How do changes in vocabulary mapping and database release versions affect cohort composition in real-world data?

Jill Hardin[1,2], Evanette K Burrows[1,2], Azza Shaoibi[1,2], Clair Blacketer[1,2]

[1]Janssen Research and Development, LLC, Raritan, NJ, USA
[2] Observational Health Data Sciences and Informatics (OHDSI), New York, NY

Background

Evaluating the impact of simultaneous changes in vocabulary mapping and database release versions presents a significant challenge in the analysis of observational data. Understanding how these factors influence cohort composition is essential for accurate epidemiological research.

Objective

This study aimed to explore the effects of concurrent changes in vocabulary mapping and database composition (releases) on cohort composition. We specifically compared incidence rates (IR) of condition and drug phenotypes (cohorts) across different vocabulary versions and database releases for 13 different data sources. We evaluated changes in the IR by calculating the absolute percent differences for 437 predefined cohorts.

Methods

We utilized 13 observational health data sources standardized to the OMOP Common Data Model (CDM) (1) version 5.4 to derive incidence rates. Two versions of each data source were created by mapping the native data to standard concepts using two vocabulary versions: February 29, 2024, and February 27, 2025. A total of 437 phenotype cohorts were then evaluated, comprising 239 clinical outcomes and 198 drug exposures across the 13 data sources x 2 vocabulary versions. The population at risk (target cohort) was constrained to subjects with an observation period between January 1, 2016, and December 31, 2022. Two open-source OHDSI tools, ATLAS (2) and the Strategus incidence rate module facilitated the development of phenotypes and the generation of incidence rates. The 437 outcomes cohorts were all developed and evaluated following OHDSI best practices. The absolute percent difference between incidence rates for each vocabulary and database release version was calculated using the formula:

$$Absolute\ Percent\ Difference \frac{(incidence\_rate\_p100py\_new\_vocab - \ incidence\_rate\_p100py\_prior\_vocab)}{incidence\_rate\_p100py\_prior\_vocab}\ x\ 100$$

Results

The categories of absolute percent differences in incidence rates between the vocabulary versions and database releases are summarized in Table 1. The majority (31 to 85%) of cohorts exhibited absolute percent differences in the >0-5% range. The databases with the largest cohort changes (absolute percent differences >5%) included Optum EHR (12.6%) and CPRD (7.3%). Figures 1 and 2 show the heatmap of the absolute percent difference values greater than 5% by database and cohort. The cohorts that had the largest absolute percent differences included heavy menstrual bleeding (menorrhagia) in Optum EHR and CPRD and epoprostenol in JMDC. The native data in the Optum EHR and CPRD databases use the SNOMED vocabulary while all other databases use ICD10. Examination of the JMDC native data showed that release to release drug mappings were added for many drugs, including epoprostenol, and that the codes were being mapped to the English database representation.

Conclusion

Cohort composition generally remains stable across data releases and updates to the vocabulary. However, in some data sources, few cohorts may be sensitive to changes in vocabulary or data releases. The observed large absolute percent differences in specific cohorts and database combinations can be attributed to shifts in clinical concepts within the SNOMED hierarchy, particularly for menstrual bleeding, and updates to the JMDC drug mapping file. Data sources that use the standard SNOMED vocabulary in its native format were more susceptible to big changes in SNOMED hierarchy.

This study emphasizes the importance of continuously adapting and understanding the implications of vocabulary and data release changes in real-world data analyses. The approach used in this study can be used to design a standardized process to assess the effect of vocabulary and data releases on existing cohort definitions (phenotypes).

Table 1: The number (%) of phenotypes by categories of absolute percent difference between incidence rates

Database	The number (%) of phenotypes by category of absolute percent difference in incidence rates											
	>0-5		>5-10		>10-20		>20-50		>50+		NA*	
	N	%	N	%	N	%	N	%	N	%	N	%
CPRD	261	59.7%	13	3.0%	12	2.7%	6	1.4%	1	0.2%	144	32.9%
France Disease Analyzer	205	46.9%	0	0.0%	2	0.5%	5	1.1%	2	0.5%	223	51.0%
German Disease Analyzer	278	63.6%	5	1.1%	2	0.5%	5	1.1%	3	0.7%	144	32.9%
Health Verity CC	370	84.7%	10	2.3%	5	1.1%	3	0.7%	0	0.0%	49	11.2%
JMDC	277	63.4%	6	1.4%	7	1.6%	6	1.4%	7	1.6%	134	30.6%
LPD Australia	136	31.1%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	301	68.8%
CCAE	367	84.0%	2	0.5%	11	2.5%	2	0.5%	0	0.0%	55	12.6%
MDCR	351	80.3%	3	0.7%	8	1.8%	1	0.2%	0	0.0%	74	16.9%

MDCD	369	84.4%	3	0.7%	7	1.6%	3	0.7%	0	0.0%	55	12.6%
Optum EHR	342	78.3%	24	5.5%	22	5.0%	7	1.6%	2	0.5%	40	9.1%
Optum DOD	372	85.1%	2	0.5%	8	1.8%	3	0.7%	0	0.0%	52	11.9%
Optum SES	372	85.1%	2	0.5%	8	1.8%	3	0.7%	0	0.0%	52	11.9%
Premier	283	64.8%	8	1.8%	6	1.4%	1	0.2%	0	0.0%	139	31.8%
*NOTE: NA bin is created when the incidence rate on new and old vocabulary are equal to 0												

Figure 1. Outcome cohorts with absolute percent difference values >5% by database

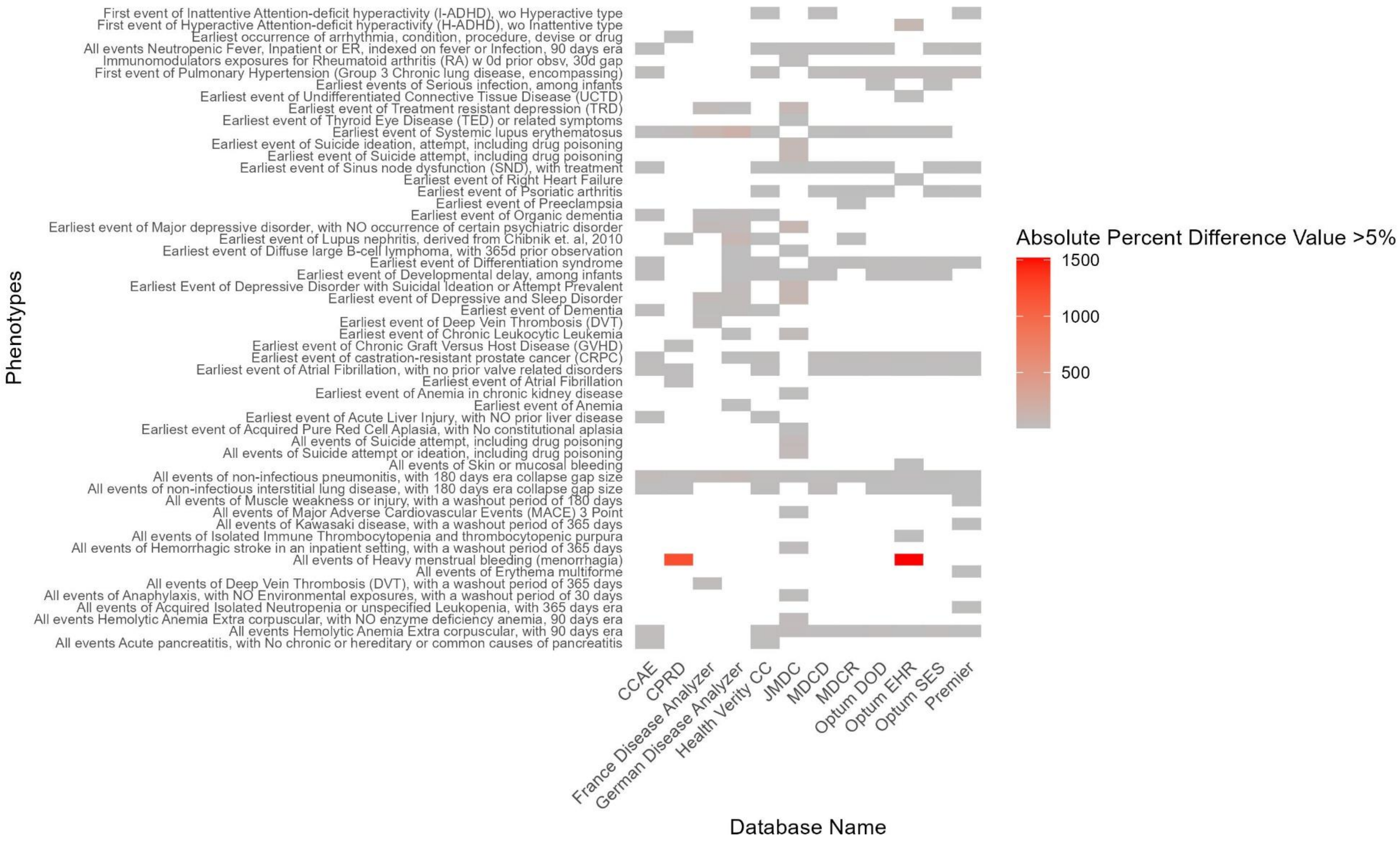
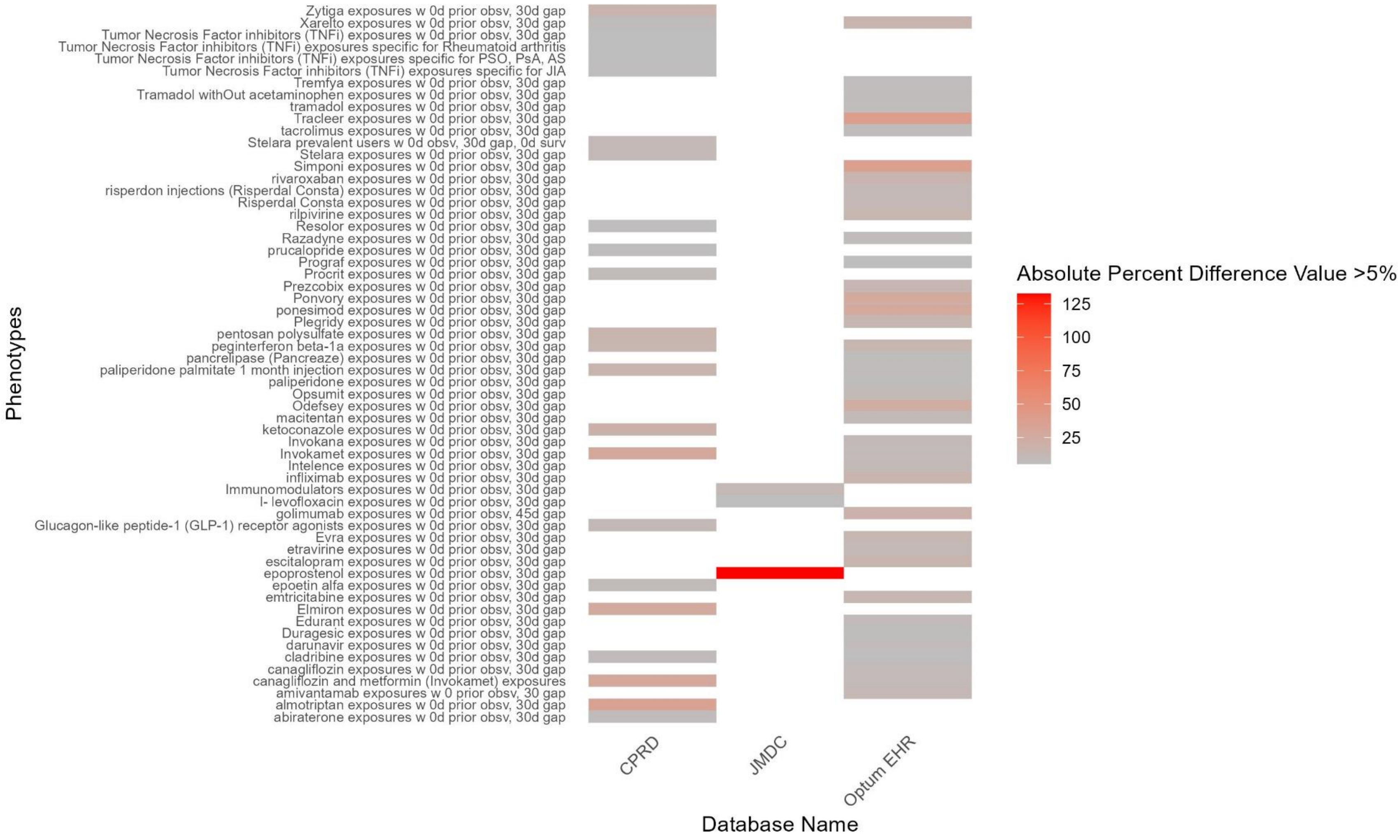

First event of Inattentive Attention-deficit hyperactivity (I-ADHD), wo Hyperactive type
First event of Hyperactive Attention-deficit hyperactivity (H-ADHD), wo Inattentive type
Earliest occurrence of arrhythmia, condition, procedure, devise or drug
All events Neutropenic Fever, Inpatient or ER, indexed on fever or Infection, 90 days era
Immunomodulators exposures for Rheumatoid arthritis (RA) w 0d prior obsv, 30d gap
First event of Pulmonary Hypertension (Group 3 Chronic lung disease, encompassing)
Earliest events of Serious infection, among infants
Earliest event of Undifferentiated Connective Tissue Disease (UCTD)
Earliest event of Treatment resistant depression (TRD)
Earliest event of Thyroid Eye Disease (TED) or related symptoms
Earliest event of Systemic lupus erythematosus
Earliest event of Suicide ideation, attempt, including drug poisoning
Earliest event of Suicide attempt, including drug poisoning
Earliest event of Sinus node dysfunction (SND), with treatment
Earliest event of Right Heart Failure
Earliest event of Psoriatic arthritis
Earliest event of Preeclampsia
Earliest event of Organic dementia
Earliest event of Major depressive disorder, with NO occurrence of certain psychiatric disorder
Earliest event of Lupus nephritis, derived from Chibnik et. al, 2010
Earliest event of Diffuse large B-cell lymphoma, with 365d prior observation
Earliest event of Differentiation syndrome
Earliest event of Developmental delay, among infants
Earliest Event of Depressive Disorder with Suicidal Ideation or Attempt Prevalent
Earliest event of Depressive and Sleep Disorder
Earliest event of Dementia
Earliest event of Deep Vein Thrombosis (DVT)
Earliest event of Chronic Leukocytic Leukemia
Earliest event of Chronic Graft Versus Host Disease (GVHD)
Earliest event of castration-resistant prostate cancer (CRPC)
Earliest event of Atrial Fibrillation, with no prior valve related disorders
Earliest event of Atrial Fibrillation
Earliest event of Anemia in chronic kidney disease
Earliest event of Anemia
Earliest event of Acute Liver Injury, with NO prior liver disease
Earliest event of Acquired Pure Red Cell Aplasia, with No constitutional aplasia
All events of Suicide attempt, including drug poisoning
All events of Suicide attempt or ideation, including drug poisoning
All events of Skin or mucosal bleeding
All events of non-infectious pneumonitis, with 180 days era collapse gap size
All events of non-infectious interstitial lung disease, with 180 days era collapse gap size
All events of Muscle weakness or injury, with a washout period of 180 days
All events of Major Adverse Cardiovascular Events (MACE) 3 Point
All events of Kawasaki disease, with a washout period of 365 days
All events of Isolated Immune Thrombocytopenia and thrombocytopenic purpura
All events of Hemorrhagic stroke in an inpatient setting, with a washout period of 365 days
All events of Heavy menstrual bleeding (menorrhagia)
All events of Erythema multiforme
All events of Deep Vein Thrombosis (DVT), with a washout period of 365 days
All events of Anaphylaxis, with NO Environmental exposures, with a washout period of 30 days
All events of Acquired Isolated Neutropenia or unspecified Leukopenia, with 365 days era
All events Hemolytic Anemia Extra corpuscular, with NO enzyme deficiency anemia, 90 days era
All events Hemolytic Anemia Extra corpuscular, with 90 days era
All events Acute pancreatitis, with No chronic or hereditary or common causes of pancreatitis
Phenotypes
CCAE
CPRD
France Disease Analyzer
German Disease Analyzer
Health Verity CC
JMDC
MDCD
MDCR
Optum DOD
Optum EHR
Optum SES
Premier
Database Name
Absolute Percent Difference Value >5%
1500
1000
500

Figure 1. Drug exposure cohorts with absolute percent difference values >5% by database



References

1. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc. 2015;22(3):553-64.
2. https://github.com/OHDSI/Atlas
3. Sena A, Schuemie M, Gilbert J (2025). Strategus: Coordinate and Execute OHDSI HADES Modules. R package version 1.3.0, https://github.com/OHDSI/Strategus, https://ohdsi.github.io/Strategus.