Chapter XIX
Improving the Quality of Healthcare Research Data Sets

Biswadip Ghosh
Metropolitan State College of Denver, USA

ABSTRACT

The goal of many healthcare research projects and evidence based medicine programs within healthcare organizations is to support clinical care team members by mining evidence from patient outcomes to support future treatment recommendations. In these research studies, the data is often extracted from secondary sources such as patient health records, benefits systems, and other nonresearch data sources. Good data is important to facilitate a good research study and to support clinical decisions using the results. Often multiple applicable healthcare data sources are available for a research study, some of which may be internal to the organization, while others may be external, such as state or national databases. This chapter attempts to develop an understanding of how the quality of data for healthcare research data sets can be established and improved when using secondary data sources, such as clinical or benefits databases, which were created without primary intentions for research use.

INTRODUCTION

Asserting data quality is a critical component of any information systems based research project (Brodie, 1980). In healthcare organizations, research is often conducted using secondary data sources such as databases set up for patient medical records, insurance billing and benefits administration. As in other domains, data quality problems are increasing in such organizational databases (Wang, Strong and Firth, 1995; Storey and Miller, 1995). Prior studies have reported that between 50-80% of records in many such databases may be inaccurate, incomplete or ambiguous (Redman, 1998). The ramifications of using poor quality data in a research study can be quite costly –incorrect results that are flawed and unusable. Information systems built upon data sets of poor data quality with missing information and processes that allow information to bypass key managers can lead to catastrophic failures (Fisher and Kingma, 2001). To build validity in research results, data

quality must be established, particularly when using secondary data sources, which are subject to multiple recording bias (Terris, Litaker and Koroukian, 2007). This chapter addresses several techniques that are suitable for implementation via organizational processes, managerial actions and technological and statistical approaches that can help build data quality. A case study is also illustrated that uses these techniques to establish data quality for data obtained from secondary data sources, such as patient records and benefits databases for use in a healthcare research project.

This chapter applies a data product lifecycle approach towards improving data quality. This approach views data as a product, and uses a value stream for the product similar to manufacturing processes (Lee and Strong, 2003). Data quality is an overarching concept that must be addressed in the entire process starting from data collection to data storage to the end production of the data product (e.g. reports and a decision model) and its usage. Data quality cannot not be simply viewed as the aggregate quality of the data elements; rather it must also include the quality of the data collection process, the support of the access methods and ways in which users manipulate the data, the actual processing of the data and the usage of the end (information) product produced from the data. Dimensions such as efficiency and timeliness of the collection phase, the completeness and accuracy of the data items, the accessibility and security of the data and the data output products by the different users, all play a large role in the definition of data quality. This concept of a data lifecycle is extremely important in the establishment of data quality. The role of different organizational stakeholders in establishing data quality and the interventions required to ensure data quality can be better evaluated using the lifecycle approach (Lee and Strong, 2003).

A typical data lifecycle for a research study using secondary data is shown in Figure 1.

Data from multiple secondary sources are extracted, transmitted and collected (step 1) into the research database (DB). The collected data is checked for inconsistencies, incompleteness and other validity problems during the database storage phase (step 2). Rather than rejecting problematic records and causing data loss, records with inconsistencies are flagged and systems are used to clean and subsequently certify the aggregated data in the research database. Appropriate updates (step 2) are made and data quality managers then certify the data set. The data is then processed (step 3) into information products, which are output (step 4) for users to access and use. Note that each step in the lifecycle must focus on the data as well as the management of the data by

Figure 1. Typical data lifecycle