



Chapter 4: Managing and Analyzing Data

This chapter first discusses how to manage MHSIP Consumer-Oriented Mental Health Report Card data. This includes both MHSIP Consumer Survey and Non-survey data. It then discusses basic approaches to data analysis.

I. Data Management

This chapter will discuss data management protocol at two administrative levels. The first, the site level, refers to the operations of the organization that is actually administering the MHSIP Consumer-Oriented Mental Health Report Card such as a CMHC, a county mental health board, or an MCO. The second, the state level, refers to the operations of any organization coordinating the efforts of multiple sites or working with data collected at such sites.

This chapter will also discuss two categories of data. The first, "data requiring entry (DRE)" refers to any type of data that requires data entry. This is likely to include consumer survey items and consumer self-report items. The second type of data, "administrative data," refers to any type of data that is taken from a database that already exists. This is likely to include enrollment/encounter data, cost/expenditure data, and clinician survey/chart review data.

Site level data management

Designating a data manger

Before undertaking data management procedures any site administering the MHSIP Consumer-Oriented Mental Health Report Card should designate a single individual to act as data manager. Investing responsibility in one person facilitates organization of data and reduces the probability of errors caused by miscommunication. Although the data manager should hold primary responsibility, other people should be familiar with the database and data management procedures in case the data manager is unavailable at any time. The data manager should maintain a codebook listing all data management decisions, procedures, and operations performed. This is an excellent way to ensure consistency in the data across time and data managers.

If resources require that responsibilities be distributed among two or more individuals, it is essential that each data manager use identical techniques. They should be trained together and should meet frequently to plan data management procedures to ensure that data managed by different individuals is compatible. A codebook, such as that mentioned above, also allows different data managers to monitor work done by others and to make sure that their own work conforms.

Tracking data: Creating data logs

As an essential part of the database codebook, the data manager should create logs that delineate each step in data management procedures. There are three types of data tracking logs: 1) an "interview log", applicable only to DRE collected by survey, 2) a "data entry log", applicable to all DRE, and 3) a "data checking" log, applicable to both DRE and administrative data.

Interview Logs

The information in an interview log tracks steps in the administration of a survey and will vary according to the method of administration. For example, a column for date of interview would not be applicable for a mail-out survey administration. An example "interview log" for a face-to-face interview methodology would include the following information:

- Identifying information for the interviewee (Please refer to the section on confidentiality below)
- Contact information for the interviewee
- Back-up contact information in case of invalid address or phone number
- Identifying information for the interviewer
- Contact information for the interviewer
- Date interview scheduled
- Date interview completed
- Reason for late interview

Data entry logs

All DRE should have a "data entry log". This type of log should contain the following information:

- Date of receipt of raw data
- Data entry person or service
- Date of completion of error checking review of raw data
- Date sent for data entry
- Date of receipt of entered data
- Date of completion of error checking of data entry
- Incidence of data entry errors
- Date data entry errors corrected
- List of errors that could not be corrected
- Incidence of missing data
- Miscellaneous information for a batch of data that should be noted for future batches (e.g., a survey that should be excluded from analysis because the respondent rescinded consent)

Data checking logs

Both DRE and administrative data should have a "data checking log". This type of log should contain the following information:

- Date of completion for each electronic error checking procedure
- Incidence of errors for each electronic error checking procedure
- Date errors corrected
- List of errors that could not be corrected
- Miscellaneous information for a batch of data that should be noted for future batches

These logs will maintain an institutional memory of data management and assist the data manager in monitoring the data. They should enable the data manager to ensure that important methodological issues are being addressed such as:

- Are sufficient interviews being scheduled?
- Are interviews being completed on time?
- Are sampling procedures being reflected in the data, that is does the sample look the way it should?
- Is data entry consistently accurate?
- Is there a high incidence of errors in the data?

Data entry procedures

The first step in data management procedure is to enter any data that is in raw form. There are several options for data entry, such as a data entry person or service, data entry software, or scanner or fax software. No matter which option is used, a database structure must be created before data can be put into electronic form. The best way to approach this task is by creating a data key.

Creating a data key

Examine the document that is going to be entered. For each item that you want to be included in the database you must define the following parameters:

1. **Decide on a variable name.** In certain database software this label is required to be eight characters or less.
2. **Define the format of the variable.** Examples of variable formats are numeric (containing only numerals), string (containing characters or numbers), and date (containing a combination of characters and numbers representing a date).
3. **Define how proper responses should be coded.** For each item with response categories, each category must be assigned a value. For example, an item inquiring about sex might be coded with a value of "0" for male and a value of "1" for female.
4. **Define how improper responses should be coded.** Often times responses such as "doesn't know", "no response", "not applicable", or "refused" will appear. Values must be assigned to each of these responses to allow monitoring of the frequency with which they occur.
5. **Define the length of the variable, that is the number of characters or numbers of which it will be composed.** For items with response categories, this is simply equal to the number of characters required for each category. For open-ended questions, this is the maximum number of characters you wish to allow for a response.

In addition to the items on the document being entered, variable information should be defined for any other data of interest, such as unique identifier, date of administration, and date of entry. Writing this information on a master document next to the item to which it corresponds makes for easy reference in later stages of data management and analysis.

Key-punching data

Once a data key has been created, you are prepared for entering the raw data into electronic form. Irrespective of the data entry method that you will be using, it is advisable to enter data in small batches as it is received (assuming it is not received all at once). In addition to making the task more manageable, it is useful in preventing large-scale errors. If a mistake occurs, it can be corrected in a small batch and be averted in future batches. If all data were entered simultaneously, the problem might be more difficult to resolve.

You should first organize the raw data. If there are different DRE documents being entered (e.g., consumer surveys, family member surveys), you should group each type together. It is useful to organize each group by unique identifier so that you can navigate the raw data and the electronic data should you encounter any difficulties.

If you will be having the data entered into a database or an ASCII file (a generic file without a particular format) or entering data with a data entry program, you must provide the data entry person with information regarding how the data should be entered. A copy of the data key should be sufficient. This will show how each item on the document will be created in a database. According to the data key, the data can be entered into an ASCII file or directly into a particular database manually or with data entry software. If you have entered the data into an ASCII file, you must then read this file into a particular database format. This procedure is built into database programs and varies according to platform. Although data entry software requires that you set up entry screens and lay out the database prior to entry, this method can restrict responses to legitimate values and thereby reduce data entry errors. Both of these approaches to data entry are resource intensive as they require a person to type all the information into a computer.

An alternative to the time-consuming method of entering data manually is to use scanning software. This type of program uses a scanner or fax to read the information from the document, eliminating the need for someone to type the information into a computer. Although this method can save time, it requires more advanced technical expertise and can be prone to technical difficulties.

Once the data has been transferred into an electronic format, you should begin checking the data for errors.

Ensuring data quality

The simplest and most straightforward means of maintaining high-quality data is to thoroughly train all individuals who will be involved in data collection and data management procedures. If all data managers are familiar with protocol and use identical procedures, potential problems will be averted. Beyond proper training, several other techniques can be employed to increase the accuracy of data.

Checking for data entry errors

For DRE, data managers can perform initial checks upon receipt of raw data prior to data entry. A sample of documents to be entered should be checked to see that forms were completed and items were filled out properly. Any errors that are discovered should be corrected prior to entering the data.

Once DRE has been entered, the data manager should perform several checks to ensure that no data entry errors occurred. A list of unique identifiers in the database should be printed and

compared with the raw data to make sure that all data was entered. A random sample of the electronic data (10% of cases) should be compared with the raw data to make sure responses were coded accurately. Any errors that are detected should be corrected. The incidence of errors should be recorded to gauge the accuracy of data entry. Another method of gauging data entry errors is to enter a random sample of the raw data in duplicate by alternate means and comparing the data. Any inconsistencies should be checked with the raw data and corrected, and the incidence of errors should be recorded.

Electronic error checks

Remaining data checking procedures are conducted on the electronic data and are applicable to both DRE and administrative data. These consist of the examination of frequency distributions for range checks and outliers and checking for logically impossible combination of responses.

The frequency distribution should be examined for all variables in the database to check proper range of response, that is that all responses fall within the realm of possible answers. While rare, it is possible that data entry personnel have entered an out-of-range value, e.g., 3 for a "sex" variable that should be coded 1 or 2. In addition, sometimes respondents may enter information that is invalid, but correctly key-entered. A frequency distribution will reveal any of these out-of-range values. Whenever such errors are found, the data manager should refer to the raw data and correct the value. If the erroneous response is not attributable to data entry error and appears on the actual document, the data manager must decide whether he or she can reasonably recode it into a valid value. If not, the item should be recoded into a system missing value.

A review of the frequency distribution also enables the data manager to check for outliers or improbable values. Check for outliers or improbable values. In some cases range and logic checks may not detect values that are possible but unlikely. For example, it is possible for an individual to be 100 years old; however, it is advisable to verify such extreme data values. This type of error check is particularly useful in examining reported rates of service utilization.

Wherever possible, data should be examined for logic errors, or logically impossible combinations of responses. For example, an individual who does not receive medication should not report involuntary muscle movements resulting from the use of psychotropic medications. These errors should be corrected by referring back to the raw data. In cases where this is not feasible, decision rules must be developed for recoding the data to values that are more likely to be accurate or to code the fields as missing.

Depending on the database software being used, these checks can be automated. If the software accepts programming, syntax files can be written to conduct them. To do so, the data manager would need to create flags for each error. For example to check for outliers in age, a variable, *agecheck*, could be created and assigned a value of zero. Then, a conditional statement, such as *if age > 100 or age < 18*, could recode the variable to a value of one. This flag marks each case where an aberrant value for age appears. The same procedure can be implemented to check for out-of-range values and logic errors.

Ensuring confidentiality

It is crucial that at each stage of data management measures are taken to ensure the confidentiality of the data. The first step should be to find out what laws govern consumer confidentiality in your locality. You should also review what statements regarding confidentiality were made in proposals, disclosure agreements, and informed consent documents. Make sure that all practices conform to the laws and the stated policies. All staff should be thoroughly trained to concord with these confidentiality guidelines.

It is also important to establish policies for handling outside requests for data. Any site administering the MHSIP Consumer-Oriented Mental Health Report Card must decide whether data will be shared with others. If data will be given to other organizations, steps must be taken to ensure that this transfer of information does not breach confidentiality.

Several technical measures should be taken to ensure confidentiality.

- Internet transmissions containing any potentially identifying information should be encrypted and sent over a secure connection to an email address accessible only to individuals who need to see the data.
- Computer file containing any potentially identifying information should be saved as an encrypted file and password protected. Only individuals who require access to this information should be aware of the password.
- All raw data containing any potentially identifying information should be stored in locked file cabinets. A single individual who is aware of project staff who might require access to this information should hold the key and monitor access to the files.

State level data management

The first task that a state-level data manager must address is merging data files from different sites. This can be problematic if sites are using different database software or different variable formats in the same type of database.

Problems of database software incompatibility can be addressed by using software packages, such as DBMS Copy or Data Junction, that convert databases from one format to another. Many database programs also give the user the opportunity to save databases in a number of different formats.

Problems related to different variable formats can be more difficult to remedy. If for example, one site saves date of interview as a date format and another saves it as an eight character string format, it may require manual correction in certain database programs. This may not be avoidable with administrative data. However, these can be averted in DRE by comparing each site's data key prior to data entry. Making sure that all data files have similar structure and will therefore be compatible will greatly facilitate database merges later on. By becoming familiar with the data management techniques of each site, the data manager can also provide technical assistance to sites by sharing the experiences of other sites. Periodic on-site reviews of a random sample of sites can help the data manager to accomplish these tasks as well as checking-up on data completeness and quality at these sites.

A state-level data manager must employ some of the same techniques used by site data managers. Upon receiving data from a site, the state-level data manager should re-run error and logic check procedures on the data and inform the site of any errors that are found. This will ensure that no errors went undetected. Similarly, frequency distributions should be re-examined to identify out-of-range values, missing values, and inconsistent data. The site data manager should be informed of any aberrations that these tests reveal. Although these operations are performed at the site level, it is prudent to conduct them again. This re-examination also allows for the comparison of data from different sites, which can identify anomalies that might not be evident in a single site's data.

Missing data

If data have been entered and error checked and there are still missing responses, procedures must be developed for handling this missing data. This is extremely important, as for many statistical analyses, missing data on a single item results in the entire survey response being discarded. This could drastically reduce the number of subjects available for further analyses. Interviewers should make every effort to get respondents to answer every item. For some analyses, "not applicable" may be a legitimate response, but it cannot be included in analysis of a

scale and must therefore be assigned a missing value. The frequency of occurrence of missing values for each item should be recorded and included in any technical reports or appendices.

A frequency distribution of the items will reveal the degree to which respondents did not answer an item (often including "not-applicable", which for this discussion will be considered missing). *We recommended that only respondents who provided valid responses for more than half of the items be kept in the analysis. Whatever strategy, is employed, however, that strategy should be clearly described in the codebook.* Tables can be constructed with the numbers of respondents with high numbers of missing item responses.

For cases with items missing a response, it is possible to impute a value for the subject that has some rationale behind it. One method simply suggests entering the sample mean of those that did answer the question. Another suggests using the mean of items measuring similar content, often in the same subscale. Another more complicated procedure involves calculating a regression equation predicting a given item response from the responses to other items. Then, using the coefficients generated by the resulting equation to weight items for which there is a response, enter the predicted value for the person not answering the item. All these techniques do have consequences, however. They tend to reduce error variance and make differences appear larger than they actually might be. To improve the situation, we recommended introducing "noise" into the prediction using patterns of variation from existing data. If you choose to do this, you may wish to consult with a local expert. This limitation should also be explicitly stated in the summary reports.

It is worth emphasizing that even the most technically advance methods for handling missing data have their limitations and drawbacks. Thus, the completeness of the data collection is absolutely critical in determining the type of conclusions you will ultimately be able to draw from the data.

II. Analyzing Data

In this section we provide an overview of data analysis tasks. This overview is intended to suggest major data analysis steps you might take. Most of the statistical procedures discussed can be conducted with personal computer statistical packages. However, if you are not familiar with statistical methods, we recommend you employ a statistical consultant.

Translate items into measures

The MHSIP consumer survey consists of a number of discrete items. Before analyzing your data, you will need to decide how to translate these items into measures. Individual items can be used individually or grouped together in indices or scales. Indices or scales can be created on the basis of their content alone, or on the basis of their statistically demonstrated relationships (e.g., through factor analysis).

Scale or measure construction should not be considered only at the data management and analysis stage, but should be considered in the planning phase as well. One reason for this is that in considering measure construction, you may decide to add items to the survey. For example, in the area of cultural competency, MHSIP has very few items making index or scale construction difficult or impossible without adding items. Another reason is that you may want to know whether different stakeholder groups have interests in specific items that should then be reported individually, as well as used in scales and indices.

The options for reporting measures are:

- Single items
- Items grouped into indicators based on their content
- Items grouped into scales on the basis of statistical scaling , factoring or scaling procedures.

The advantage of **single items** is that their interpretation is more straight forward. Another advantage is that they tell us about specific issues that may be of special concern for which there are not multiple measures (e.g., self-help participation, cultural competency).

The disadvantage of single items is that they may not be as reliable. For example different raters might agree on the overall cultural competency of a program, but disagree about a single item.

The most detailed way to report the results of single items would be to simply report the frequency of each response to each item. This may lead to an overwhelming amount of data, and so you may choose to reduce the 5-response choices into positive and negative responses, or something similar. This will allow you to report the percentage of persons who responded positively to an individual item.

One advantage of items grouped into **indicators and scales** is that they are more reliable measures of the construct underlying the items.

However, a disadvantage is that you must have multiple items which leads to more burdensome, intrusive and expensive measurement instruments.

Scales based on statistical procedures are more likely to be reliable than ones based on content alone, however, the latter indicators can be reliable.

The MHSIP Consumer-Oriented Mental Health Report Card originally envisioned using survey items, single or in combination, to compute indicator scores. For example, Access Indicator #8 is "The percentage of consumers for whom services are readily available" and is computed by the percentage of persons responding positively to the following survey items:

- I was unable to get the services I thought I needed
- I was able to see a psychiatrist when I wanted to
- Staff were willing to see me as often as I felt it was necessary

This strategy exemplifies items grouped into indicators based on their content. The specific options for computing results given this option are similar but somewhat more complex to those of single items; i.e., you can report simple frequencies of the summation of responses for each item in the scale, or you can group responses into categories such as positive and negative. If you decide to categorize responses, you will need to determine values that correspond to your categories. For example, for the indicator above, the greatest possible score one could have if he or she responds "strongly agree" on each item is 15. If you decided to report the percentage of consumers responding positively on this indicator, you might decide that anyone with a score of 10 or greater has responded positively—this is equal to "I am neutral" responses to two items and "Agree" to one item.

One difficulty with grouping items into indicators based strictly on their content that has emerged through testing is that for some indicators, the items that together comprise the measure are not highly correlated and thus the reliability as measured by alpha coefficient for internal consistency is low. This suggests that the items are not tapping closely related constructs that together comprise a larger indicator, as hypothesized, but are rather discrete constructs.

The third option, items grouped into scales based on statistical procedures, has been explored in several states. The most common statistical procedure used has been factor analysis, a procedure that groups items that are highly related with each other and together reflect an underlying construct into factors. These analyses have consistently yielded a four factor solution consistent with the theorized domains of the MHSIP Consumer-Oriented Mental Health Report Card: access, quality/appropriateness, satisfaction, and outcome.

If you select the third option, you could use the factors that have been previously demonstrated to compute your scores, or you could conduct a factor analysis of your own to confirm that the structure of the survey is consistent in your locality. If you choose to use factors that have been previously demonstrated, refer to Appendix F for a list of items that correspond to the four factors. Using this list of items and factors to compute your scores will facilitate comparison across systems as more states, counties and localities implement the MHSIP Consumer-Oriented Mental Health Report Card.

Again, if you select this option you will still need to decide exactly how you want to compute your results. Your options are the same as if you grouped items based on their content discussed above.

Address validity and reliability

Validity refers to whether the measures obtained actually measure what they are intended to measure. Reliability refers to whether measures repeatedly and consistently do so. There are multiple types of validity and reliability.

In evaluation materials, data on the validity and reliability of measures is usually discussed in the instruments section, if it exists prior to an application and in the results section if it is part of the study design.

If certain types of validity and reliability have been previously demonstrated for measures that are used without modification, it is usually not necessary to repeat these tests. Reliability as measured by internal consistency has been well established at acceptable levels for the MHSIP Consumer Survey in a number of states. We are not aware of other types of empirical testing of validity and reliability for the MHSIP Consumer-Oriented Mental Health Report Card measures.

If you are using a measure for which no validity or reliability testing has been done, or if you have modified a measure, you may wish to do your own validation and reliability analysis. Below is a guide to different analysis you might undertake. Additional detail is given in the references provided.

Reliability

We recommended that each site reporting MHSIP results present their own reliability data. Even if the identical instrument is used, local differences in the way words are perceived may introduce unwanted variation. There are several types of reliability that may be calculated.

Internal consistency is usually calculated using **Cronbach's alpha**, ranging from 0 to 1. Most researchers consider 0.7 as a minimally acceptable criterion. If a measure has more than one dimension or scale, it is usual practice to calculate alpha for each scale. Some statistical programs also calculate the correlation of each item with the scale total (having removed that item from the scale). This information allows the researcher to see if individual items are much less affiliated with the scale being investigated than others. The results of the internal consistency analysis may be compared with the results from other applications (e.g., the five state study results).

A second form of reliability is **test-retest** reliability. This method assumes stability over short amounts of time of the construct being measured, although that assumption may not be valid in all cases. After the survey is given once, and sufficient time has passed that respondents will not remember their answers, but their situation will not have changed, the survey is given a second time. Measures of test-retest reliability such as correlation coefficients are then computed. The size of the correlation you expect should depend on how stable you expect the survey to be. A prudent value for a test-retest measure of association would be .6 or higher.

If non-survey items are included that reference narrative in materials such as clinical case records or, in the case of surveys, allow the respondent to write a sentence or paragraph, and those responses are to be coded so that Likert scale items correspond to those codes, it is necessary to also perform **inter-rater reliability** analyses. The most common case like this within the MHSIP Report Card is the clinician rating of the Abnormal Involuntary Movement Scale (AIMS) and the Child and Adolescent Functional Assessment Scales (CAFAS). If the criteria for each level of coding is spelled out explicitly in a scoring manual and raters are trained to use the manual, it is likely that two raters will score responses in a similar way. However, it is necessary to demonstrate that agreement with an empirical test. Often, some number of case records or surveys, minimally 25-30, are scored by two or more raters. Then, by performing an analysis similar to that used in repeated measures designs, it is possible to show that the variance due to raters is very small. As with the coefficient alpha, a statistic with range between 0 and 1 is calculated, where 1 means perfect agreement. If you do not have experience with this type of reliability testing, it may be useful to consult with a local psychometrician.

Validity

The **content validity** of the MHSIP Consumer-Oriented Mental Health Report Card and Consumer Survey has been established by including items based on the domains, concerns and indicators described in the original MHSIP task force report. In the case where a state wishes to add or delete items, it will be necessary to show the correspondence between the revised survey and the domains, concerns and indicators. In some situations, local agencies may wish to expand certain sections of the survey because of special interest in the issues. In that case, the analyses described in the reliability section above must be performed.

The validity of the MHSIP Consumer-Oriented Mental Health Report Card and Consumer Survey also might be tested in other ways. Some other commonly tested forms of validity might be examined as described below:

- The **predictive validity** of MHSIP scores might be tested by examining whether low scores predicted persons switching providers or plans or filing grievances.
- The **discriminant validity** of MHSIP scores might be tested by examining whether persons who made complaints about their care scored lower than persons who did not, or whether scores for groups of persons served by different organizations differed.
- The **convergent validity** of MHSIP scores might be tested by examining whether other performance measures (e.g., HEDIS mental health measures) correlated with MHSIP scores. For this type of test, it is important to be confident that the two performance measures are intended to measure closely related concepts.

Control for differences in populations served

If your performance measurement plan involves comparing entities such as health or behavioral health plans or substate units with each other or with benchmark data, you should control for differences in persons served. Controlling for such differences is often referred to as "risk adjustment" or "case mix adjustment." Controlling for differences in populations served will increase your confidence in attributing differences in performance indicator scores to the performance of the service provider organization or MCO. Some experts recommend only adjusting for variables that correlate with both group membership and dependent variables. We recommend controlling for any variables observed or theorized to differentiate among groups.

There are at least four major approaches to risk adjustment:

- **Subgroup analyses:** In this approach, only subgroups that are similar in theoretically indicated ways (e.g., diagnostic groups) are directly compared.
- **Regression approaches and analysis of covariance:** These approaches use regression methods to compare populations by statistically controlling for the effects of variables that distinguish groups before comparing groups scores.

- **Propensity scores:** In this method, variables that distinguish groups are statistically identified using methods like logistic regression. Then individuals are given propensity scores (literally propensity to be in one group or another) based on their values for these variables. Then scores between groups are compared only for persons with similar propensity scores.

It cannot be emphasized too strongly that comparing entities on performance measures is not valid unless some form of risk adjustment is implemented. For more detailed information on risk adjustment methods, see:

Iezzoni, L.I. (Ed.) (1994). *Risk Adjustment for Measuring Health Care Outcomes*. Health Administration Press: Ann Arbor, MI.

Rosenbaum, P.R. (1995) *Observational Studies*. Springer-Verlag: New York.

Different types of comparisons

Once results are computed, you can take one of two basic approaches, or a combination of the two, to make comparisons:

- Comparison of results across the organizations whose performance is being assessed by the performance measurement system in a relative manner,
- Comparison of results against benchmarks, derived prior to or separately from the collection of performance measurement data in this system.

Comparison of organizations to other organizations in the system

The first approach is fairly straightforward and will be familiar to most. Very simply, this approach compares the results from one organization to another in a relativistic way. This type of comparison leads typically to ranking of organizations with respect to their competitors or peers.

Comparing scores from different groups involves comparing group averages or other measures of central tendencies (between group differences) to measures of within group variability. In general, larger between group differences coupled with smaller within group variability are more meaningful. Most statistical software available for personal computers enables you to compare scores from different groups. Statistical methods for comparing groups are also discussed in most statistic texts.

It is worth reiterating here the point made above, that comparisons between groups should not be made without carefully considering inter-group comparability and employing some form of risk adjustment for any differences observed or theorized to be of importance described above.

Comparison against benchmarks

Underlying the use of benchmarks to interpret performance measurement data is the assertion that a desired, minimum, or standard level of performance can be stated prior to collection of performance measurement data. Then, performance measurement data can be compared to the benchmark—organizations that fall short of the benchmark are deemed unsatisfactory even if they compare favorably with other organizations. More complex benchmarks can include a range of acceptable values instead of a single reference point.

Benchmarks can be found or derived from many sources. No one source for benchmarks is likely to meet the needs of all indicators and measures within a performance measurement system—different sources of benchmarks may be more likely to support certain indicators.

Potential sources of benchmarks

- Humanistic values
- Public opinion
- Expert judgments
- Statistical norms
- Historical performance measurement data
- Scientific evidence of links to outcomes (i.e., for structure and process indicators)

Humanistic values

Humanistic values are most often used as benchmarks for indicators of untoward, sentinel events or critical incidents. For example, for an indicator like number of consumers who are homeless, humanistic values may place the benchmark for this indicator at zero. Another case where humanistic values may be the basis for benchmarks is for indicators about fundamental rights; e.g., from a humanistic point of view, a system's performance should be questioned until all consumers have an individual service plan.

Public opinion

Public opinion may be used as a source for a benchmark when there is a strong and broad opinion about any aspect of service system performance. For example, the public may believe that every person should have a primary care physician, thus leading to a benchmark for this indicator at 100 percent.

Expert judgements

Most benchmarks being used currently are the result of expert judgements. There is a wide range of methods for arriving at expert judgements ranging from a single expert deciding based solely on his or her experience to scientific approaches like consensus panels on which several experts assemble with the literature on a topic and through facilitated discussion come to a consensus judgement. You should take advantage of instances in which a systematic, scientific strategy has been used to derive an expert judgement on a topic found within your performance measurement system.

Statistical norms

Statistical norms are benchmarks derived from substantial testing of a measure on a population. These are sometimes found with widely used survey instruments such as the SF-36 (Ware). It is important to recognize that statistical norms will often be derived for different populations and that it only makes sense to adopt them as benchmarks in your performance measurement system is when your population is similar enough to the tested population. For example, an instrument may have norms for a healthy population and a disabled population. It would be important to understand the characteristics of the different populations to see if either would be appropriately applied to your performance measurement system.

If you know something about the comparability of your groups with the groups on which the norms are based, you can compare the values for MHSIP Consumer Survey measures. However, you must also consider the variability in the norms and your data. The more variable scores are within groups, the less meaningful any differences between the scores for those groups and norms or other groups. Most statistics text books explain how to compare scores to norms given measures of intra-group variability.

Historical Performance Measurement data

In the absence of statistical norms for many measures, historical performance measurement data can be used as benchmarks. A benchmark from historical performance measurement data can be simply how the individual organization performed in the prior year, can be the average performance across organizations in the prior year, or any other type of standard based on some organization or organizations prior performance.

Chapter 4: Recommendations

- Designate a single individual to act as a data manager
- The data manager should create logs that delineate each step in data management procedures
- A database structure must be created before data can be put into electronic form; the best way to approach this task is by creating a data key
- Enter data in small batches as it is received; this is useful in preventing large scale errors
- To maintain high-quality data, thoroughly train all individuals who will be involved in data collection and data management procedures
- Once the data is entered, the data manager should perform several checks such as examination of frequency distributions for range checks and outliers and checking for logically impossible combinations of responses
- Ensure the confidentiality of the data; for example establish policies for handling outside requests for data
- Periodic on-site reviews of a random sample of sites is recommended; this will help the data manager check on data completeness and quality
- It is recommended that only respondents who provide valid responses for more than half of the items be kept in the analysis. The strategy should be clearly described in the codebook.
- Introduce "noise" into the imputation of missing data using patterns of variation from existing data. You may wish to consult with a local expert on this task.
- We recommend that each site reporting MHSIP results present their own reliability data

- We recommend controlling for any variables observed or theorized to differentiate among groups