# Current status of OMOP-CDM in OHDSI APAC regions : Lessons for Data Quality Assessment

Chungsoo Kim
on be half of OHDSI APAC Study Team

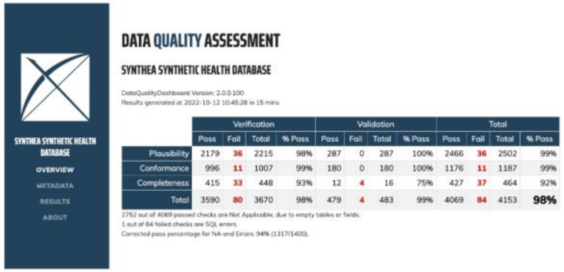
2022-11-13

# Background

- The adoption of OMOP-CDM in the Asia-Pacific (APAC) region is increasing.
- A lot of database in the APAC region is still in the conversion stage or just after the conversion is completed, therefore, an extensive quality assessment is needed.
- For preventing errors on ETL process, there are tools (Achilles Heel, DQD) have been developed in the OHDSI community  to check the quality of data.[1), 2)]

OHDSI/**Achilles**

Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES) - descriptive statistics about a OMOP CDM database



DATA QUALITY ASSESSMENT

SYNTHEA SYNTHETIC HEALTH DATABASE

DataQualityDashboard Version: 2.0.0.100
Results generated at 2022-10-12 10:45:28 in 15 mins

| | Verification | | | | Validation | | | | Total | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pass | Fail | Total | % Pass | Pass | Fail | Total | % Pass | Pass | Fail | Total | % Pass |
| Plausibility | 2179 | 36 | 2215 | 98% | 287 | 0 | 287 | 100% | 2466 | 36 | 2502 | 99% |
| Conformance | 996 | 11 | 1007 | 99% | 180 | 0 | 180 | 100% | 1176 | 11 | 1187 | 99% |
| Completeness | 415 | 33 | 448 | 93% | 12 | 4 | 16 | 75% | 427 | 37 | 464 | 92% |
| Total | 3590 | 80 | 3670 | 98% | 479 | 4 | 483 | 99% | 4069 | 84 | 4153 | 98% |

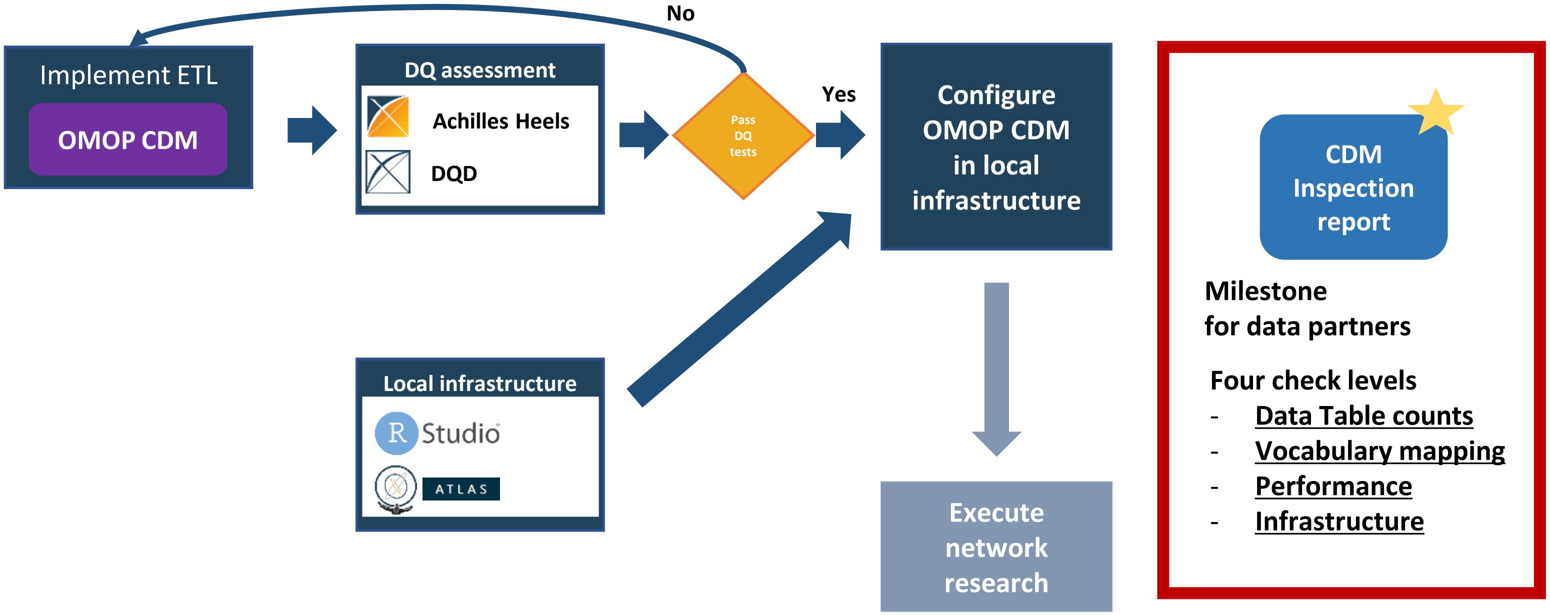
2752 out of 4069 passed checks are Not Applicable, due to empty tables or fields.
1 out of 84 failed checks are SQL errors.
Corrected pass percentage for NA and Errors: 94% (1317/1400).

*1) Callahan TJ, Bauck AE, Bertoch D, Brown J, Khare R, Ryan PB, Staab J, Zozus MN, Kahn MG. A Comparison of Data Quality Assessment Checks in Six Data Sh aring Networks. EGEMS (Wash DC). 2017 Jun 12;5(1):8. doi: 10.5334/egems.223. PMID: 29881733; PMCID: PMC5982846.*

*2) Clair Blacketer, Frank J Defalco, Patrick B Ryan, Peter R Rijnbeek, Increasing trust in real-world evidence through evaluation of observational data quality, J ournal of the American Medical Informatics Association, Volume 28, Issue 10, October 2021, Pages 2251–2257, https://doi.org/10.1093/jamia/ocab132*

# Background

- However, the current **quality assessment is usually conducted during ETL.**
- Macroscopic statistics including data count, distribution of data, composition of data, vocabulary mapping status may also be included as a quality control items.
- CDM inspection report is for checking four levels of converted OMOP-CDM : Data table counts, Vocabulary mapping, Performance, Infrastructure



# Objectives

### What is this study for?

- Collecting CDM Inspection reports from OHDSI APAC community

### Why is this study needed?

- To check the current status of OMOP-CDMs, to get insights from the our CDMs, and to seek quality improvement point.

### What is the final goal?

- It could provide a basic reference of statistics which can be used for future CDM conversion.
- Disclosure of current status of conversion, contents, and data distribution of CDMs of the OHDSI APAC community.

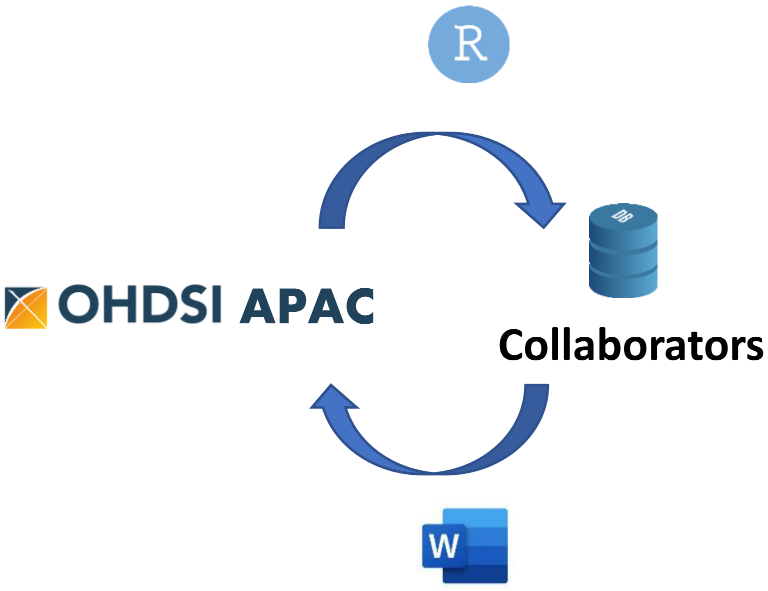
# Methods

- Data sources: CDM databases from OHDSI APAC community
- Collecting inspection reports from each site.
- R package for automatically creating inspection reports.

- Collectibles
  - Number of record, person
  - Number of unique concepts per person
  - Source-CDM mapping ratio
  - Proportion of standard concepts in mapped codes
  - Drug mapping level (granularity)
  - Frequent concept list in each domain
  - Achilles heel result (error / notification / warnings)

# Methods

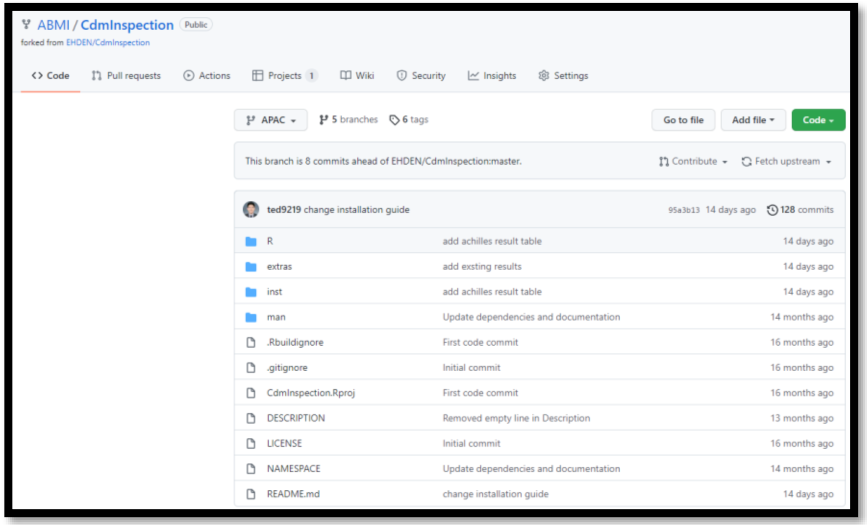
## CDM Inspection package



Collaborators

## CDM Inspection report

Prerequisite : Achilles Heel
https://github.com/ohdsi/achilles



ABMI / CdmInspection Public
forked from EHDEN/CdmInspection

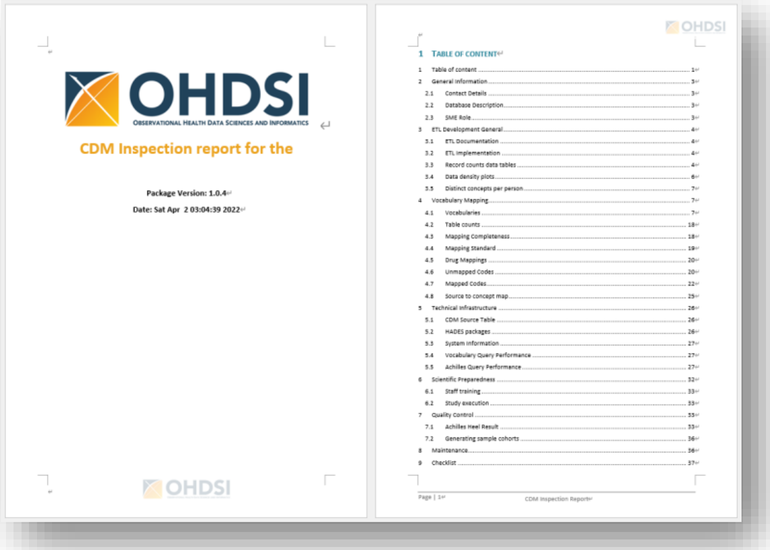
APAC · 5 branches · 6 tags

This branch is 8 commits ahead of EHDEN/CdmInspection:master.

ted9219 change installation guide · 95a3b13 · 14 days ago · 128 commits

| | | |
|---|---|---|
| R | add achilles result table | 14 days ago |
| extras | add exsting results | 14 days ago |
| inst | add achilles result table | 14 days ago |
| man | Update dependencies and documentation | 14 months ago |
| .Rbuildignore | First code commit | 16 months ago |
| .gitignore | Initial commit | 16 months ago |
| CdmInspection.Rproj | First code commit | 16 months ago |
| DESCRIPTION | Removed empty line in Description | 13 months ago |
| LICENSE | Initial commit | 16 months ago |
| NAMESPACE | Update dependencies and documentation | 14 months ago |
| README.md | change installation guide | 14 days ago |

https://github.com/ABMI/CdmInspection/tree/APAC



CDM Inspection report for the

Package Version: 1.0.4

Date: Sat Apr 2 03:04:39 2022

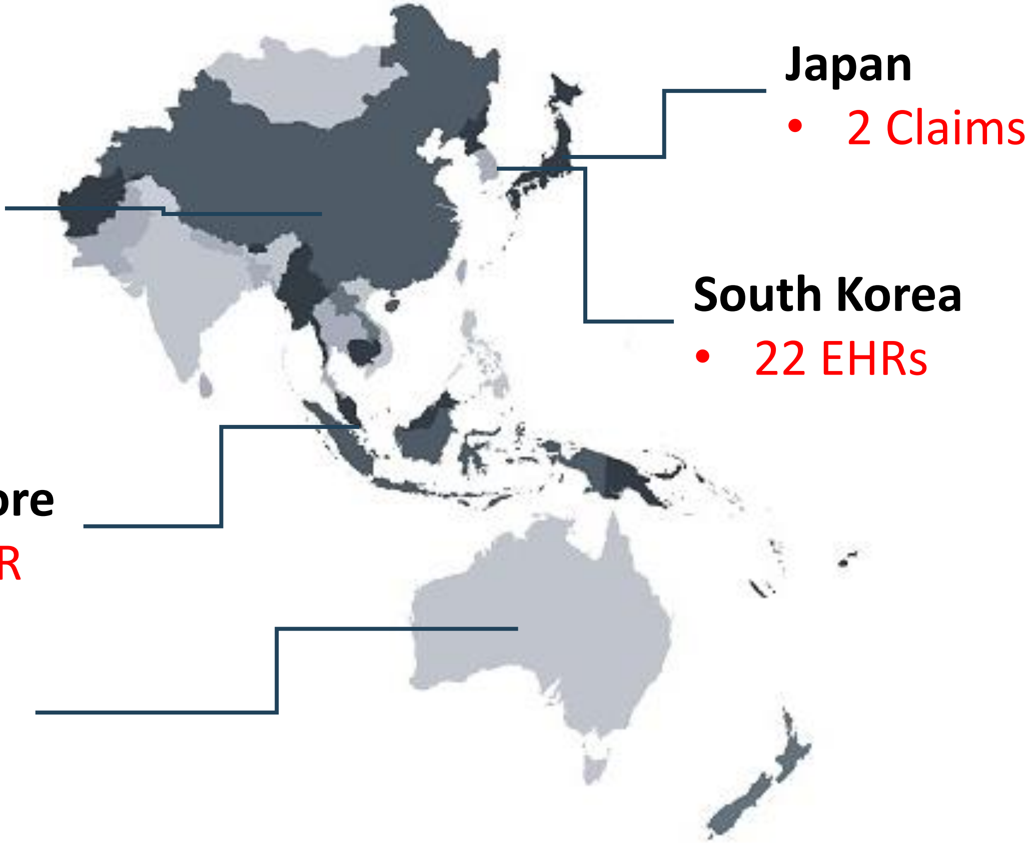
1 TABLE OF CONTENT

1 Table of content ........ 1
2 General information ........ 3
2.1 Contact Details ........ 3
2.2 Database Description ........ 3
2.3 SME Role ........ 3
3 ETL Development General ........ 4
3.1 ETL Documentation ........ 4
3.2 ETL implementation ........ 4
3.3 Record counts data tables ........ 4
3.4 Data density plots ........ 6
3.5 Distinct concepts per person ........ 7
4 Vocabulary Mapping ........ 7
4.1 Vocabularies ........ 7
4.2 Table counts ........ 18
4.3 Mapping Completeness ........ 18
4.4 Mapping Standard ........ 19
4.5 Drug Mappings ........ 20
4.6 Unmapped Codes ........ 20
4.7 Mapped Codes ........ 22
4.8 Source to concept map ........ 25
5 Technical infrastructure ........ 26
5.1 CDM Source Table ........ 26
5.2 HADES packages ........ 26
5.3 System information ........ 27
5.4 Vocabulary Query Performance ........ 27
5.5 Achilles Query Performance ........ 27
6 Scientific Preparedness ........ 32
6.1 Staff training ........ 33
6.2 Study execution ........ 33
7 Quality Control ........ 33
7.1 Achilles Heel Result ........ 33
7.2 Generating sample cohorts ........ 36
8 Maintenance ........ 36
9 Checklist ........ 37

Page | 1 CDM Inspection Report

# Methods

**Data partners**
**27 databases in 5 Countries**



**Japan**

- 2 Claims

**China**

- 1 EHR

**South Korea**

- 22 EHRs

**Singapore**

- 1 EHR

**Australia**

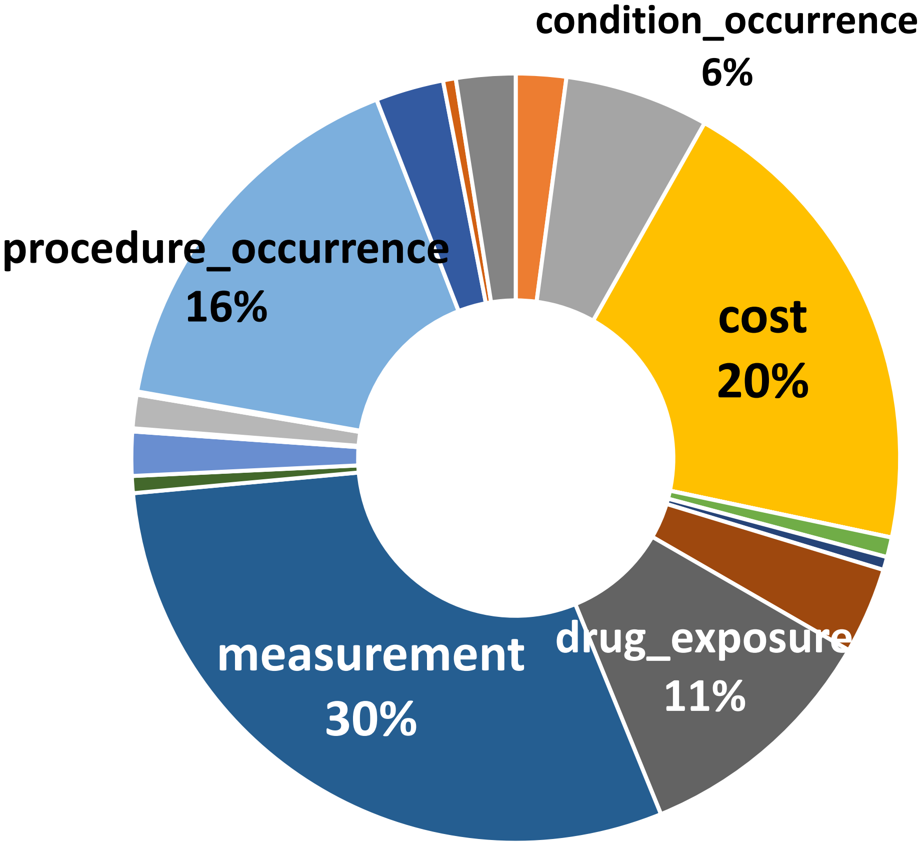
- 1 EHR

# Results

*– Statistics Summary*

**Total number of patients = 49,438,422 (49M)**

**Total data records = 37,690,792,027 (37 Billion)**



# Results

*– Statistics Summary*

By data type

By data period

Claims

General hospital

Tertiary hospital

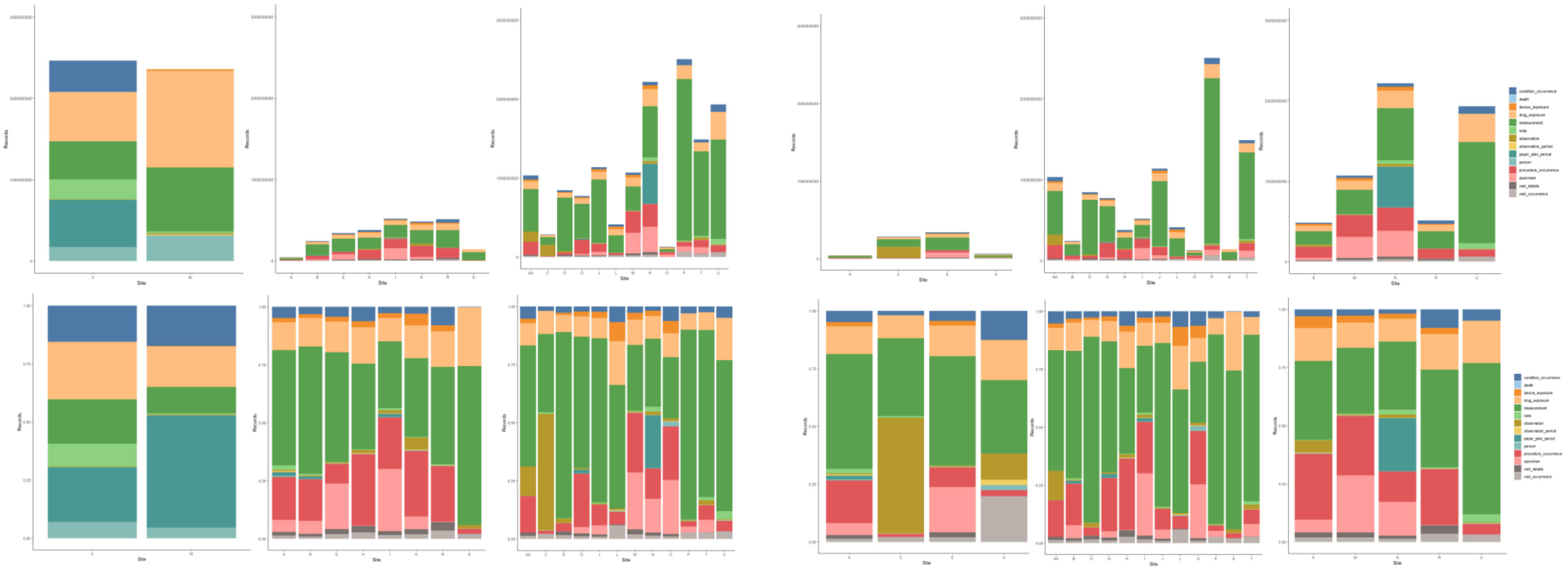
< 10 yr

10-20 yr

> 20 yr

Counts

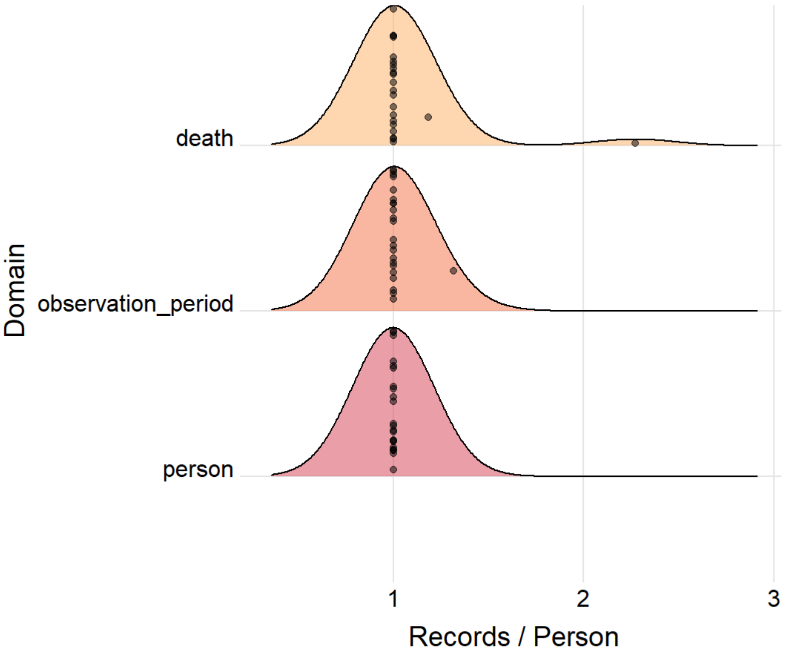
Prop.

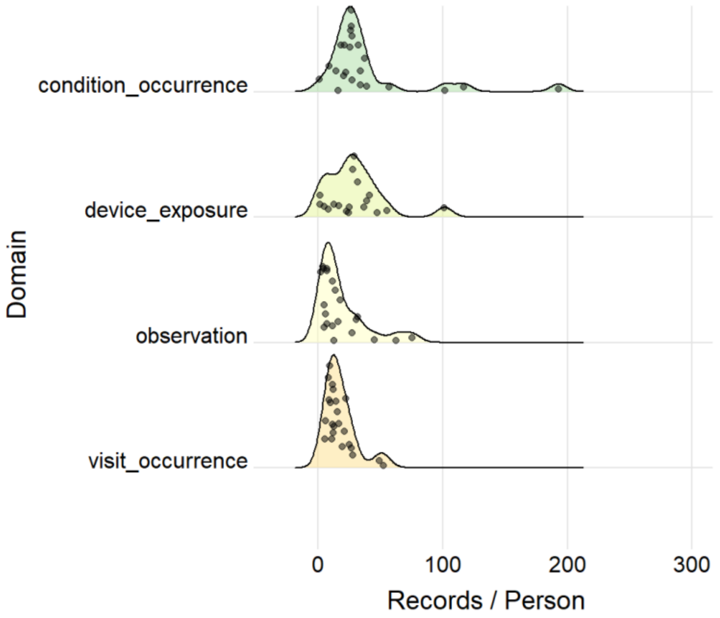


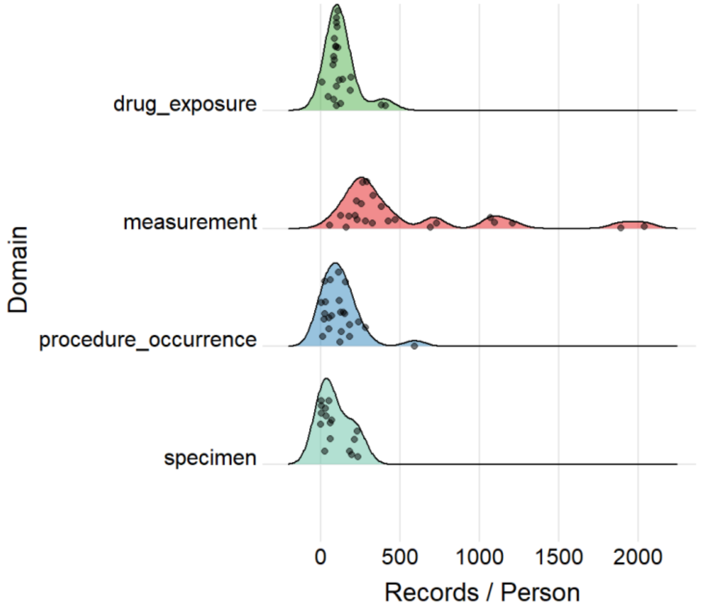
**Records proportion by data type and data periods between domains in each database**
Each institution has a different ratio of the number of records for each domain. If a specific domain is abnormally high, a quality check process could be required.

# Results

## – *Statistics Summary*



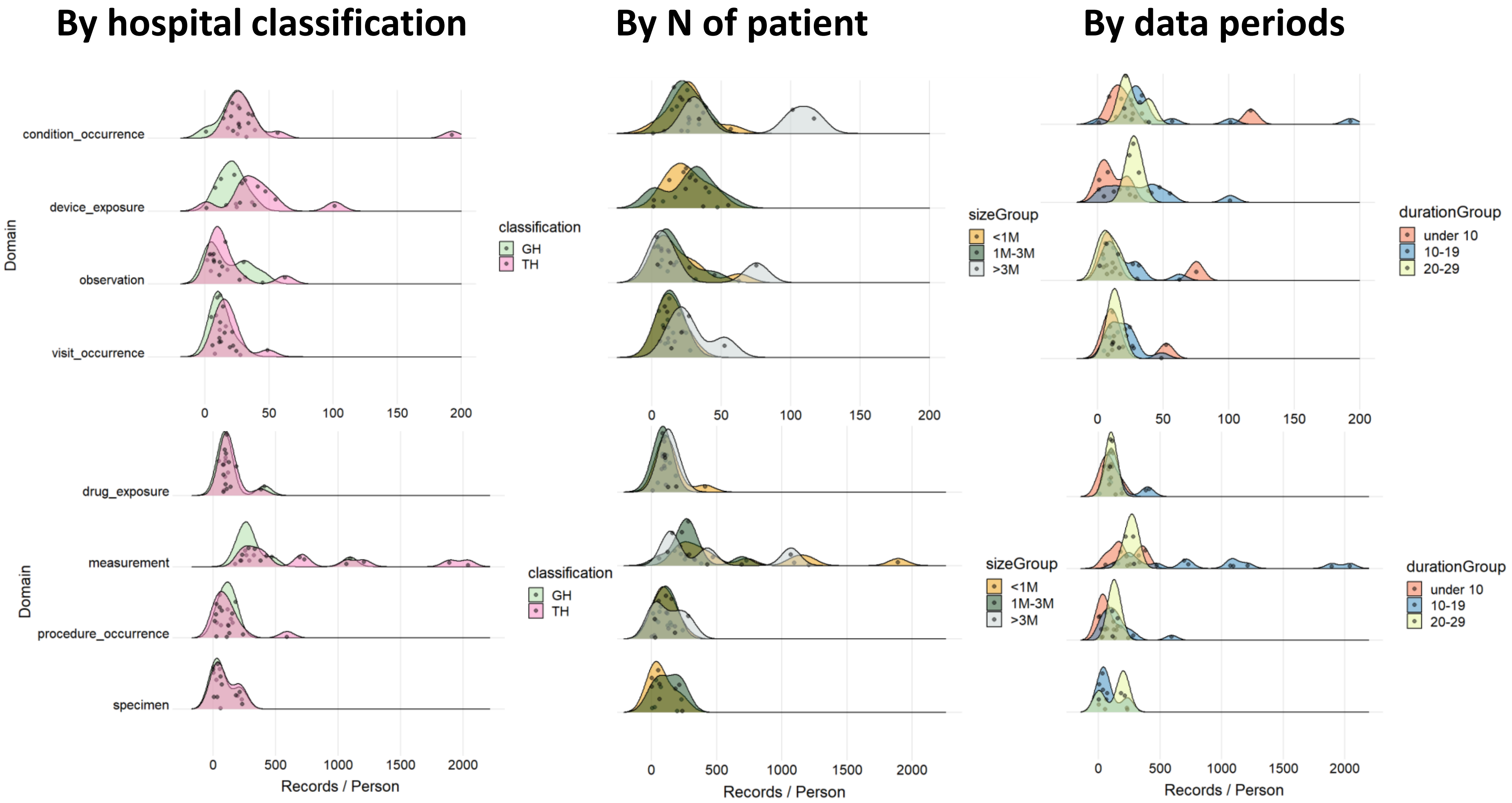




**Distribution of the records to person ratio in each domain**
The records to person ratio has a specific distribution for each domain. A quality check could be needed if you have outliers compared to other databases.

# Results

## – *Statistics Summary*



By hospital classification

By N of patient

By data periods

**Figure 3. Distribution of the records to person ratio in each domain** The records to person ratio has a specific distribution for each domain. A quality check could be needed if you have outliers compared to other databases.

# Results
*– Mapping*

## Summary results of mapping

| Domain | Mapping codes / source codes | Mapped records / total records | Mapped as standard / Mapped records |
|---|---|---|---|
| | Median [Q1, Q3] | Median [Q1, Q3] | Median [Q1, Q3] |
| Condition occurrence | 97.5 [88.8, 99.5] | 99.5 [94.1, 100.0] | 100.0 [99.1, 100.0] |
| Device exposure | 58.7 [49.1, 83.1] | 77.8 [66.1, 94.7] | 79.9 [54.9, 100.0] |
| Drug exposure | 83.8 [73.9, 90.1] | 96.2 [94.7, 98.0] | 98.3 [97.6, 99.0] |
| Measurement | 50.1 [24.8, 84.0] | 92.4 [65.7, 99.5] | 100.0 [99.7, 100.0] |
| Measurement-unit | 96.2 [50.7, 100.0] | 96.7 [35.8, 100.0] | 100.0 [98.4, 100.0] |
| Measurement-value | 11.4 [0.58, 38.1] | 7.5 [4.0, 30.0] | 100.0 [100.0, 100.0] |
| Observation | 100.0 [98.8, 100.0] | 100.0 [94.7, 100.0] | 100.0 [100.0, 100.0] |
| Observation-unit | 100.0 [0, 100.0] | 90.1 [0.0, 100.0] | 97.8 [44.4, 100.0] |
| Observation-value | 50.0 [50.0, 100.0] | 89.7 [73.9, 100.0] | 100.0 [100.0, 100.0] |
| Procedure occurrence | 62.3 [51.8, 92.0] | 32.4 [20.8, 90.0] | 100.0 [97.5, 100.0] |
| Visit occurrence | 100.0 [100.0, 100.0] | 100.0 [100.0, 100.0] | 100.0 [100.0, 100.0] |

*Above 90% of condition, drug, visit, observation mapping were conducted already!*

# Results
## *– Mapping*

### Summary results of mapping

| **Domain** | **Mapping codes / source codes** | **Mapped records / total records** | **Mapped as standard / Mapped records** |
|---|---|---|---|
| | Median [Q1, Q3] | Median [Q1, Q3] | Median [Q1, Q3] |
| Condition occurrence | 97.5 [88.8, 99.5] | 99.5 [94.1, 100.0] | 100.0 [99.1, 100.0] |
| **Device exposure** | **58.7 [49.1, 83.1]** | **77.8 [66.1, 94.7]** | **79.9 [54.9, 100.0]** |
| Drug exposure | 83.8 [73.9, 90.1] | 96.2 [94.7, 98.0] | 98.3 [97.6, 99.0] |
| **Measurement** | **50.1 [24.8, 84.0]** | **92.4 [65.7, 99.5]** | **100.0 [99.7, 100.0]** |
| Measurement-unit | 96.2 [50.7, 100.0] | 96.7 [35.8, 100.0] | 100.0 [98.4, 100.0] |
| **Measurement-value** | **11.4 [0.58, 38.1]** | **7.5 [4.0, 30.0]** | **100.0 [100.0, 100.0]** |
| Observation | 100.0 [98.8, 100.0] | 100.0 [94.7, 100.0] | 100.0 [100.0, 100.0] |
| Observation-unit | 100.0 [0, 100.0] | 90.1 [0.0, 100.0] | 97.8 [44.4, 100.0] |
| Observation-value | 50.0 [50.0, 100.0] | 89.7 [73.9, 100.0] | 100.0 [100.0, 100.0] |
| **Procedure occurrence** | **62.3 [51.8, 92.0]** | **32.4 [20.8, 90.0]** | **100.0 [97.5, 100.0]** |
| Visit occurrence | 100.0 [100.0, 100.0] | 100.0 [100.0, 100.0] | 100.0 [100.0, 100.0] |

*Measurement, Device and procedure would be a good target of the next mapping*

# Results

**Drug mapping**

**Branded: ~ 60 %**

**Ingredient: ~ 40%**

| Vocabulary | Classification | N of records | Mapped records / Drug records, Mean ± SD |
|---|---|---|---|
| AMT | Substance | 13,335 | 0.0 ± 0.0 |
| ATC | ATC 2nd | 860,982 | 0.0 ± 0.1 |
| | ATC 3rd | 1,981,429 | 0.1 ± 0.1 |
| | ATC 4th | 10,767,782 | 0.4 ± 0.4 |
| | ATC 5th | 15,594,137 | 0.5 ± 0.8 |
| EDI | Drug Product | 5,319,678 | 0.2 ± 0.8 |
| HCPCS | HCPCS | 90 | 0.0 ± 0.0 |
| NDFRT | Pharma Preparation | 592 | 0.0 ± 0.0 |
| RxNorm (Extension) | Brand Name | 12,621 | 0.0 ± 0.0 |
| | Branded Drug | 972,954,865 | 31.8 ±26.6 |
| | Branded Drug Box | 1,265 | 0.0 ± 0.0 |
| | Branded Drug Comp | 6,247,648 | 0.1 ± 0.2 |
| | Branded Drug Form | 145,530,254 | 0.8 ± 2.3 |
| | Branded Form | 362,260 | 0.0 ± 0.2 |
| | Branded Pack | 405,282 | 0.4 ± 0.7 |
| | Clinical Dose Group | 31 | 0.0 ± 0.0 |
| | Clinical Drug | 745,935,213 | 21.4 ± 25.5 |
| | Clinical Drug Box | 61,761 | 0.0 ± 0.0 |
| | Clinical Drug Comp | 46,549,865 | 1.6 ± 6.2 |
| | Clinical Drug Form | 149,604,612 | 1.9 ± 2.4 |
| | Clinical Pack | 32,272 | 0.0 ± 0.0 |
| | Dose Form | 438,734 | 0.0 ± 0.1 |
| | Ingredient | 181,837,620 | 6.8 ± 16.6 |
| | Marketed Product | 508,260,612 | 4.5 ± 7.8 |
| | Precise Ingredient | 79,856 | 0.0 ± 0.0 |
| | Quant Branded Box | 145,461 | 0.0 ± 0.0 |
| | Quant Branded Drug | 541,892,481 | 15.5 ± 14.6 |
| | Quant Clinical Drug | 322,553,181 | 8.3 ± 12.3 |
| SNOMED | Pharma/Biol Product | 3,601,700 | 0.4 ± 2.0 |
| VA Product | VA Product | 96 | 0.0 ±0.0 |
| Undefined | Undefined | 52,699,439 | 2.5 ± 2.8 |

# Results

*– Achilles results*

## Achilles heel results

| Site, n | Type | | |
|---|---|---|---|
| | **Error**, median [Q1-Q3] | **Notification**, median [Q1-Q3] | **Warning**, median [Q1-Q3] |
| **22** | **3.0 [0.3-15.0]** | **8.0 [7.0-8.0]** | **20.0 [17.0-21.0]** |

## Common errors : 35% related to the observation period

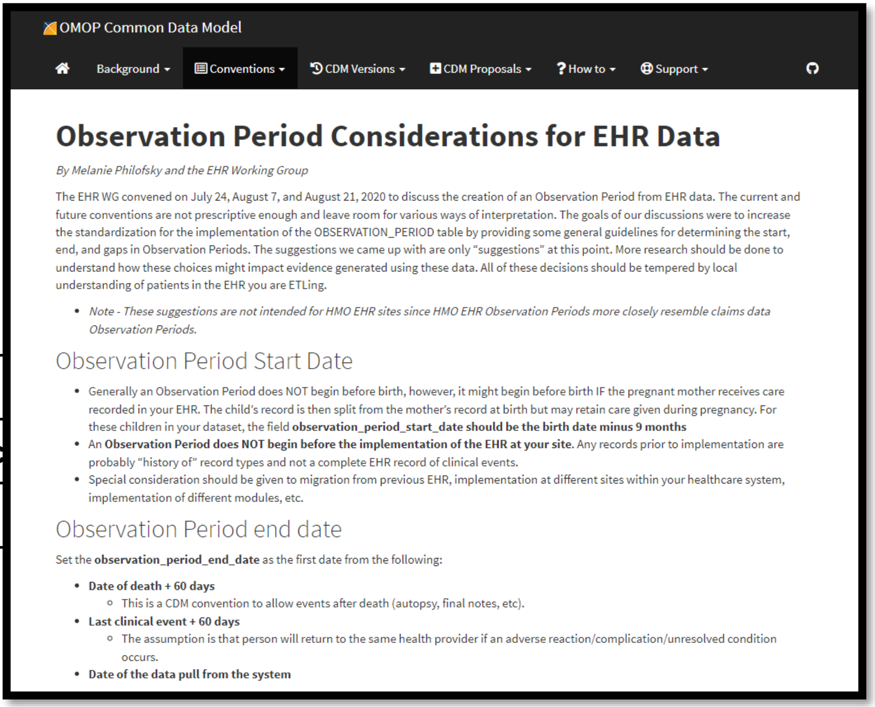
| Name | Counts |
|---|---|
| ERROR: 103 - Distribution of age at first **observation period** (count = 1); min value should not be negative | 6 |
| ERROR: 410-Number of condition occurrence records outside valid **observation period**; count (n=1245850) should not be > 0 | 6 |
| ERROR: 600-Number of persons with at least one procedure occurrence, by procedure_concept_id; concepts in data are not in correct vocabulary | 6 |
| ERROR: 710-Number of drug exposure records outside valid **observation period**; count (n=10842) should not be > 0 | 6 |
| ERROR: 101-Number of persons by age, with age at first **observation period**; should not have age < 0, (n=29) | 5 |
| ERROR: 114-Number of persons with **observation period** before year-of-birth; count (n=111) should not be > 0 | 5 |
| ERROR: 301-Number of providers by specialty concept_id; 2 concepts in data are not in correct vocabulary (Specialty) | 5 |
| ERROR: 810-Number of observation records outside valid **observation period**; count (n=2124) should not be > 0 | 5 |
| ERROR: 814-Number of observation records with no value (numeric, string, or concept); count (n=149329) should not be > 0 | 5 |
| ERROR: 8-Number of persons with invalid location_id; count (n=162923) should not be > 0 | 5 |

# Results

## – *Achilles results*

### Achilles heel results

| Site, n | Error, median [Q1-Q3] | Notific |
|---|---|---|
| 22 | 3.0 [0.3-15.0] | |



## Observation Period Considerations for EHR Data

*By Melanie Philofsky and the EHR Working Group*

The EHR WG convened on July 24, August 7, and August 21, 2020 to discuss the creation of an Observation Period from EHR data. The current and future conventions are not prescriptive enough and leave room for various ways of interpretation. The goals of our discussions were to increase the standardization for the implementation of the OBSERVATION_PERIOD table by providing some general guidelines for determining the start, end, and gaps in Observation Periods. The suggestions we came up with are only “suggestions” at this point. More research should be done to understand how these choices might impact evidence generated using these data. All of these decisions should be tempered by local understanding of patients in the EHR you are ETLing.

- *Note - These suggestions are not intended for HMO EHR sites since HMO EHR Observation Periods more closely resemble claims data Observation Periods.*

### Observation Period Start Date

- Generally an Observation Period does NOT begin before birth, however, it might begin before birth IF the pregnant mother receives care recorded in your EHR. The child’s record is then split from the mother’s record at birth but may retain care given during pregnancy. For these children in your dataset, the field **observation_period_start_date should be the birth date minus 9 months**
- An **Observation Period does NOT begin before the implementation of the EHR at your site.** Any records prior to implementation are probably “history of” record types and not a complete EHR record of clinical events.
- Special consideration should be given to migration from previous EHR, implementation at different sites within your healthcare system, implementation of different modules, etc.

### Observation Period end date

Set the **observation_period_end_date** as the first date from the following:

- **Date of death + 60 days**
  - This is a CDM convention to allow events after death (autopsy, final notes, etc).
- **Last clinical event + 60 days**
  - The assumption is that person will return to the same health provider if an adverse reaction/complication/unresolved condition occurs.
- **Date of the data pull from the system**

### Common errors

| Name | Counts |
|---|---|
| ERROR: 103 - Distribution of age at first **observation period** (count = 1); min value should not be negative | 6 |
| ERROR: 410-Number of condition occurrence records outside valid **observation period**; count (n=1245850) should not be > 0 | 6 |
| ERROR: 600-Number of persons with at least one procedure occurrence, by procedure_concept_id; concepts in data are not in correct vocabulary | 6 |
| ERROR: 710-Number of drug exposure records outside valid **observation period**; count (n=10842) should not be > 0 | 6 |
| ERROR: 101-Number of persons by age, with age at first **observation period**; should not have age < 0, (n=29) | 5 |
| ERROR: 114-Number of persons with **observation period** before year-of-birth; count (n=111) should not be > 0 | 5 |
| ERROR: 301-Number of providers by specialty concept_id; 2 concepts in data are not in correct vocabulary (Specialty) | 5 |
| ERROR: 810-Number of observation records outside valid **observation period**; count (n=2124) should not be > 0 | 5 |
| ERROR: 814-Number of observation records with no value (numeric, string, or concept); count (n=149329) should not be > 0 | 5 |
| ERROR: 8-Number of persons with invalid location_id; count (n=162923) should not be > 0 | 5 |

# Discussion

*– Lessons*

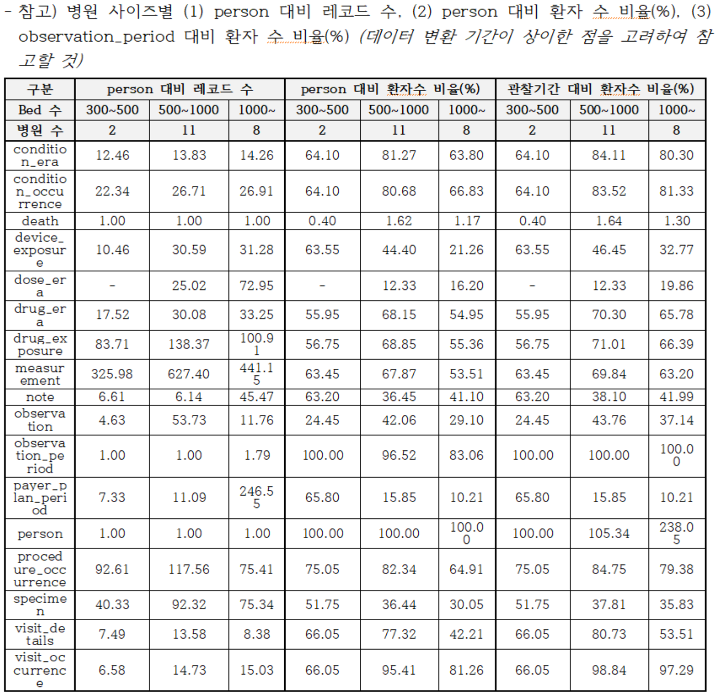
- This was a first study collecting detailed summary of OMOP-CDM and investigating the macroscopic aspects in the AP region.
- In this study, the summary of data, the mapping status and quality of data could be estimated by collecting inspection reports on the OMOP-CDM database of about 27 institutions.

- This can provide us not only insight on data quality but also giving us a reference that can help other sites, especially new institution who want to do conversion their data to CDM.

# Discussion

## – *Lessons*

- In Korea chapter, we are using the CDM inspection report for the quality consulting of the OMOP-CDM.
- Based on the collected results, now it is possible to provide a reference range to new institutions for CDM conversion.



- 참고) 병원 사이즈별 (1) person 대비 레코드 수, (2) person 대비 환자 수 비율(%), (3) observation_period 대비 환자 수 비율(%) *(데이터 변환 기간이 상이한 점을 고려하여 참고할 것)*

| 구분 | person 대비 레코드 수 | | | person 대비 환자수 비율(%) | | | 관찰기간 대비 환자수 비율(%) | | |
|---|---|---|---|---|---|---|---|---|---|
| Bed 수 | 300~500 | 500~1000 | 1000~ | 300~500 | 500~1000 | 1000~ | 300~500 | 500~1000 | 1000~ |
| 병원 수 | 2 | 11 | 8 | 2 | 11 | 8 | 2 | 11 | 8 |
| condition_era | 12.46 | 13.83 | 14.26 | 64.10 | 81.27 | 63.80 | 64.10 | 84.11 | 80.30 |
| condition_occurrence | 22.34 | 26.71 | 26.91 | 64.10 | 80.68 | 66.83 | 64.10 | 83.52 | 81.33 |
| death | 1.00 | 1.00 | 1.00 | 0.40 | 1.62 | 1.17 | 0.40 | 1.64 | 1.30 |
| device_exposure | 10.46 | 30.59 | 31.28 | 63.55 | 44.40 | 21.26 | 63.55 | 46.45 | 32.77 |
| dose_era | - | 25.02 | 72.95 | - | 12.33 | 16.20 | - | 12.33 | 19.86 |
| drug_era | 17.52 | 30.08 | 33.25 | 55.95 | 68.15 | 54.95 | 55.95 | 70.30 | 65.78 |
| drug_exposure | 83.71 | 138.37 | 100.91 | 56.75 | 68.85 | 55.36 | 56.75 | 71.01 | 66.39 |
| measurement | 325.98 | 627.40 | 441.15 | 63.45 | 67.87 | 53.51 | 63.45 | 69.84 | 63.20 |
| note | 6.61 | 6.14 | 45.47 | 63.20 | 36.45 | 41.10 | 63.20 | 38.10 | 41.99 |
| observation | 4.63 | 53.73 | 11.76 | 24.45 | 42.06 | 29.10 | 24.45 | 43.76 | 37.14 |
| observation_period | 1.00 | 1.00 | 1.79 | 100.00 | 96.52 | 83.06 | 100.00 | 100.00 | 100.00 |
| payer_plan_period | 7.33 | 11.09 | 246.55 | 65.80 | 15.85 | 10.21 | 65.80 | 15.85 | 10.21 |
| person | 1.00 | 1.00 | 1.00 | 100.00 | 100.00 | 100.00 | 100.00 | 105.34 | 238.05 |
| procedure_occurrence | 92.61 | 117.56 | 75.41 | 75.05 | 82.34 | 64.91 | 75.05 | 84.75 | 79.38 |
| specimen | 40.33 | 92.32 | 75.34 | 51.75 | 36.44 | 30.05 | 51.75 | 37.81 | 35.83 |
| visit_details | 7.49 | 13.58 | 8.38 | 66.05 | 77.32 | 42.21 | 66.05 | 80.73 | 53.51 |
| visit_occurrence | 6.58 | 14.73 | 15.03 | 66.05 | 95.41 | 81.26 | 66.05 | 98.84 | 97.29 |

# Discussion

*– Limitations*

- Although data quality improvement is continuously being made; it was a result evaluated at a specific time point (cross sectional).

- Because most of the results were from South Korea, it may not be appropriate to apply  to other countries.

- Due to the limited number of reports from claims data, it was not possible to compare them sufficiently with EMR database.

# Conclusion

- Appropriate quality of OMOP-CDM is directly related to the quality of the real-world evidence, so continuous quality management is extremely required.

- In order to improve data quality, considering the macroscopic aspect was helpful.

- It is painful to disclose our information to others, but it has greatly helped improve the quality through discussion. This efforts must be continue.

OHDSI

Observational Health Data Sciences and Informatics

# Thank you for listening!