BACKGROUND
Deficiencies in Observational Data Context
- Observational patient data often entails significant adaptation and abstraction from transactional systems (1) to adhere to the source country’s privacy laws, information security policies, and industry standards.
- The transformative nature of these preparation steps is rarely fully captured in vendor-provided documents.
- Users need to discover these nuances through review and hands-on utilization, resulting in a high likelihood of underhandled quality issues and undocumented contextual nuances that can bias study results.
- Critical information is left unknown to novice users or forgotten by seasoned researchers.

Metadata and Annotations
- Metadata refers to “the information we create, store, and share to describe things” (1).
- An annotation is a type of metadata in which an “intentional and topical value-adding note” is tagged to a data element that helps explain “structure, function, location, and provenance” (2).
- Annotations can be made on multiple levels (data set, domain, event concept id)
- Two types of annotations:
  - Structured metadata that can be programmatically derived.
  - Unstructured metadata that are best understood by data analysts.

Current Status in Observational Health Data Sciences and Informatics (OHDSI)
- No formalized construct to store metadata
- Users need to discover these nuances through review and hands-on utilization, resulting in a high likelihood of underhandled quality issues and undocumented contextual nuances that can bias study results.

CASE STUDIES
Two case studies, utilizing OptumClininformatics® DataMart (OPTUM) claims, demonstrate the need for OHDSI sites to:
1. Adopt the forthcoming metadata repository standard in the OMOP CDM.
2. Enact annotation of data anomalies or extract, transform, & load (ETL) choices as standard practice to prevent avoidable study design mistakes.
3. Consider an “interventional” annotation table to store suggestions on how to handle data anomalies once identified.

CASE STUDY 1: Social Security Death Master File
- OPTUM sources death events from the Social Security Administration (SSA)’s Death Master File (DMF), which consisted of death records from both national- and state-level systems.
- In November 2011, the SSA stopped including death information whose source was solely state-level records (A).
- Before this change, the incidence of confirmed death status in OPTUM was as high as 1.6 records per 1000 patients, which then dropped to about 0.4 records per 1000 patients (Figure 1).

CASE STUDY 2: ICD9CM to ICD10CM Migration
- The overhaul of diagnosis claims in the US to switch from ICD9CM to ICD10CM began in October 2015.
- Major drop in prevalence of conditions like “malaise and fatigue” (concept id 439926) could cause confusion or be neglected altogether in concept set / cohort design (Figure 2).

PHASES OF ANNOTATION IMPLEMENTATION
To prevent the case studies illustrated from going undetected 3 phases of annotation implementation are required:

**PHASE 1: Formalization of a system that identifies notable data elements**
- Both automated processes and manual observations can provide annotation candidates.
- Trend anomalies are identifiable using Achilles (Figures 1 and 2).
- The development of an algorithm that highlights anomalies to data custodians would ensure that all trend-related annotation opportunities are identified.

**PHASE 2: Adopting standards for annotation storage**
- The CDM Metadata table can suitably store annotations from the case study examples (Table 1).

**PHASE 3: Utilization of annotations in applications**
- Given the cataloguing of notable events and suggestions on how to address them, the opportunity to guide Atlas users away from avoidable design flaws becomes possible (Figures 3 and 4).

CONCLUSIONS
- 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, nor are solutions available on how to handle their presence.
- Human- and algorithm-generated annotation allows expression of this information in applications like Atlas.
- We recommend that OHDSI sites adopt both observational and interventional annotations as standard practice and store them in the upcoming metadata repository to help researchers avoid flawed study design, particularly when conducting studies against multiple CDM data sets.

CONFLICT OF INTEREST STATEMENT
Ajit A. Londhe and Erica A. Voss are full time employees of Janssen Research and Development, a unit of Johnson and Johnson. The work on this study was part of their employment. They also hold pension rights from the company and own stock and stock options.

ACKNOWLEDGEMENTS
Vojtech Huser’s work was supported by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ National Cancer Institute National Center for Biomedical Communications (NCBC).