the following covariates are included:

age (in years)

male (1 = male, 0 = otherwise)

depend (1=dependent, 0 = not)

spouse (1=spouse, 0 = not)

anx_pt (1 = anxiety diagnosis, 0 = not)

dep_pt (1 = depression diagnosis, 0 = not)

sub_pt (1 = substance use disorder, 0 = not)

sch_pt (1 = schizophrenia diagnosis, 0 = not)

adj_pt (1 = adjustment disorder, 0 = not)

bip_pt (1 = bipolar disorder, 0 = not)

per_pt (1 = personality disorder, 0 = not)

gen_pt (1 = induced by general medical condition, 0 = not)

With 12 covariates, 12 variance and 66 covariance parameters must be estimated making the model computationally very intensive. Preliminary analyses suggest that the covariance parameters were rarely significant and could be eliminated from the case-mix adjustment model. Thus, there was no evidence that the association between case mix adjusters and profile scores was greater for some providers than for others. Elimination of the covariance parameters is accomplished by eliminating the type = un option from the random statement. The resulting model contains only the 12 variance components.

In addition to the variance-covariance modeling options, the SAS mixed model procedure will yield the best linear unbiased predictions (BLUPs) of provider scores, adjusting for case-mix covariates. The BLUPs are "shrunken" estimates that bring provider estimates closer to the mean. Hofer and colleagues point out that this adjusts the provider profile measure for unreliability.[5]

## F. Standard Error of Profile Scores

Standard error estimates are necessary to test whether the difference between the jth provider and the grand mean of all providers is statistically significant. It has already been stated that HLMs result in better standard error estimates than that derived from OLS estimates.

Figure 1 plots error estimates for the 73 psychiatrists from the HLM models generated by SAS PROC MIXED against simple standard error estimates. Unadjusted HLM error estimates are linearly related to simple averaging and tend to be somewhat smaller in magnitude than for simple averaging.

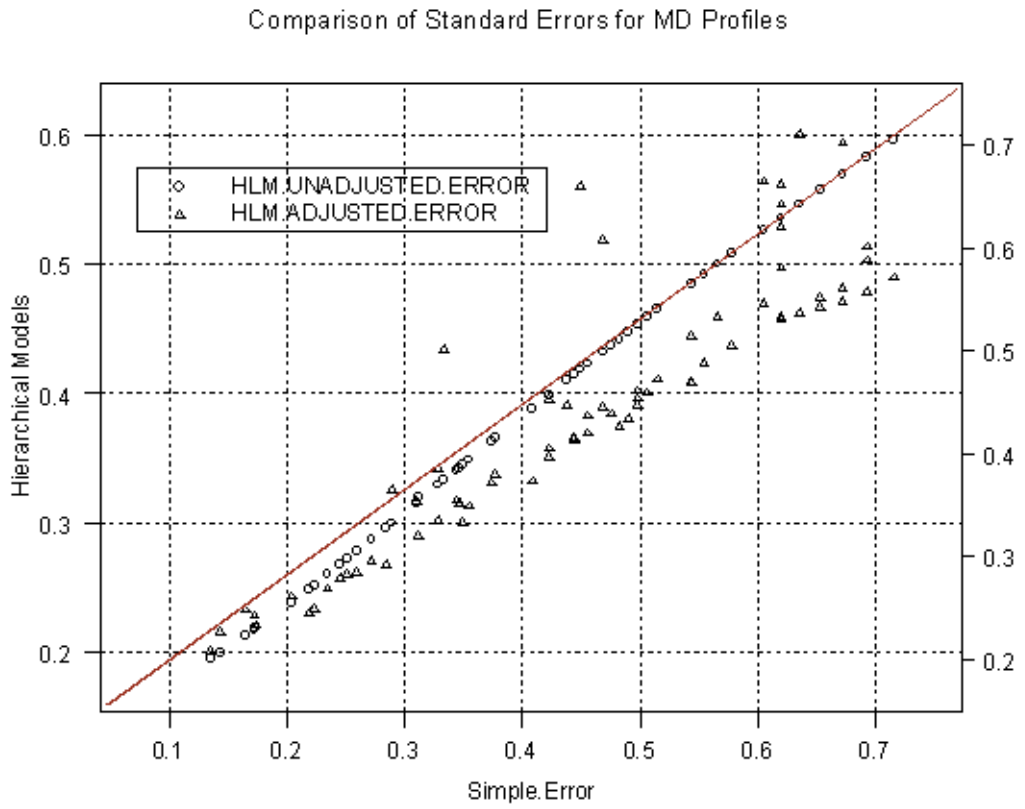



**Figure 1. Comparison of Standard errors from three statistical models**

In most cases, standard error of the HLMs is reduced, suggesting that the addition of random provider effects and covariates improve the prediction model. In rare cases, the HLM model increased provider error estimates. Figure 1 shows how the addition of covariates may affect error estimates for individual providers as well. Results also highlight the importance of model specification and the potential effects of irrelevant covariates on inferences about providers. Averaging across providers, the standard error was estimated to be .45 visits for a simple average, .41 visits for the HLM without covariates, and .44 visits for the HLM with covariates.

As expected, standard error shows an inverse relationship to the size of the providers case load for all models. Standard error as a function of number of patients per provider is shown in Figure 2.

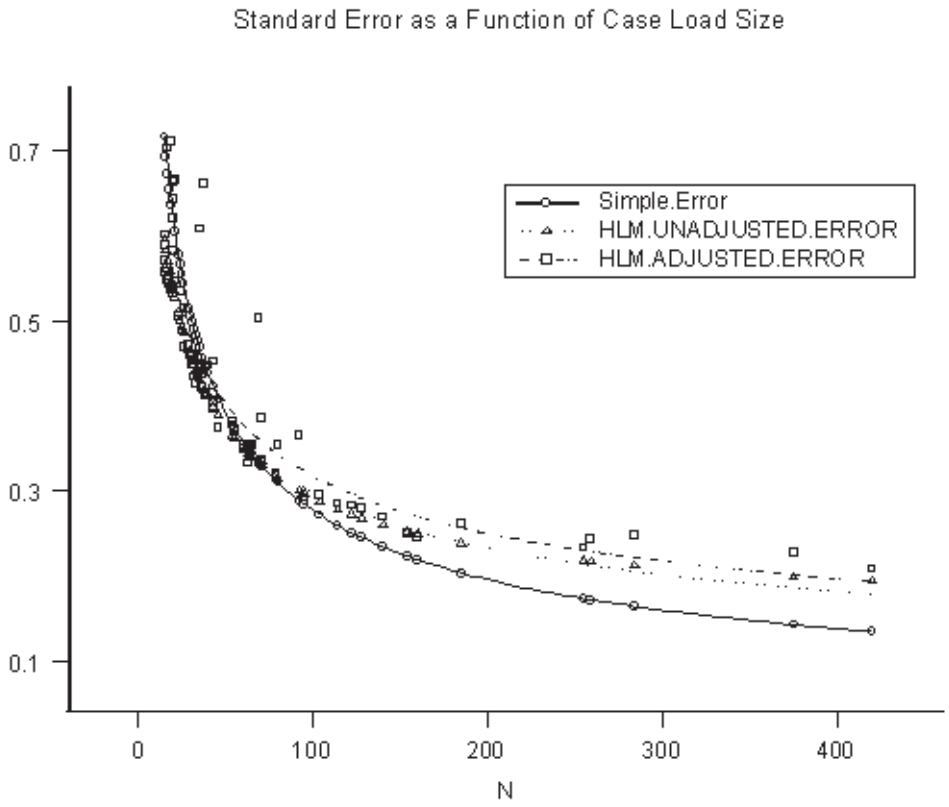



**Figure 2. Standard error as a function of caseload size.**

## G. Reliability of Provider Profiles

HLMs also permit estimates of the reliability of profiles. Hofer and colleagues (1999) demonstrated that profiles of Primary Care Physician practices were considerably unreliable, and that few providers were statistically different from the mean when case-mix and provider unreliability was accounted for by HLMs. The concerns raised by this study apply to any profiling effort in which case load sizes are small or provider effects are small relative to other sources of variation. Estimates of profile unreliability can be obtained by substituting the ICC obtained from an HLM into the Spearman-Brown Prophecy formula:

$$reliability = \frac{ICC * n}{1+(n-1) * ICC}$$

where n is the case load size of the provider. Using the data from UBH, reliability was estimated for psychiatrists, psychologists, and social workers and compared to that of Hofer and colleagues in Figure 3.

Findings suggest that profile reliability for outpatient mental health visits is higher than that found by Hofer for visits to Primary Care Physicians. ICCs for MDs, PhDs, and Masters- level therapists were .18, .12, and .14 respectively as contrasted with an ICC of .04 for visits to Primary Care Physicians treating diabetes patients. Whereas Hofer estimated that reliable physician profiles were not produced until the number of patients per

provider approached 100, data on mental health providers suggest that reliability exceed .80 when provider profile n's are between 20 and 25 patients as shown in Figure 3. Higher ICCs and reliabilities derive from the stronger provider practice effect observed in this data than in earlier research and suggest a need to understand this variability through provider profiling and other means.

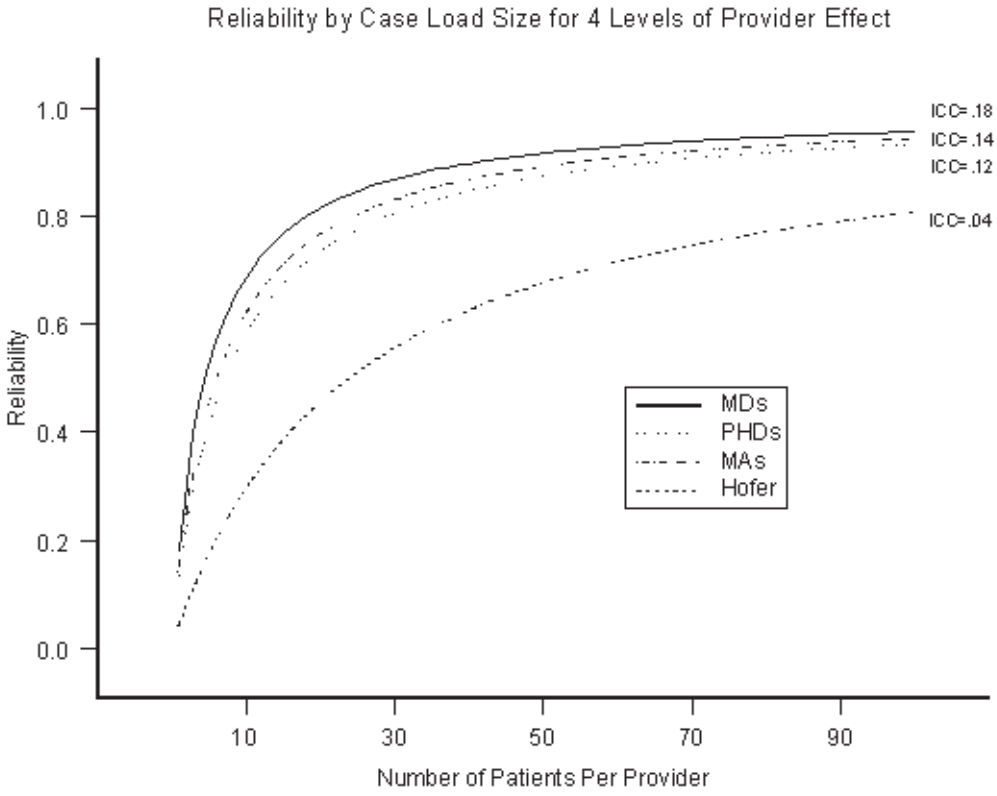



**Figure 3. Reliability of case load size for psychiatrists, psychologists, and masters-level therapists versus that for diabetes care**

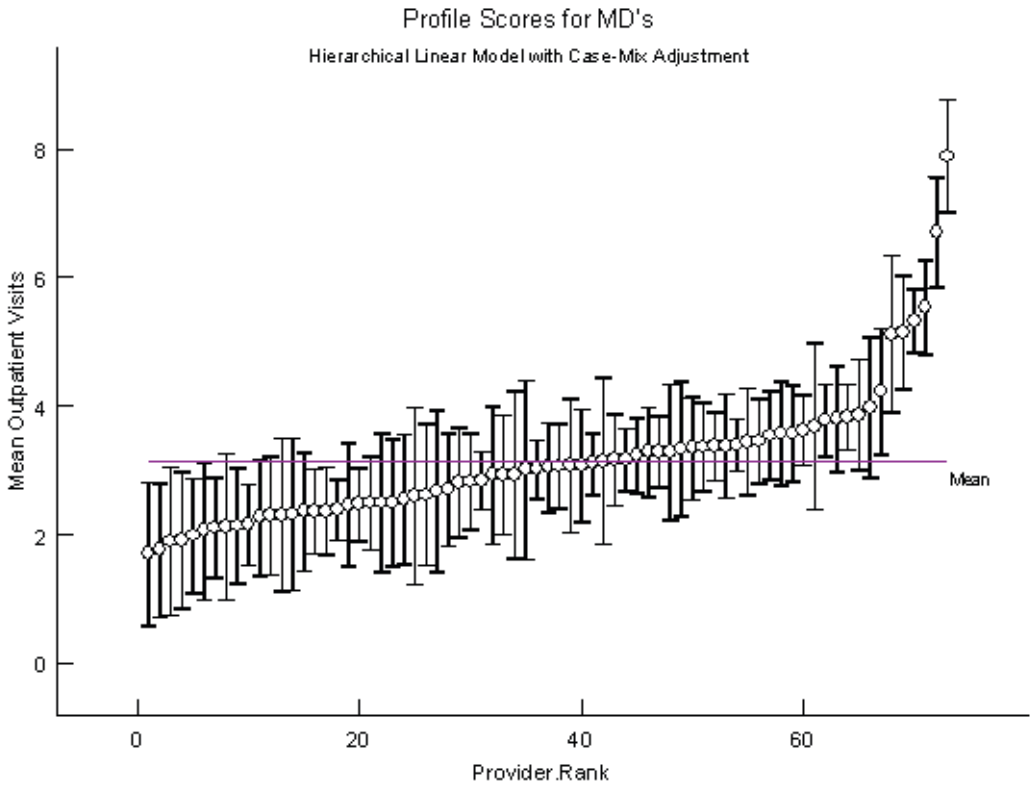



**Figure 4. Profile scores for psychiatrists**

The mean provider profile scores for the 73 MDs in this sample is shown in Figure 4. Standard error estimates are used to construct confidence intervals. Standard error estimates are multiplied by 1.96 so that confidence intervals that do not contain the mean indicate a provider who is statistically different from the mean at $p < .05$. In contrast to the findings of Hofer and colleagues, many of the physicians at the extremes are statistically different from the mean of the providers (Mean = 3.14).

# IV. Using HLMs to Make Inferences about Providers

Ultimately, systems of care that profile providers wish to draw conclusions about whether providers are different from the overall provider mean. The resulting estimates and conclusions are compared across models for the 73 psychiatrists in the sample. For each psychiatrist and each model, we use a traditional level of statistical significance to conclude whether they are "above the mean," "below the mean," or "not different" from the average provider. Agreement between models was computed using the kappa statistic.

The impact of case-mix adjustment on profile scores is shown in Figure 5. Figure 5 plots the adjusted and unadjusted HLM estimates against profile scores derived from simple averaging. The effect of adding providers as a random effect to the model on profile score estimates can be observed in the unadjusted HLM estimates. Because of the relatively higher ICCs observed in this study, most provider estimates are not dramatically affected by shrinkage due to unreliability although there is a consistent and expected trend for profile score estimates to be pulled closer to the mean. The added effect of the covariates is more dramatic as can be seen in the adjusted

HLM estimates. In some cases, the adjustment is dramatic – pulling profile scores estimates much closer to the mean.

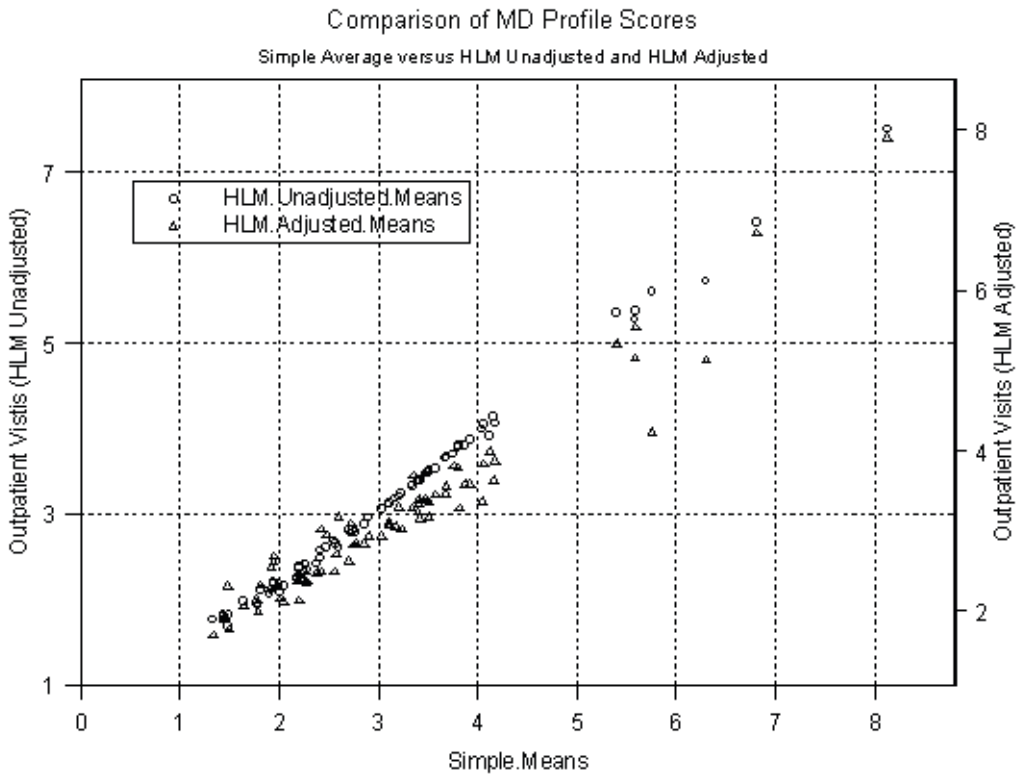



**Figure 5. Relationship of profile scores computed by different methods**

The effects of the covariates on profile score estimates suggest that conclusions about providers may vary considerably depending on the model employed. The percentage of providers identified as above, below, or not different than the average provider for each model is shown Table 3. Conclusions from the simple average model identified a total of 46% of the providers as above or below the mean. Adding covariates using OLS regression lowered the percentage of providers considered above or below the mean to 43%. Using HLMs without covariates had a similar effect, causing 43% of providers to be considered above or below the mean of their peers. However, the use of covariates in an HLM model reduced that number to 30% of providers.

Agreement between the models was measured using the kappa statistic. The effect of correcting for profile unreliability is measured by the kappas between the Simple Average and the HLM without covariates. Adjusting for profile unreliability lowers kappa from 1.0 to .88. Divergence between models is somewhat greater with the addition of covariates using OLS, which drops agreement levels to .83. Finally, as expected, adjusting for profile unreliability and covariates lowers agreement levels to .77.

Table 3. Comparison of Profile Conclusions about Providers

| | Above % (n) | Below % (n) | Not Different % (n) | Kappa |
|---|---|---|---|---|
| Simple Average | 16 (12) | 30 (22) | 53 (39) | |
| OLS with Covariates | 14 (10) | 29 (21) | 58 (42) | |
| HLM without Covariates | 18 (13) | 25 (18) | 57 (42) | |
| HLM with Covariates | 12 (9) | 18 (13) | 70 (51) | |
| Simple vs. HLM without Covariates | | | | .88 |
| Simple versus OLS with Covariates | | | | .83 |
| HLM without Covariates vs. HLM with Covariates | | | | .77 |

It should be noted that the patient covariates used in this report are rudimentary even though they illustrate the effects of case-mix adjustment with HLMs. More complex models need to consider the effect of covariates, not only on provider estimates, but also on error estimates.

# V. Resources For Hierarchical Linear Models

There are an increasing number of resources for learning about and developing HLM. Major statistical packages such as SAS, STATA, S-Plus, and R, in addition to some specialized programs, contain programs for estimating these types of models.[11] The present paper has used PROC MIXED in SAS to estimate these types of models. For binary and other non-normally distributed profile measures, PROC NLMIXED and a macro routine called GLIMMIX are available. Excellent resources for learning about HLM procedures in SAS are two articles by Judith Singer that can be downloaded from her web site at: http://gseweb.harvard.edu/%7Efaculty/singer/.

For individuals and institutions that do not license SAS software, an open source statistical software package known as "R" estimates HLM models and non-linear variants in its "NLME" package. Written in the S statistical language, R estimates many of the same type of HLM models as SAS and is freely downloadable at: http://www.r-project.org for Windows, Unix, and MAC-OS operating systems.

Finally, an enormous and growing amount of information about software, common modeling problems, and free statistical advice for "*newbies*" is available on the internet. A good starting point is the Centre for Multilevel Modelling Web site: http://multilevel.ioe.ac.uk/index.html which contains links to multilevel modeling sites, software, mailing lists, publications, and references.

# VI. Conclusions

This paper reports on the application of Hierarchical Linear Models (HLMs) to profiling mental health providers. Although methods based on simple averaging or ordinary least squares (OLS) regression are more widely used, HLMs offer important advantages. The most fundamental advantage is that HLMs account for patient and provider sources of variation leading to improved profile score and error estimates over that of OLS. Results presented in this manuscript indicate that adjustments for profile unreliability and patient characteristics significantly impact the conclusions drawn regarding individual providers.

In the present study, application of HLMs led to a number of important observations. Among the most important is that provider practice effects for outpatient mental health sessions appears to be stronger than in a previous study looking at outpatient sessions by primary care physicians for diabetes care (Hofer et al. 1999). Provider variance estimates accounted for 13-18 percent of the total variation in outpatient mental health sessions. Stronger practice effects in mental health than in other areas of health care may result from the absence of well-defined practice guidelines. Where guidelines exist for certain clinical conditions, they are rarely quantitative and specific about the appropriate length of treatment.

Another important observation is that case–mix adjustment covariates can have a dramatic effect on the conclusions reached about individual providers. The addition of case-mix variables can have two kinds of effects on provider profile statistical tests. Both effects were observed in the present study. First, case-mix adjustment can substantially affect the generated profile scores. For some providers in this study, adjusted profile scores were pulled substantially closer to the grand mean. The proportion of providers considered equivalent to the mean was 53 percent using simple averaging but increased to 70 percent when HLM models were used with a limited number of case mix covariates. In contrast to the findings by Hofer and colleagues, profiles reached a threshold of reliability (.80) at substantially lower caseload sizes.

Second, the specification of covariates influences error estimates as well. In general, error estimates were improved due to the addition of provider effects and covariates. For some providers, however, errors increased, suggesting that covariates were irrelevant to the prediction equation and degraded its performance. Although this did not occur in a large number of cases and the average error rate across providers was reduced by the addition of covariates, the finding points out the potential problem that irrelevant covariates may increase the rate at which providers with unique practice effects fail to be identified.

Generalizations of hierarchical models form the basis for other important advances in provider profiling. First, outcome variables do not need to be continuously distributed as in average outpatient visits. The SAS GLIMMIX macro can be used to estimate profile scores for nominal measures (Littell et al. 1996). Second, these models can be generalized to take advantage of repeated measures and multiple outcome measures per provider (McClellan & Staiger 1999). When these data are available, they can result in improved profile scores. Finally, hierarchical models have been used to derive empirical Bayes estimates of provider performance (Normand et al. 1997). Empirical Bayes estimators allow profiles to move away from conclusions based on traditional tests of statistical significance by allowing probabilistic statements about the likelihood that individual providers have exceeded either a normative standard of performance or a standard determined by expert consensus or guideline.

# APPENDIX REFERENCES

Bindman A. (1999) Can Physician Profiles Be Trusted? American Journal of Medicine; 281(22):2142-2143.

Christensen CL, Morris CN. (1997) Improving the statistical approach to health care provider profiling. Annals of Internal Medicine; 127(8):764-768.

De Leeuw J, Kreft IGG. (2001) Multilevel Modelling of Health Statistics. Leyland AH, Goldstein H, editors. Software for Multilevel Analysis. New York: Wiley.

DeLong ER, Peterson ED, DeLong DM, Muhlbaier LH, Hackett S, Mark DB. (1997). Comparing risk-adjustment methods for provider profiling. Statistics in Medicine; 16:2645-2664

Feinglass J, Martin GJ, Sen A. (2000). The financial effect of physician practice style on hospital resource use. Health Services Research; 26:183-205.

Hofer TP, Hayward RA, Greenfield S, Wagner E, Kaplan SH, Manning WG. (1999) The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. JAMA; 281(22):2098-2105.

Littell RC, Milliken AG, Stroup WW, Wolfinger.R.D. (1996) SAS System for Mixed Models. Cary, NC: SAS Institute Inc.

McClellan M, Staiger D. The Quality of Health Care Providers. (1999). Cambridge, MA, National Bureau of Economic Research. Working Paper 7327.

Normand SL, Glickman ME, Gatsonis CA. (1997) Statistical methods for profiling providers of medical care: issues and applications. American Statistical Association Journal; 92:803-814.

Powe NR, Weiner JP, starfield B, Stuart M, Baker A, Steinwachs DM. (1996) Systemwide provider performance in a Medicaid program. Profiling the care of patients with chronic illnesses. Medical Care; 34:798-810.

Singer JD. (1999) Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. Journal of Educational and Behavioral Statistics; 24:323-355.

Spoeri RK, Ullman R. (1997) Measuring and reporting managed care performance: lessions learned and new initiatives. Annals Internal Medicine; 127:726-732.