User's Guide: Issues in Data Coding & Validation

Understanding exactly how data in a secondary dataset was collected and coded is essential. This information will usually be available in the dataset documentation, including a manual of study variables and an often-separate compendium of questionnaires given to the patient (or, for data collected by chart abstraction, an abstraction manual). In addition to determining what data was collected, it is important to know on whom the data was collected. For example, was it collected on all study participants, only on study participants in a special sub-study, only on study participants who answered "yes" to another question, and so forth.

Several approaches are useful in assessing the validity of data items of interest. First, and most important, is common sense based on a critical reading of the data collection methods. Second, many datasets borrowed data collection methods and questionnaire items from previously validated work, or conducted their own validation assessments. Description of these validity assessments is often found in the dataset documentation or is published in monographs and peer-reviewed journal articles available as links from the dataset website. Third, in some cases independent investigators have examined and published papers on the validity of dataset items. These can be found through a list of published papers provided on the dataset's website or through a Medline search. Finally, the dataset's technical help desk can often be useful in answering these questions.