APPENDIX II-E

CRITERIA FOR EVALUATION OF DATA SOURCES

Timeliness

Many population-based data sets are extremely useful when comparing state data against other state and national findings. However, due to the complexity of collecting, cleaning and preparing large data sets, there is frequently a 2 to 3 year delay in the availability of these data for use by local health agencies. Thus, in order to monitor the impact on health status or outcomes of the rapid transition of Medicaid enrollees into managed care plans, local agencies may need to collect and/or analyze more timely data. For example, rather than use the state birth certificate data to monitor perinatal health, a local health jurisdiction would choose to analyze its own birth certificate data it is collecting.

Geographic Specificity

Data sets differ in their level of geographic specificity. Some data sets, such as the Federal Census of Population and Housing, contain ZIP code and census tract level information. However, population-based data are frequently collected through surveys of only a sub-sample of population, e.g., the National Hospital Discharge Survey or the National Health Interview Survey. Sample representativeness and generalizability may limit the utility of such data sets for evaluating local conditions. For example, national data set samples may not be representative of a state, regional or local population. States often address these barriers by using the standardized instrument from a national survey and sampling a much larger local population. However, additional local sampling requires additional financial resources.

Specificity of Demographic Data

National population samples are often limited, by the need for statistically valid data, to a smaller number of race/ethnic groups that have large enough numbers of people in each. For example, the size and content of the samples in the Youth Risk Behavior Survey or the High School Senior Survey on drug use do not allow results to be calculated specifically for Asian/Pacific Islander populations.

In addition, age is often reported as a categorical age group rather than a continuous variable (e.g., data on age group categories are gathered - rather than data on actual years of age). At the national level, for example, mortality reports frequently aggregate age into broad ranges such as 0 to 24 years for injuries, or 15 to 24 years for motor vehicle related indicators. This makes the age specific analysis that is necessary for program planning and evaluation more challenging, if not impossible.

As with geographic sampling, states often attempt to correct these limitations by using a standardized instrument and oversampling the groups of particular interest.

Data Consistency and Standardization

In order to adequately compare health outcome measures from one jurisdiction to another, or to monitor changes in health outcome measures over time, it is necessary to compare similar groups using standardized variables. Standardization requires that state data collection efforts use identical definitions and standard instruments where possible. This may be a challenge given that the existing data available from multiple sources frequently measures the same construct in different ways. For example, race/ethnicity may be collected and/or coded using different categories: Asians may be reported as an aggregate or by specific categories such as Japanese, Chinese, Southeast Asians, etc.

Whether a measure is recorded by self-report or "assigned" by the data collector may also lead to discrepancies across data sources.

Differences in methodologies for combining variables (or codes) into groups, may also lead to inconsistent comparisons across data sets. This frequently occurs with grouping of diagnostic codes in hospital discharge data where, for example, different researchers measure heart disease using different diagnostic ICD-9 code groupings. Similarly, age groupings are frequently not comparable across datasets.

In some cases, the data may be collected in a more detailed fashion and reported in categories that are not useful or consistent with other data being utilized. In this case, obtaining the raw data and recoding the variables may solve the problem. In other cases, the data set may simply not be helpful and a primary data collection effort will be necessary.

Availability Over Time

Some data collection efforts occur at specific intervals and are not available every year. Most notable is the US Census, which is only collected every ten years. In order to supplement this, states collect intercensal samples and make projections for each year. However, these samples may not be detailed enough to provide data at the ZIP code or census tract level for all ages or race/ethnic groups. This limits the ability to generate rates for these groups between census years. Many local health jurisdictions get population estimates through private companies that do projections for the private sector to use in marketing.

National or state survey data are also collected for specific time periods. Other surveys may be done only once due to the one-time availability of resources or the political climate. An example of this might be toxicology screening of newborns for in utero exposure to drugs or alcohol. In order to use a particular data set for ongoing monitoring, resources would have to be identified to repeat the data collection effort on an ongoing or periodic basis.

Ability to Identify Individuals or Events

Many service utilization data sources were developed for billing purposes. They contain records for encounters, admissions, or visits, as opposed to records for unique individuals and are therefore subject to duplication of information. Before deciding to use one of these data sets it is important to define the unit of analysis under consideration (billed claims,

episode of care, or individual client) and to determine if indeed this information can be derived from the data source.

Often deterministic and probabilistic record linkage strategies would have to be developed and utilized to obtain client-specific data within a billing data set and across data sets. A local jurisdiction needs to decide whether it has the resources for such an effort. A more long-term strategy would include developing a unique personal identifier using, for example, client identification numbers or a set of standardized variables.

Adequate Sample Size

Many national or state surveys that contain data only for a sample of the population are not readily applicable to smaller geographic areas or subpopulations such as smaller racial, ethnic or age groups. In some cases the data can be weighted to produce usable estimates. For example, most national surveys produced by the National Center for Health Statistics require special software to weigh records to produce national estimates. The cost of acquiring this software and the technical skill needed to use it represent barriers to the proper utilization of data sources for which weighting is required. However, in many cases there is no way to produce local estimates. In these cases, the community may decide to use a survey instrument to do primary data collection or to pay the survey developers to take a larger sample from the community of interest. Many counties have contracted with UCLA for expanded samples for the California Health Interview survey.

Sample Validity

Some survey data are collected on convenience samples. For example, the High School Senior Survey collects self-reported data on students who are enrolled and in attendance at a school on a particular day. This methodology excludes those students who have dropped out of school or who attend school sporadically from participating in this survey. Thus, it would be inaccurate, for example, to use this sample to estimate overall teen drug use since it excludes those teens more likely to have frequent, chronic or severe drug use.

Potential for Use of Instrument for Primary Data Collection

There are many indicators for which no data source exists. In these cases the decision needs to be made about whether it is feasible to develop a primary data collection instrument. An important consideration in making the decision to do primary data collection is whether there is a standardized, validated instrument that has been used in a similar population. If so, can it be acquired for a reasonable price and does the agency have the expertise and the resources to collect and analyze the data using this instrument?

In cases where there is no instrument, a key question is "Does the agency have the expertise to design an instrument or can it afford to hire a consultant to develop one?" If the answer to both is no, the planning group should probably not choose the indicator.