Name:	Ajit Londhe
Affiliation:	Janssen Research & Development
Email:	alondhe2@its.jnj.com
Presentation type(s):	**Poster**

The Impact of Data Quality Annotations on Observational Data Research

Ajit A. Londhe, MPH [1,2], Vojtech Huser, MD, PhD [2,3], Erica A. Voss, MPH [1,2]
[1] Janssen Research & Development, Titusville, NJ, [2] Observational Health Data Sciences and Informatics, New York, NY, [3] National Institute of Health, Bethesda, MD

Abstract

Research based upon observational data often requires deep understanding of the source data, temporal trends in source code vocabularies, changes in source data collection processes, and the transformative steps executed by the data custodian prior to usage. At present, this understanding is not routinely stored in a formalized manner, which can make uninformed study design possible. A metadata repository to augment the OMOP CDM could support the storage and retrieval of human-authored annotations. These annotations would include extract-transform-load (ETL) choices, vendor-supplied data collection details, observations from data set characterization, and experiences from researchers; all of which would ensure that future studies leveraging the data set adjust for known limitations or nuances therein. This poster will examine case studies to explore the practical value of leveraging a metadata repository to store annotations, and propose a set of best practices for OHDSI sites to implement.

Introduction

Observational data are often adapted from transactional systems into secondary data sets (1), and are governed by the source country's laws, policies, and industry standards. Consequently, these data sets can be rife with data quality issues and contextual nuance unbeknownst to novice users, and within OHDSI are rarely catalogued by data set experts in a formalized manner. To mitigate the impact of data quality issues or nuances, effective annotations of a data set can provide necessary context to a researcher. In data research, an annotation should be "an intentional and topical value-adding note linked to an extant information object" that helps explain "structure, function, location, and provenance" (2). Two case studies, utilizing Optum Clinformatics™ DataMart (OPTUM) as an example of commercially available administrative claims data, demonstrate the need for OHDSI sites to 1) enact annotation of data anomalies or ETL choices as standard practice to prevent avoidable study design mistakes and 2) adopt a metadata repository standard to hold these annotations.

Case Study 1: Social Security Administration Death Master File

With the enactment of Section 205(r), the Social Security Administration (SSA) limited the distribution of the full Death Master File (DMF) to only government agencies beginning in November 2011 (3). In OPTUM, the impact of this change was significant, as it originally sourced death information from both the DMF and the National Death Index. Prior to the change, the incidence of confirmed death status was as high as 1.6 records per 1000 patients, which then dropped to about 0.4 records per 1000 patients. While the reduced death coverage is easily identifiable through the OHDSI tool Achilles, the cause of the drop is not available as a resource. Table 1 shows how the metadata repository standard proposed by Huser et al. (4) could store a human-authored annotation highlighting the DMF data loss. This metadata could be exposed to users in the OHDSI tool Atlas (Figure 1 and Figure 3), suggesting adjustments to study parameters may be necessary.

Case Study 2: ICD9CM to ICD10CM Migration

The overhaul of diagnosis claims submissions in the United States to switch from ICD9CM to ICD10CM began in October 2015. To a US claims novice, the major drop in prevalence of conditions like "malaise and fatigue" (concept id 439926) around this event could cause confusion or be neglected altogether.

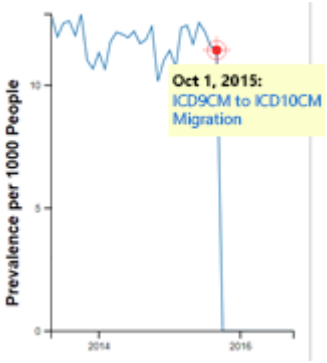
However, with appropriate context catalogued (Table 1), any OPTUM user that examines the annotations could appreciate the prevalence drop (Figure 2 and Figure 4), and could then attempt to design their concept set to include standard concepts with corresponding ICD10CM codes as descendants.

Table 1: Proposed Metadata Table with Annotations. Uses the metadata table proposal from Huser, et al. (4) to store human-authored annotations.

metadata concept id	metadata type concept id	name	value as string	value as concept id	metadata datetime	metadata date
44819056	1	Death	Loss of access to Social Security Administration Death Master File	NULL	2011-11-01	2017-08-23
439926	19	Malaise and fatigue	ICD9CM to ICD10CM migration	NULL	2015-10-01	2017-08-23

Figure 1: Mockup of Death Prevalence before and after SSA Section 205(r)

Figure 2: Mockup of "Malaise and Fatigue" Prevalence



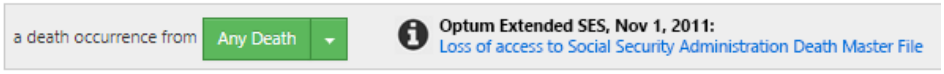


Figure 3: Mockup of Atlas Cohort Designer with Death Annotation

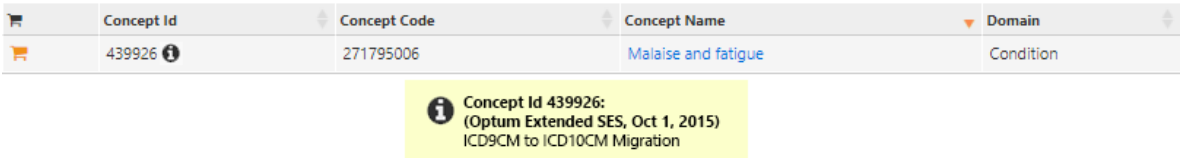


Figure 4: Mockup of Atlas Concept Set Designer with Condition Annotation

Conclusion

In both case studies, major shifts in data prevalence are visible in Achilles, but the circumstances around their existence are not immediately clear to novice users. Human annotation allows expression of this information. We recommend that OHDSI sites adopt annotations as standard practice and store them in the proposed metadata repository to help researchers avoid flawed study design, particularly when conducting studies against multiple CDM data sets.

References

1. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. eGEMs. 2016;4(1).
2. John Macmullen W. Annotation as process, thing, and knowledge: Multi-domain studies of structured data annotation. in ASIST Annual Meeting, (Charlotte, NC, 2005), ASIST, in review; 20052005.
3. Requesting The Full Death Master File (DMF): Social Security Administration; 2013 [updated 2013. Available from: https://www.ssa.gov/dataexchange/request_dmf.html.
4. Huser V, Londhe A, Voss E. Metadata Proposal GitHub2017 [Available from: https://github.com/OHDSI/CommonDataModel/issues/79.