Fidelity Assessment of the Clinical Practice Research Datalink Transformation to the OMOP Common Data Model with a Replication Study

Amy Matcho[1], Patrick B. Ryan, PhD[1]

[1]Janssen Research & Development, LLC, Titusville, NJ

BACKGROUND

- The Clinical Research Practice Datalink (CPRD), an electronic health record from the United Kingdom, is one of the primary observational databases used for epidemiological studies. However, its unique structure and coding present challenges in analysis.
- The OMOP Common Data Model (CDM) facilitates efficient, comparable and systematic large-scale drug safety analysis across disparate databases. [1]
- Open question remained whether CPRD could be converted to the CDM, and to what extent that conversion might present issues that would limit its use.

OBJECTIVES

I. Transform CPRD into the OMOP CDM Model V4 and quantify quality of mappings and drug exposure duration imputation.

II. Conduct a population-based replication study on the current raw CPRD data and the transformed CDM. Verify the CPRD CDM instance created is an acceptable approximation of the raw data by comparing results of both studies.

METHODS

I. Janssen 2013 CPRD CDM Transformation:

- Utilized improved Multilex drug code mappings to RxNorm in OMOP standard dictionaries.
- Some additional data included over prior efforts:
  - Procedure.
  - Entire body of lifestyle (smoking, BMI, etc.), disease and scoring data from the CPRD 'Additional' file.
- 93% of drug exposure durations in CPRD unpopulated; necessary to impute.
  - Most common durations in the data for unique combinations of product, quantity, numeric daily dose and number of packs were used for the imputation, mode of duration for product only was used for non-existing combinations.

II. Original Published Study Definition:

- Use of nonsteroidal anti-inflammatory drugs (NSAIDs) and the risk of first-time acute myocardial infarction by Schlienger et al in 2002. [2]
- Case-control analysis with conditional logistic regression model.

Cases:
- Incident Acute Myocardial Infarction (AMI) patients.
- Excluded any patients with prior history of metabolic or cardiovascular disease predisposing to AMI >60 days before AMI.
- 3 years of data prior to index date.

Controls:
- Matched on year of birth, gender, practice and calendar time using index date of case
- Excluded any patients with prior history of metabolic or cardiovascular disease predisposing to AMI >60 days before case AMI.
- 3 years of data prior to case index date.

Exposure Definition:
- Only traditional type NSAID ingredients.
- Categories of Users:
  - Current User: supply of last prescription prior to index date ended at
  - or after index date.
  - Recent User: supply ended between 1 to 29 days before index date.
  - Past User: supply ended 30 or more days prior to index date.
  - Non Users: had no NSAID records prior to index date.

RESULTS

I. CDM Transformation

- Completeness of data mappings in CDM:
  - 89.5% of all drug exposures in data mapped
  - 99.5% of all conditions in data mapped
  - 99.0% of all procedures in data mapped
  - 92.0% of all observations in data mapped

- Top 100 source code mappings for select domains were examined for accuracy and completeness.
  - Unmapped drug exposures primarily due to United Kingdom ingredients and formulations not available in the US.
  - Unmapped observations can be attributed to source data from the CPRD 'Additional' file with no matches in the Logical Observation Identifiers Names and Codes (LOINC) dictionary.

Table 1: Top 100 CPRD source code mapping accuracy and completeness

Domain	% CDM Data	Mapped Accurately	Unmapped % CDM Data
Drug Exposure	42.0	93.0	7 (1.5)
Condition	38.0	100.0	0 (0.0)
Procedure	76.0	98.0	2 (0.5)
Observation	75.0	93.0	7 (3.0)

- Drug exposure duration imputation validation yielded 69,244 imputations (6% of data) out of ~300,000 that were potentially inaccurate.
- Valid imputations defined as:
  - Duration equal to or within 5 days of quantity/numeric daily dose (57% of data).
  - Common durations of 28 and 30 (33% of data).
  - Numeric daily dose equal to 0 and duration equal to quantity (4% of data).

II. Replication Study

- We compared demographics and patient characteristics, NSAID exposures, and risk of first-time AMI with NSAIDs between raw data study and CDM study.
- Also compared NSAID exposure categories between original published study and raw data study to validate drug exposure duration imputation which was used for raw data study and CDM.

Table 2: Patient characteristic percentages

		Original Study		Raw CPRD Data		CPRD CDM	
		Cases	Controls	Cases	Controls	Cases	Controls
Age	<40	2.8	2.8	2.1	2.2	2.1	2.2
	40-49	12.6	12.6	11.2	11.4	11.2	11.4
	50-59	25.0	25.2	25.2	25.4	25.2	25.4
	60-69	37.0	36.8	35.9	36.1	35.8	36.1
	70-75	22.6	22.6	25.6	25.0	25.6	25.0
Sex	Male	74.0	73.9	74.2	74.4	74.1	74.4
	Female	26.0	26.1	25.8	25.6	25.9	25.6
Smoking	Non	32.6	47.2	24.9	33.6	24.9	33.3
	Current	33.2	19.6	28.4	15.9	28.4	15.5
	Ex	11.3	10.3	8.2	7.0	8.1	6.8
	Unknown	22.9	22.9	38.5	43.5	38.5	44.4
BMI	<25	26.7	32.3	20.9	21.0	20.9	21.3
	25-29.9	33.2	30.5	21.8	21.8	21.8	20.8
	>=30	11.7	9.2	8.6	6.6	8.6	6.5
	Unknown	28.4	28.0	48.7	50.6	48.7	51.5

Table 2: Comparison of case characteristic proportions:
- **Raw data vs. CDM**: Age, sex, smoking status and BMI case proportions remained constant across both studies.

Table 3: Percentage NSAID exposure stratified by current, recent past and past categories, duration (number of prescriptions) and ingredient

	Original Study		CPRD Raw Data		CPRD CDM	
	Cases	Controls	Cases	Controls	Cases	Controls
Non-users	45.3	47.5	57.1	63.2	58.5	63.3
Current NSAIDs	7.3	6.3	6.2	4.2	5.4	4.1
1-4 Rx	1.0	0.8	1.3	1.0	1.3	0.9
30+ Rx	2.7	2.2	2.0	1.1	1.7	1.4
Recent past NSAIDs	3.6	2.9	3.2	2.1	2.8	1.9
1-4 Rx	0.8	0.8	1.0	0.8	1.0	0.9
30+ Rx	1.1	0.4	0.7	0.3	0.5	0.1
Past NSAIDs	43.8	43.4	33.4	30.5	33.3	30.7
1-4 Rx	29.7	30.5	26.2	24.8	25.9	24.6
30+ Rx	1.2	0.6	0.5	0.4	0.9	0.5
Current Users:						
Ibuprofen	1.8	1.6	2.1	1.1	2.0	1.2
Diclofenac	2.9	2.1	1.9	1.2	1.7	1.3
Piroxicam	0.3	0.2	0.3	0.3	0.3	0.3
Ketoprofen	0.5	0.4	0.2	0.3	0.1	0.2
Indomethacin	0.5	0.4	0.5	0.3	0.5	0.3
Naproxen	0.6	0.8	0.5	0.5	0.4	0.5

Table 3: Comparison of case NSAID exposure proportions:
- **Raw data vs. CDM:** Majority of NSAID exposure proportions for cases were <1% lower for all user categories in the CDM study. Slight differences can be explained by a very small amount of unmapped NSAIDs and mapped NSAIDs without ingredient relationships in the Multilex dictionary mappings to RxNorm.
- **Original published study vs. Raw data**: Less NSAID exposures found overall in raw data study, but assignments to NSAID user categories remained similar across both studies.

Table 4: Odds ratios of first-time AMI in NSAID users adjusted for smoking status, BMI, HRT and aspirin (reference group non-users)

	Original Study	CPRD Raw Data	CPRD CDM
Current NSAIDs	1.17 (0.99–1.37)	1.64(1.35,2.00)	1.39(1.13,1.70)
1-4 Rx	1.30 (0.87–1.93)	1.21(0.82,1.80)	1.40(0.92,2.11)
30+ Rx	1.21 (0.94–1.55)	1.90(1.32,2.73)	1.12(0.79,1.60)
Recent past NSAIDs	1.26 (1.01–1.57)	1.54(1.18,2.01)	1.57(1.17,2.10)
1-4 Rx	0.95 (0.61–1.48)	1.32(0.84,2.07)	1.24(0.79,1.96)
30+ Rx	2.71 (1.75–4.22)	3.11(1.63,5.97)	3.68(1.52,8.87)
Past NSAIDs	1.04 (0.96–1.13)	1.19(1.09,1.30)	1.14(1.05,1.25)
1-4 Rx	1.02 (0.93–1.12)	1.15(1.04,1.27)	1.11(1.01,1.23)
30+ Rx	2.33 (1.57–3.46)	1.89(0.93,3.81)	1.61(0.95,2.72)
Current Users:			
Ibuprofen	1.17 (0.87–1.58)	1.72(1.23,2.40)	1.61(1.14,2.28)
Diclofenac	1.38 (1.08–1.77)	1.86(1.30,2.66)	1.46(1.01,2.11)
Piroxicam	1.65 (0.78–3.49)	0.93(0.40,2.15)	0.83(0.38,1.84)
Ketoprofen	1.39 (0.77–2.51)	0.80(0.29,2.23)	0.67(0.22,2.05)
Indomethacin	1.03 (0.58–1.85)	1.97(0.99,3.93)	2.11(1.05,4.25)
Naproxen	0.68 (0.42–1.13)	0.93(0.48,1.80)	0.70(0.36,1.37)

Table 4: Comparison of associations between NSAIDS and first-time AMI:
- **Raw data vs. CDM:** Adjusted odds ratios for first-time AMI in NSAID user categories are similar.

CONCLUSIONS

- CRPD can be accurately transformed into the OMOP CDM, with minimal information loss across drugs, conditions and observations. The majority of CPRD source codes were successfully mapped to CDM concepts.
- The CPRD CDM replication analysis was easier to perform due to the standardized structure of the data and useful derived constructs such as the drug era file. Quality of additional analyses will be improved as validated algorithms within the CDM can be leveraged.
- The CPRD CDM can be a valuable part of future efforts to compare CPRD to other observational databases.
- The same duration imputation was used in the raw data study and CDM study in order to validate directly against a prior study's methods. The categorization of NSAID exposure into 'current', 'recent past' and 'past' users required the imputation in the raw data study and the categorizations compared favorably with the original published study.
- Though our objective did not involve an exact replication of the original published study on the raw data, any effort in that direction was hindered by the fact that underlying source system changes were implemented after the authors completed their analysis.

REFERENCES

1. Overhage, J.M., et al., *Validation of a common data model for active safety surveillance research.* J Am Med Inform Assoc, 2012. 19(1): p. 54-60.
2. Schlienger, R. G., et al., *Use of nonsteroidal anti-inflammatory drugs and the risk of first-time acute myocardial infarction.* Br J Clin Pharmacol, 2002. 54(3): p. 327-332.

This study was conducted by Janssen Research & Development, LLC.