Data Standards ≠ Data Quality

Meredith Nahm and W. Ed Hammond
Duke Center for Health Informatics, Duke University, Durham, NC, USA

Abstract

The relationship between data quality and data standards has not been clearly articulated. While some informaticians directly state that data standards increase data quality, others describe situations where this is not the case. Depending on the type of data standard and the impact of data quality considerations, both arguments may in fact be correct. We apply a dimensional definition of data quality to clearly articulate the relationship between data standards and data quality, and provide a framework for data quality planning.

Introduction

Unfortunately, the language we use to talk about data quality and data standards lacks specificity. We posit that each impacts data quality differently. Quality likewise is an imprecise term. There are at least two world views from which quality is approached: 1) quality as conformance to specifications (Crosby) vs. 2) quality as fitness for use (Deming, Juran). While these need not be different, in the absence of a clear definition of what quality means for a particular deliverable, significant expectation differences between suppliers and customers can result. Lastly, there are different perspectives of what data quality actually is.

We explore each of these in depth and use the explicit treatment of each to provide a framework describing the relationship between data standards and data quality.

Types of Data Standards

By deconstructing 1) the types of data standards according to the aspects of data or reality represented and 2) the dimensions of data quality, the inequality of data standards and data quality becomes apparent. With this approach, we can examine a data standard and understand as well as impact on data quality.

Case 1

A quick examination of the data in the Admission Discharge Transfer (ADT) messages at our institution reveals strings on the following:

999999999
999-9999
NO PHONE

Registration information is used in many ways; one is the ADT message that among other things is used to notify relevant parts of the facility of a new patient. The patient identification (PID) section of the standard contains a same phone number attribute (phone number—home, data type XTN [extended telecommunication number]); it is not required, and the standard doesn’t enforce other constraints. For data exchange and notification within the facility, the data were of fine quality. For other users like patient contact, use in entity resolution, billing, and administration, the data were of poor quality, and had a high inaccuracy rate. The ADT message is one of the most relied upon and fundamental standards in health care—thus it has had a large positive impact on health care. The ADT message, however, was not developed to address data accuracy, content and format for exchange; therefore, any expectation that the ADT standard would improve the accuracy dimension of data quality would be misguided.

Types of Data Quality

Data quality has been defined by the information quality community as a multidimensional concept. Some dimensions such as accuracy are properties of individual data values and are context independent, while others, e.g., relevance, are dependent on the context of use. Data quality dimensions are used to define and articulate expectations of data, and fit with both “fitness for use” and “conformance to specification” perspectives of quality. When necessary and sufficient dimensions have been articulated and paired with acceptance criteria, the gap between the “fitness for use” and “conformance to specification” perspectives diminishes.

Data quality dimensions have been articulated for regulatory decision-making in clinical trials for marketed therapeutics: Accurate, Legible, Contemporaneous, Original, and Attributable, referred to as the ALCODA criteria. Some dimensions have not yet been articulated for health care contexts. However, that use of data by someone other than who collected them, i.e., secondary use, necessitates additional dimensions such as unambiguous definition, as well as context of data collection. Until necessary and sufficient definitions of data quality for health care have been articulated, the term “data quality” with respect to health care data has little meaning. We posit that for clinical decision-making, five attributes are important: 1) to have accuracy within some wide acceptance criteria, 2) to come from a trusted origin, 3) to be reliably assessed/documented for similar patients in routine care, 4) to have consistent meaning across health care, and 5) to be immediately accessible. If these are the important dimensions, then in a case where the data were accurate, trusted, and reliable but not accessible, a clinician could certainly say the data were of low quality. Unfortunately, standard data elements specifying definition and measurement method may impact dimensions 1 and 3, but probably not 2 and 4. Different types of standards impact data quality in different ways/dimensions.

References


Acknowledgment

This work was supported by National Library of Medicine Grant K23LM011128-01A1. The responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

Case 2

In the recent Institute of Medicine meeting on Sharing Clinical Research Data, Dr. Seyfert-Margolis reviewed a case study in which data from 101 Impact New Drug Applications (INDAs) were converted to the Clinical Data Interchange Standards Consortium (CDISC) Submission Data Tabulation Model (SDTM). She described a situation in which conversion to the SDTM standard resulted in reduction of the information content. She presented an example of multi-racial patients and said, “We saw a lot of interesting stuff around racial categorization.” The CDISC SDTM standard follows the Office of Management and Budget (OMB) five-category race and two-category ethnicity standard. Dr. Seyfert-Margolis described mapping of data from mixed-race individuals. The person would be coded as one of the other, and the remaining information would be placed in another column in “an additional information field which became quite complicated when we started looking through it.” She explained that “even though it was standardized it didn’t ensure quality because those categories were not capturing the actual racial identity of the patients,” and she made the point that conversion to a standard has the potential to adversely impact data quality and accuracy. She further emphasized that in the case of mapping data to a standard for an intended use data, the scientific questions of the intended use drive the details of the conversion. The CDISC SDTM standard is an information model with an associated standard terminology. Therapeutic area-specific data elements are not defined; thus, especially in the more normalized SDTM domains, the semantics are underspecified. So, while the standard most definitely improves consistent formatting of data for regulatory submission, it has no mechanism to affect accuracy and, through information reduction, may in fact decrease the ability of the reported data to accurately reflect the actual state of the patient at the time of the study. It is only through considering the aspects standardized and the dimensions of data quality of impact that we are able to clearly articulate what the standard will and will not impact.