Toward High-Quality Big Data to Support Population Health: Identifying the Data Quality Problems with Medicaid Datasets

Yili Zhang, MS
Co-presenters: Pratik Tamakuwala, BE; and Gunes Koru, PhD

Abstract

Problem Statement: Big data analytics hold tremendous potentials to improve health outcomes at the population level by improving care delivery at a reduced cost. However, data quality problems are commonly encountered in research studies leveraging big data. A lack of data quality can result in imprecise, useless, or even misleading results, which detract from the quality of reports produced and decisions made to improve population health. Therefore, it becomes important to develop strategies to help improve the quality of big data in an effective and efficient manner. For this purpose, the investigation, classification and identification of the "data defects" is a necessary first step. Data defect refers to a discrepancy between the actual and expected values held by a data item that requires a corrective change. In this study, we focused on the first step to improve the quality of data stored in the Provider and Procedure Subsystems of a Medicaid Management Information Systems (MMIS). More specifically, we classified and detected the data defects in these MMIS subsystems. **Methods**: The datasets subject to defect detection consist of eleven tables for the Provider Subsystem with more than 1.5 million records, and eight tables for the Procedure Subsystem with more than 700 thousand records. The methodical steps involved reviewing all of the data-related documents, performing a descriptive analysis to better understand the data, conducting a literature review to define a taxonomy of data defects, and developing a data quality toolkit (DQT) to detect the data defects automatically and efficiently. **Results**: The taxonomy for data defects includes four major categories: Syntax violation, semantic violation, missing data, and duplicate data. These major categories are further divided into twelve subcategories. For this defect taxonomy, DQT detected more than three million data defects in the MMIS data. Fifty-nine percent of the data defects fall in to the syntax violation category and thirty-six percent of data defects fall in to the missing-data category. Most of the syntax violation defects were about the invalid values of certain Medicaid codes in the MMIS data, and the semantic violations mostly occurred due to the presence of invalid dates in the dataset. Defects related to invalid syntax should be the focus of future initiatives for data quality improvement. **Significance**: Medicaid data, a type of big data, have been utilized in various healthcare applications and population health analytics for various purposes. Examples of foci include improving the quality of myocardial infarction care, improving prescription drugs outcomes, providing a resource for epidemiologic studies, and estimating the prevalence and medical care costs for various diseases. However, substantial problems with the quality of the existing MMIS data reduce their usefulness for population health analytics purposes. Thus, effective data maintenance and cleaning become crucial to improve the quality and utility of the Medicaid data. So far, there has been no study which created a taxonomy of defects for the MMIS data and detected the data defects automatically. This research takes the first step to make the big Medicaid data an even more useful resource in population health decision making