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, LLC, Titusville, NJ, [2]Observational Health Data Sciences and Informatics (OHDSI), New York, NY, [3]National Institute of Health, Bethesda, MD

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 unhandled data quality issues and undocumented contextual nuance 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** in the OMOP Common Data Model (CDM) as of version 5.2.0.
- A recent metadata storage proposal has been approved for use in an upcoming release of the OMOP CDM (3).

CASE STUDIES

Two case studies, utilizing Optum Clinformatics® 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 (4).
- 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).

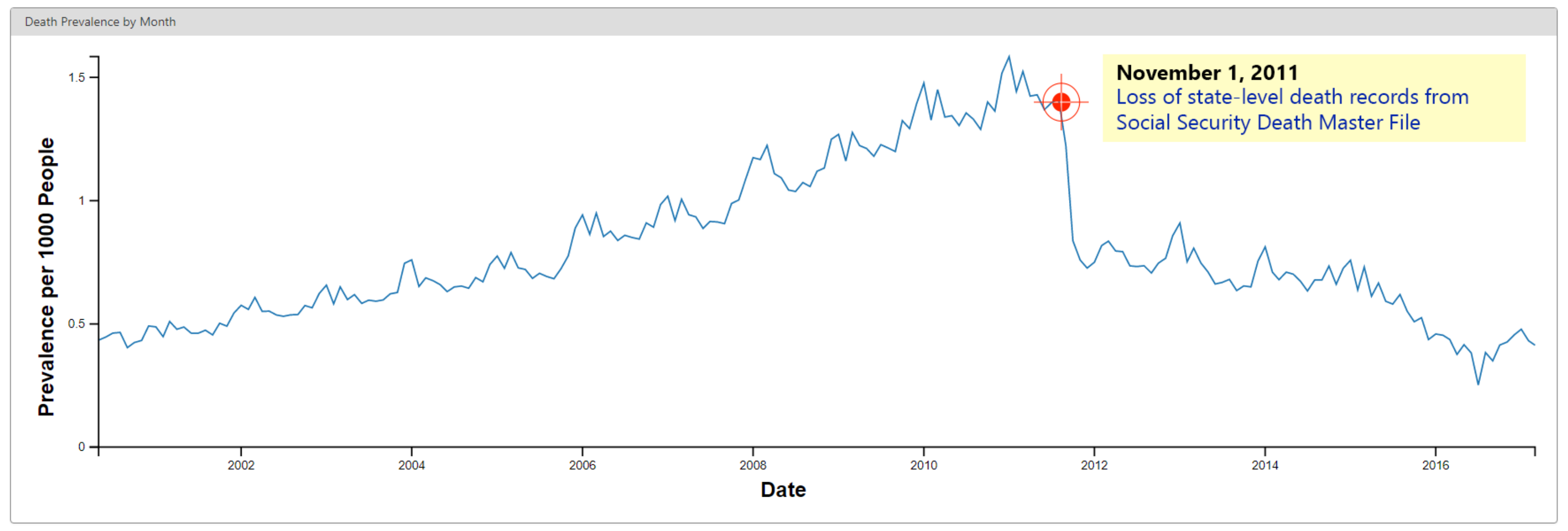


Figure 1: Achilles chart of death records per 1000 people in OPTUM. The incidence of death records dropped significantly in November 2011, due to external policy changes that are not readily apparent.

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).

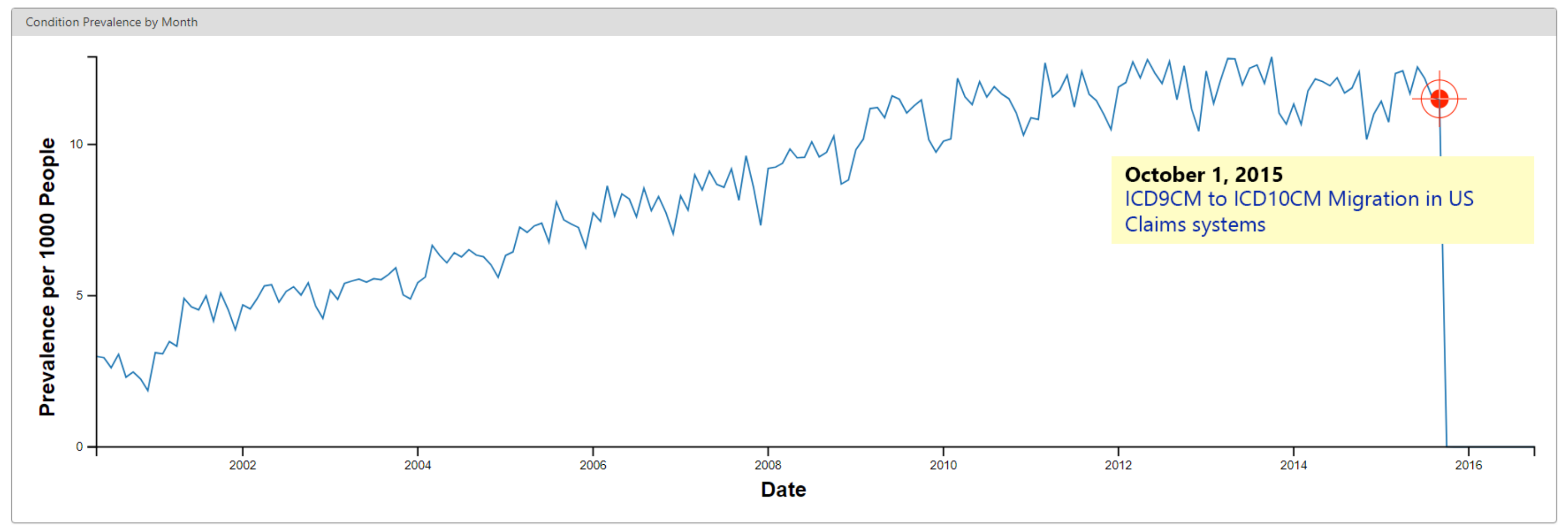


Figure 2: Achilles chart of the prevalence of SNOMED concept 439926, "malaise and fatigue." The switch from ICD9CM to ICD10CM necessitates consideration of broader concept set lists that include concepts that map to the newer source vocabularies.

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 possible 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).

metadata id	metadata concept id	metadata type concept id	name	value as string	value as concept id	metadata datetime	metadata date
1	44819056	1	Death	Loss of state-level death records from Social Security Death Master File	NULL	2011-11-01	2017-08-23
2	439926	19	Malaise and fatigue	ICD9CM to ICD10CM migration	NULL	2015-10-01	2017-08-23

Table 1: Demo of how annotations could be stored in the CDM Metadata table. Comments, tagged with concept ids and dates where applicable, help in explaining an unclear behavior in the data.

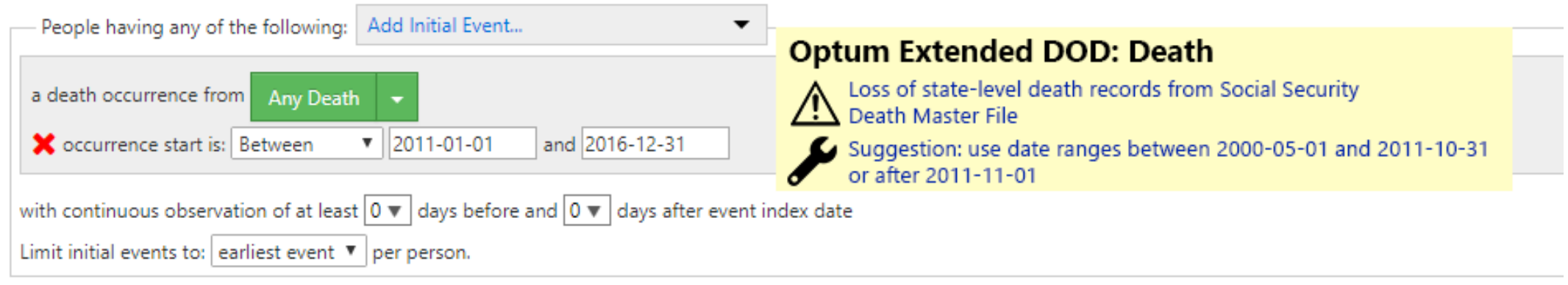
- While informative, the **CDM Metadata table will not help provide solutions** to the problems it describes.
- If the annotations stored in the CDM Metadata table could be considered "observational," then **an "interventional" annotation table could store suggestions** on how best to handle such situations when designing studies that require their data elements (mocked up in Table 2).

metadata id	suggested concept id	valid start date	valid end date
1	0	2000-05-01	2011-10-31
1	0	2011-11-01	2099-12-31
2	4272240	2015-10-01	2099-12-31
2	4223659	2015-10-01	2099-12-31

Table 2: Potential "interventional" annotation table that could be utilized to store solutions for the data quality issues identified in the CDM Metadata table. Suggested concept ids can be stored for metadata records about problematic concepts. Valid start and end dates provide temporal boundaries around the solution.

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).



a death occurrence from Any Death

occurrence start is: Between 2011-01-01 and 2016-12-31

with continuous observation of at least 0 days before and 0 days after event index date

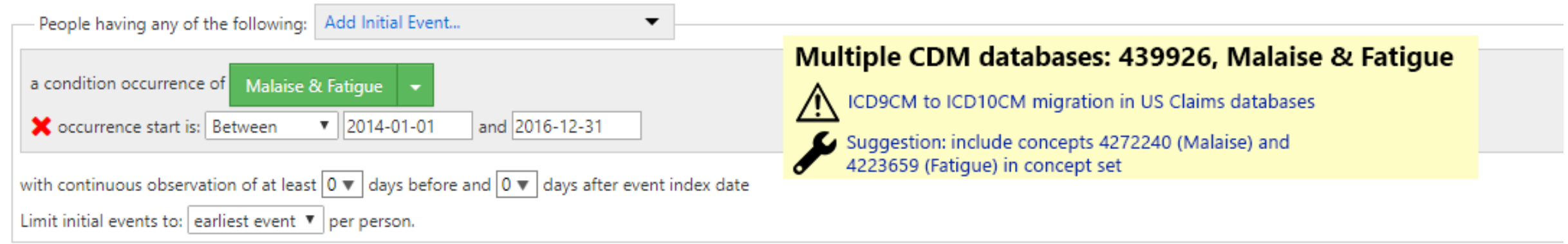
Limit initial events to: earliest event per person.

Optum Extended DOD: Death

⚠ Loss of state-level death records from Social Security Death Master File

Suggestion: use date ranges between 2000-05-01 and 2011-10-31 or after 2011-11-01

Figure 3: Mockup of Atlas cohort designer with annotation and suggestion for Case Study 1. The loss of state-level death records from the SSA DMF is highlighted to the user because they selected a death criteria with a date range that included the date annotated in the CDM Metadata table. The suggested solution is to instead utilize death date windows that begin and end before the DMF change, or begin and end after it.



a condition occurrence of Malaise & Fatigue

occurrence start is: Between 2014-01-01 and 2016-12-31

with continuous observation of at least 0 days before and 0 days after event index date

Limit initial events to: earliest event per person.

Multiple CDM databases: 439926, Malaise & Fatigue

⚠ ICD9CM to ICD10CM migration in US Claims databases

Suggestion: include concepts 4272240 (Malaise) and 4223659 (Fatigue) in concept set

Figure 4: Mockup of Atlas cohort designer with annotation and suggestion for Case Study 2. The migration of US Claims databases from ICD9CM to ICD10CM is highlighted to the user because they selected a condition criteria with a concept that has been annotated to have a data anomaly on October 1, 2015. The suggested solution is to instead build a concept set that includes analogous concepts that include ICD10CM source codes.

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)/ Lister Hill National Center for Biomedical Communications (LHNCBC).

REFERENCES

1. Riley, Jenn. 2017. "Understanding Metadata." {National Information Standards Organization}. Accessed August 7. http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf.
2. John Macmullen W. Annotation as process, thing, and knowledge: Multi-domain studies of structured data annotation. in ASIST Annual Meeting,
3. Huser V, Londhe A, Voss E. Metadata Proposal GitHub2017 [Available from: https://github.com/OHDSI/CommonDataModel/issues/79.
4. Winn D. National Cancer Institute2012. [cited 2017 2017/10/9]. Available from: https://epi.grants.cancer.gov/blog/archive/2012/05-24.html.

Contact: ALondhe2@ITS.JNJ.COM