Siriraj Informatics and
Data Innovation Center

Mahidol University
Faculty of Medicine
Siriraj Hospital

Lessons Learned Adopting OHDSI/OMOP

at Siriraj Hospital

Natthawut ‘Max’ Adulyanukosol
Deputy Director of Siriraj Informatics and Data Innovation Center (SiData+)

24 April 2024

Mahidol University
Faculty of Medicine
Siriraj Hospital

Siriraj's Data and Challenges

pre-OMOP

EHR data over 20 years in a centralized data warehouse

- 2.5M total patients
- 49M clinic visits
- 1.9M admissions
- 140M diagnoses
- 235M lab
- 104M drug events

Can't let researchers access raw EHR data
(privacy, security, messiness)

2º Data Usage
- Research
- Data Analytics
(mainly Tableau)

Irreproducible research,
Less trust in results

Can't effectively share analytics scripts/ dashboards

Network research studies 😫

Takes time to prepare data
(inclusion/exclusion criteria + formatting)

Takes time to analyze data

Most of the time researchers want quick feasibility assessment

Siriraj believes OMOP CDM can help

~~Can't let researchers access raw EHR data (privacy, security, messiness)~~

Structured data structure with proper de-identification

OMOP CDM

ATLAS helps cohort generation
i2b2 helps feasibility assessment
GenAI helps SQL code generation

2º Data Usage
- Research
- Data Analytics
(mainly Tableau)

Sharable analytics scripts/dashboards

Reproducible research, More trust in results

Use publicly available codes: Official HADES R libraries, Community contribution

Network research studies 😍

~~Takes time to prepare data (inclusion/exclusion criteria + formatting)~~

~~Takes time to analyze data~~

~~Most of the time researchers want quick feasibility assessment~~

OMOP CDM vs HL7 FHIR

They address different purposes

Observational Medical Outcomes Partnership Common Data Model (OMOP CDM)	vs	Health Level International (HL7) Fast Healthcare Interoperability Resources (FHIR)
Store data and enable reproducible observational retrospective studies	Main Purpose	Exchange health information for services
Relational Database 🗃️	Data format	JSON Document 🖨️
Researchers RWD Surveillance	Target Users	Clinicians Providers
2009	1st released	2014

Siriraj Data Team

Mahidol University
Faculty of Medicine
Siriraj Hospital

Siriraj Timeline

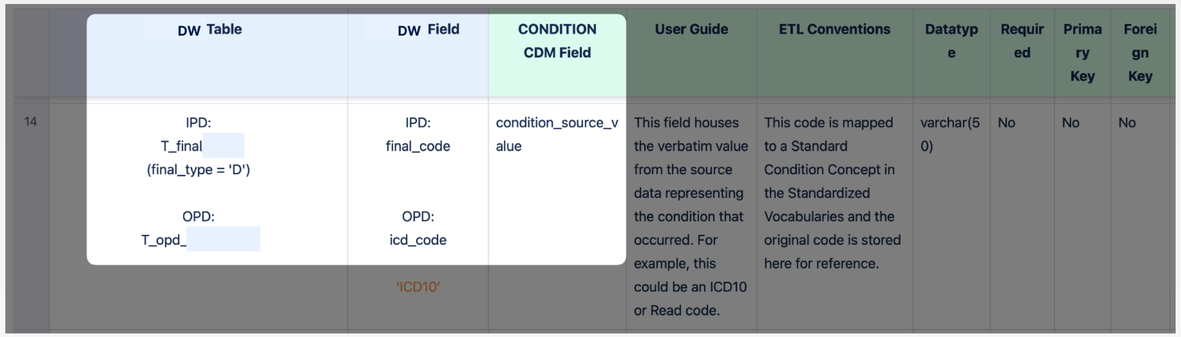
We did not spend full efforts throughout the years, i.e., Timeline could be shortened

1. 2020: Found OMOP CDM
2. 2021: Joined ETL workshop by OHDSI APAC (Mui, Seng Chan, Selva, Jing)
3. 2021: Mapping our data warehouse structure to OMOP CDM (Data Scientists + Data Analysts; mapping docs)
4. 2022: Data Transformation v1 (Data engineers; work presented at OHDSI Symposium 2022)
5. 2023: Data Transformation v2 (rewrote the conversion logic from scratch) + Data anonymization (SANT)
6. 2024
    - 6.1. Data Transformation v3 (migrate from dbt to SQLMesh)
    - 6.2. Prototypes with research projects (Clinical researchers)
    - 6.3. Secure Research Environment
    - 6.4. Atlas Deployment
    - 6.5. More code mapping

Mahidol University
Faculty of Medicine
Siriraj Hospital

Data Conversion into OMOP CDM

1. **Mapping Documentation:** Well-structured Thoughts from deep understanding in
    1.1. Source data structure
    1.2. OMOP CDM
    1.3. Clinical meaning behind data
    1.4. Expectation of researchers



	DW Table	DW Field	CONDITION CDM Field	User Guide	ETL Conventions	Datatype	Required	Primary Key	Foreign Key
14	IPD: T_final (final_type = 'D') OPD: T_opd_	IPD: final_code OPD: icd_code 'ICD10'	condition_source_value	This field houses the verbatim value from the source data representing the condition that occurred. For example, this could be an ICD10 or Read code.	This code is mapped to a Standard Condition Concept in the Standardized Vocabularies and the original code is stored here for reference.	varchar(50)	No	No	No

2. **Mapping Code:** Well-structured SQL Scripts for ETL
    2.1. Extraction
    2.2. Transformation
    2.3. Loading

ETL Tool presented
at OHDSI Global Symposium 2022
(see next slide for the poster)

dbt

Currently migrating to SQLMesh
and will be sharing lessons learned

SQLMesh

Using dbt—a free and open-source software framework—to transform data into OMOP CDM in the ETL process

PRESENTER: Thanapat 'Thane' Pitchayarat
thanapat.pit@mahidol.edu

"An organized approach to build a maintainable ETL pipeline for the OMOP CDM with minimal cost while keeping our data engineers sane 😍"

INTRO:

- The conversion of medical data into the OMOP CDM format requires a managed data engineering pipeline commonly referred to as the extract, transform, and load (ETL) process.
- The main transformation tasks in a typical OMOP CDM conversion include combining data from multiple sources, changing the original data models to match the OMOP CDM, retrieving the concept IDs of source values, and mapping the source concept IDs to the standard IDs.
- The complexity of the data transformation SQL scripts may grow rapidly beyond manageable. To keep the ETL pipeline maintainable, Siriraj Hospital uses dbt™ to transfrom its data to the OMOP CDM.
- dbt™ (shortened from data build tool) is a free and open-source software (FOSS) framework available at https://www.getdbt.com. It could be applied to data transformation at other institutions.

METHODS:

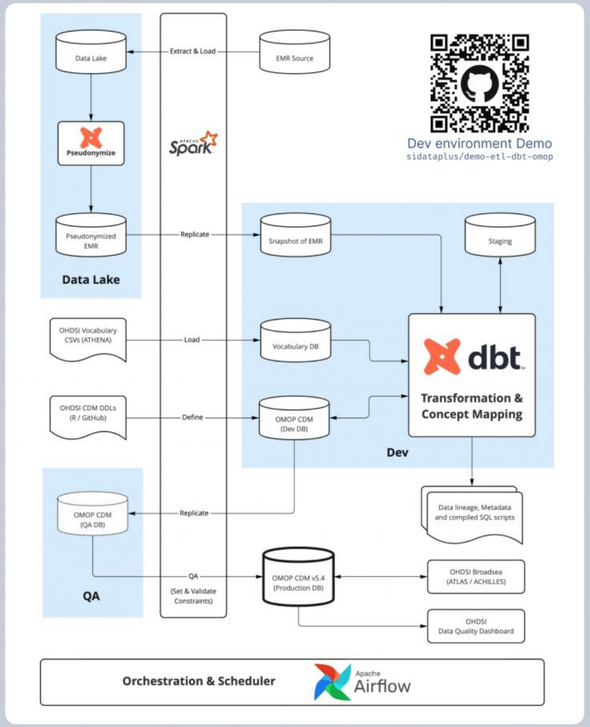
The data conversion pipeline at Siriraj Hospital can be summarized as:

1. Extraction of data from hospital sources with Apache Spark
2. Load the data into data lake and Development enviroment with Apache Spark
3. Transform the data to match the OMOP CDM specifications with dbt
4. Load the OMOP CDM-ed data into QA and Production environments

Each step is containerized with Docker. All steps are ochestrated and scheduled by Apache Airflow. Codes are version controlled with GitHub.

dbt:

- dbt comes with a command-line interface with commands that compile SQL scripts and execute the code on the connected database engines, as well as a graphical user interface.
- The core library of dbt is a Python package that supplements traditional SQL scripts with Pythonic Jinja templating.
- With the Jinja templating,
  - any frequently used SQL command can be packaged as a modular macro that can take parameters similar to a Python function, and;
  - the Jinja tags enable data lineage tracking that can be visualized on an interactive web application generated by dbt command. The web application referred to as dbt documentation also presents metadata, such as table & field descriptions, data testing conditions, upstream and downstream tables. The metadata are partly generated automatically and can be added manually as YAML files.
- To verify data quality, dbt can run automated tests during transformation execution or on demand.
- Given the popularity of dbt in the enterprise analytics space, there are many tools that can be integrated with dbt, namely Airflow for data pipeline orchestration, GreatExpectations for data quality, and DataHub for data catalog.



Simplified architectural diagram of the OMOP CDM conversion pipeline at Siriraj Hospital.

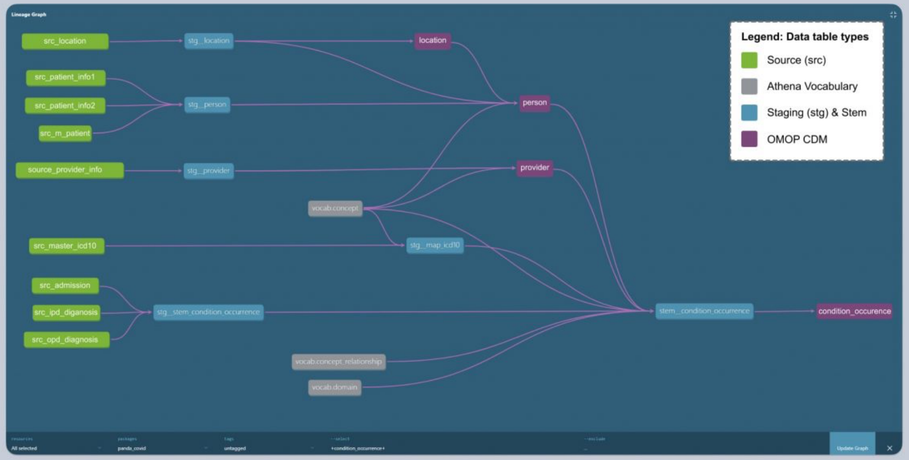


Table data lineage automatically generated by dbt. Each node represents a table or a view of data. Each linking edge represents a data flow from the source(s) to its destination(s), with data transformation in between. Each of the data transformation step is programmed as an SQL SELECT script.

Full paper

(a)

```sql
-- dbt_project/models/cdm/PERSON.sql

SELECT
  person.patient_id AS person_id,
  gender_concept.concept_id AS gender_concept_id,
  -- ...,
  race_concept.concept_id AS race_concept_id,
  -- ...,
  -- the rest of SELECT statement omitted for brevity
  -- please refer to OMOP CDM PERSON table for CDM fields
FROM {{ ref('stg__person') }} AS person
{{ map_concept(cdm_table='person', concept_code_field='gender_concept_code',
               vocabulary_id='gender') }}
{{ map_concept(cdm_table='person', concept_code_field='race_concept_code',
               vocabulary_id='race') }}
```

(b)

```sql
-- dbt_project/macros/map_concept.sql

{%- macro map_concept(cdm_table="", concept_code_field="", vocabulary_id="") -%}

LEFT JOIN {{ source('vocab', 'concept') }} AS {{vocabulary_id}}_concept
ON {{cdm_table}}.{{concept_code_field}} = {{vocabulary_id}}_concept.concept_code
  AND {{vocabulary_id}}_concept.vocabulary_id = '{{vocabulary_id}}'
  AND {{vocabulary_id}}_concept.standard_concept = 'S'

{%- endmacro -%}
```

Simplified SQL snippets (a) to create the CDM PERSON table with data from a staging table joined with the vocabulary concept tables via macros (b) to set a macro template for concept mapping. These SQL snippets with Jinja tags are to be compiled and submitted to the database engine by dbt.

CONCLUSION:

- dbt is a promising free and open-source software framework that massively facilitates the data conversion process into OMOP CDM.
- dbt programmatically manages the SQL transformation scripts in the ETL process, and consequently enhances the maintainability of the data pipeline.
- Data engineers with proficiency in SQL and Python could learn dbt in a few days and probably take a few weeks to implement dbt in the pipeline.

REFERENCES:

1. dbt Labs, Inc.. dbt-core [Internet]. 2022. Available from: https://github.com/dbt-labs/dbt-core
2. OHDSI. WhiteRabbit [Internet]. 2022. Available from: https://github.com/OHDSI/WhiteRabbit
3. OHDSI. Rabbit-in-a-Hat [Internet]. 2022. Available from: http://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html
4. OHDSI. Usagi [Internet]. 2022. Available from: https://github.com/OHDSI/Usagi
5. The Pallets Projects. Jinja [Internet]. 2022. Available from: https://palletsprojects.com/p/jinja/
6. Superconductive Health, Inc.. Welcome to great expectations [Internet]. 2022. Available from: https://greatexpectations.io/
7. Calogica. dbt_expectations [Internet]. 2022. Available from: https://hub.getdbt.com/calogica/dbt_expectations/0.1.2/
8. dbt Labs, Inc.. Success Stories [Internet]. 2022. Available from: https://www.getdbt.com/success-stories/
9. Apache Software Foundation. dbt Cloud Operators [Internet]. 2022. Available from: https://airflow.apache.org/docs/apache-airflow-providers-dbt-cloud/stable/operators.html
10. DataHub Project. dbt [Internet]. 2022. Available from: https://datahubproject.io/docs/generated/ingestion/sources/dbt/

Thanapat Pitchayarat, Gun Pinyo, Watcharaporn Tanchotsrinon, Somkid Khamsrimuang, Chalita Issarasittiphap, Chaiyanun Bootnumpech, Noppon Siangchin, Kanphitcha Promma, Nattachai Bovornmongkolsak, Prapat Suriyaphol, Natthawut Adulyanukosol
Siriraj Informatics and Data Innovation Center (SiData+), Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

Mahidol University
Faculty of Medicine
Siriraj Hospital

SiData
Siriraj Informatics and Data Innovation Center

OHDSI

https://www.ohdsi.org/wp-content/uploads/2022/10/2-Pitchayarat-OHDSI2022Poster-Adulyanukosol-scaled.jpg

Suggested Order of Conversion

modifiable

1. Local concept master
2. person
3. visit_occurrence
4. observation_period
5. location
6. care_site
7. provider
8. death
9. condition_occurrence
10. observation
11. procedure_occurrence
12. drug_exposure
13. condition_era
14. drug_era
15. measurement
16. cost
17. payer_plan_period

Code Mapping

	Siriraj	Standard Concepts
Condition (Diagnosis)	ICD-10-TM	SNOMED, ICDO3
Procedure	ICD-9-CM	SNOMED, CPT4, HCPCS, ICD10PCS, ICD9Proc, OPCS4
Measurement (Lab)	Thai Medicines Terminology (TMT) → SNOMED-CT structure	SNOMED, LOINC
Drug	Thai Medical Laboratory Terminology (TMLT) → LOINC code	RxNorm, RxNorm Extension, CVX

Lessons Learned

1. OMOP CDM can be daunting at first sight. But its core idea is easy to grasp. Then, its vast details will intimidate you again ...for good.
2. Data transformation/conversion requires understanding in source EHR data, OMOP CDM, clinical meaning of data, and expectations of the data for research purposes.
    2.1. The clinical and research understandings are mostly tacit knowledge.
3. Local concept codes are just fine for early internal research.
4. DQD helps basic QC, but needs willing clinical researchers to go through data quality in depth.
5. Need to engage researchers on the new way of working.
    5.1. old way: asked data analysts to prepare data into a specific format, time-consuming for both analysts and researchers, lacks granular details of data
    5.2. new way: researchers have more efficient access to granular data, GUI tools (ATLAS+i2b2) help filter wanted data, GenAI knows OMOP CDM structure well to help write code to format data as needed

What's next at Siriraj

1. Researchers Engagement
    1.1. Eating our own dog food
    1.2. Early adopters
2. Increase Organizational Capacity, e.g.,
    2.1. Johns Hopkins
    2.2. Stanford
3. Release our tools/resources publicly
    3.1. code on GitHub: http://github.com/sidataplus
    3.2. handbook: https://omop.sidata.plus

Building organizational capacity for observational research within a health system

JOHNS HOPKINS SCHOOL of MEDICINE · OHDSI

PRESENTERS: **Paul Nagy, Mary Grace Bowring**

INTRODUCTION The Johns Hopkins OHDSI research community was formed to help clinical researchers take advantage of this opportunity. We approached the institutional adoption of OHDSI as a socio-technical endeavor benefiting from social solutions and providing new technical methods.

We leverage the work of Patterson et al. to highlight the sources of influence necessary to enact effective change within an institution and enable adoption of OHDSI practices.

APPROACH Patterson describes sources of influence using the main categories of motivation and ability:

Motivation: 'Will this be worth it?'
Ability: 'Can I do this?'

These categories are subdivided into **organizational, team, and individual** levels that encompass the **six sources of influence**. Organizational ability refers to changes in the environment that allow for organizational change. Team or social ability refers to the need to find strength in numbers to enact change. Individual ability refers to the need to surpass your current skill level and develop proficiency. Organizational motivation refers to extrinsic rewards and incentives that are built into the environment or organization. Team motivation refers to peer pressure and how we can harness that for change. Individual motivation refers to making the behavior desirable.

We delineated activities implemented at one institution to support researchers in their use of OHDSI through an application of the six sources of influence model.

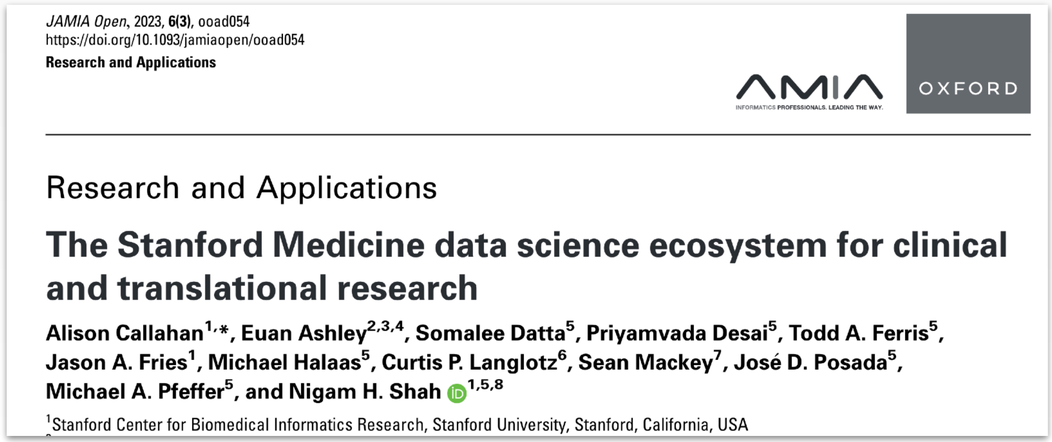
One institution's approach to **empowering** researchers **to learn and conduct observational research**

JH OHDSI adoption strategy

ORGANIZATIONAL ABILITY	TEAM ABILITY	INDIVIDUAL ABILITY
Data: Up-to-date EHR data available **Tools**: OHDSI tools **Tools**: R/Python/SQL **Support**: Clinical research core data service team	**Team science:** Teams channel with interdisciplinary group **Networking**: Partner with OHDSI institutions **Registry creation:** OMOP sub-setting	**Online training:** EHDEN Academy, office hours **Graduate courses**: Observational research, data science
ORGANIZATIONAL MOTIVATION	**TEAM MOTIVATION**	**INDIVIDUAL MOTIVATION**
Awareness: Institutional website **Leadership:** Support for OHDSI **Support**: Grant letters of support **IRB**: Enable easier process	**Peer Mentoring**: Weekly calls **Networking**: Participation in OHDSI working groups **Data Science**: Graduate student project partnering	**Data:** Get data faster **Publications**: Produce robust, reproducible publications **Grants**: Grant template library **Data**: Get multi-institutional data

We aim to accelerate the use of OHDSI by creating value for our researchers and our organization. This framework can be adopted to support clinicians and researchers as they incorporate OHDSI into their research efforts.

Mary Grace Bowring, Michael Cook, Star Lui, Khyzer Aziz, Aki Nishimura, Paul Nagy
Johns Hopkins University School of Medicine, Baltimore, US



JAMIA Open, 2023, **6(3)**, ooad054
https://doi.org/10.1093/jamiaopen/ooad054
Research and Applications

AMIA INFORMATICS PROFESSIONALS. LEADING THE WAY. · OXFORD

Research and Applications

The Stanford Medicine data science ecosystem for clinical and translational research

Alison Callahan[1,*], Euan Ashley[2,3,4], Somalee Datta[5], Priyamvada Desai[5], Todd A. Ferris[5], Jason A. Fries[1], Michael Halaas[5], Curtis P. Langlotz[6], Sean Mackey[7], José D. Posada[5], Michael A. Pfeffer[5], and Nigam H. Shah [1,5,8]

[1]Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California, USA

What's next in Thailand

OHDSI Thailand Chapter

- Activities
- Community Support
- Research
- Funding

Please reach me at natthawut.adu@mahidol.edu, 02 4141 369