Framework for Curating and Applying Data Elements within Continuing Use Data: A Case Study from the Durham Diabetes Coalition

Shelley A. Rusincovitch, Bryan C. Batch, MD, Susan Spratt, MD, Bradi B. Granger, PhD, Ashley A. Dunham, PhD, Lisa P. Davis, PhD, Stephanie Brinson, Jeffrey M. Ferranti, MD, MS, Howard C. Shang, Robert M. Califf, MD

Duke Medicine, Durham, North Carolina

Abstract
Data within a continuing use context (also known as secondary use) can require translation into the variables necessary for project analysis. We have developed and applied a framework in which:

1. Project objectives inform the curation of data elements.
2. Data elements are rendered into system-readable metadata.
3. Metadata are applied to the source data and used to produce data sets.

This process distinguishes between data sets and source data. Data sets contain project-specific variables that are structured for analytic activities. This can differ from source data, which may be stored in a structure dictated by the original source system for data collection, or in a data structure contrary to what is desired for analysis. Data elements mediate this translation, and the process of curation refines their definitions and associated attributes. This framework improves analysis workflow through the application of best practices, consistent processes, and centralized decision-making.

Introduction and Background
The Durham Diabetes Coalition (DDC) is a quality improvement project with the goal of improving outcomes for the diabetic population within Durham County, North Carolina. The DDC utilizes the Duke Medicine enterprise data warehouse (EDW). The EDW is populated with Electronic Health Record (EHR) and financial data from clinical encounters across our health system. Our objective was to curate a set of data elements that would be applied to the EDW in order to create data sets. In turn, these data sets are the basis for analytic activities including hypothesis development, geospatial analysis, risk algorithm development, financial modeling, and statistical analysis.

Methods
We have utilized a cross-functional team for centralized review and decision-making, with two main considerations in data element curation:

1. Refinement of the data element definition, and
2. Specific attributes associated with the data element, including representation and identification.

These criteria are rendered into system-readable metadata, which are applied in the programmatic generation of data sets. By storing the data element criteria as metadata, the data set generation process is reusable and can be utilized more efficiently on a prospective basis. Documentation is an important component of this process, and supports versioning for effective change management. In the data sets, variables within a single data domain can be derived (such as a variable to designate a diagnosis of kidney disease as indicated by any one of more than 30 ICD-9 diagnosis codes). Multiple domains can also be combined into composite variables (for example, a single diabetes variable that indicates that conditions were met within diagnosis, lab result, and/or medication data).

Results and Discussion
Our first phase of work, completed in April 2012, included the curation of 91 data elements and 4 associated metadata tables. We processed a five-year period of data from 2007 through 2011 and generated the resulting data set with a population of 244,317 patients. We anticipate that multiple data sets will be defined for different types of analytic activity. We also anticipate creating data sets with different levels of patient identifying information (PHI) to meet the HIPAA minimally-necessary standard for PHI data disclosure. All data sets will be derived from the same source of data elements and retain consistency of definition.