

Module 6: Working with Secondary Data

Common Measures Training
Chelmsford, MA
September 28, 2006

Goals of Module

- Discuss general quality issues relevant to secondary data
- Discuss general rules for when and how to use secondary data
- Identify secondary data sources and quality issues likely to arise in Common Measures project

Definition of Secondary Data (Reprise)

Secondary data are...

- Any data that are *already collected*
- Any other data *not collected with purposes of the Common Measures project in mind*

Functional definition: any data whose quality you cannot influence.

Note: Facility-submitted data can be primary or secondary

Examples: Secondary Data Sources

- Compliance databases
- Permitting or notification databases
 - E.g., air source registration
- Self-certification or inspection results from *last year's* ERP
- ERP data collected this year by a non-participating agency
- Member survey conducted by trade association
- Environmental sampling data collected for another purpose
- Other sources this project might consider?



Secondary Data and Data Quality

- Unlike primary data, little or no control over quality.
- How to handle secondary data:
 - Accept or reject it, *based on DQOs*
 - E.g., does it meet our representativeness requirement?
 - If accepted, use it properly and transparently



DQI Issues for Secondary Data

- Some common quality issues with secondary data...



Completeness

- Facility universe could be **incomplete**, or **geographically limited**
- Data based on a **different definition of the sector** could be incomplete for your purposes
- Data might only be available for a **smaller number of indicators** than desired



Bias

- **Consider the source:** From industry, or other interested party?
- **Consider the motive:** Were inspectors looking to find fault, or to gather a holistic impression of sector performance?



Representativeness

- In a compliance database, problem facilities might be **over-represented**—or **under-represented**.
- Data could range from **non-random sample** to **census**



Comparability

We look at **every DQI** through the lens of comparability as well

- No two data collection efforts are identical
- But there must be a certain amount of similarity for data sets to be compared



Evaluating Acceptability

Consider the range of today's topics, e.g.,

- Timing (How long ago? What time of year?)
- How was the universe defined?
- What indicators were used?
- How were indicator questions phrased?
- How were data gathered and processed?
- Were data quality objectives established? Were they met?



Availability of Information to Assess Acceptability

- Especially in the case of older secondary sources, you might find a **lack of information** about the circumstances under which data were collected.
- However, many regulatory data sets will have **quality plans** associated with them, which are good source of information.



Evaluating Data Sources

Find at least one DQI issue per data source:

- Compliance databases
- Permitting or notification databases (e.g., air source registration)
- Self-certification or inspection results from *last year's* ERP
- ERP data collected this year by a non-participating agency
- Member survey conducted by trade association
- Environmental sampling data collected for another purpose
- Other sources this project might consider?



For more information...

Contact Michael Crow

- E-mail: mcrow@cadmusgroup.com
- Phone: 703-247-6131


