APAC Scientific Forum

November 7, 2024

Agenda

- 2024 OHDSI APAC Symposium News
- Community-wide ETL Project: Project Closeout

2024 OHDSI APAC Symposium

- Collaborator Showcase
  - Acceptance notifications have been sent out on **October 30**
  - Authors of abstracts selected for posters should come to the symposium with your printed posters
  - Posters should be prepared and printed in **portrait (vertical) orientation** and no larger than **A0 (84.1 cm x 118.9 cm / 33.1 in x 46.8 in) size**
  - Authors of abstracts additionally selected for lightning talks must confirm your attendance at the symposium **by November 15**
  - Confirmed lightning talk speakers will be asked to submit your slides in advance

Registration

Flat fee of 488 SGD (~375 USD), inclusive of food and beverages

Early-bird
https://www.gevme.com/83213998/registration/buy?tickets-345330=1

Student
https://www.gevme.com/67642123/registration/buy?tickets-348888=1

Day 1 (December 4) → **Day 2-3 (December 5-6)** → **Day 4-5 (December 7-8)**

Tutorial @ NUS

Main Conference @ Marina Bay Sands

https://www.gevme.com/83213998/registration/buy?tickets-337255=1

Datathon @ NUS

- Early-bird: +150 SGD (~115 USD)
- Normal: +199 SGD (~153 USD)
- Student: +50 SGD (~38 USD)
- Inclusive of food and beverages

APAC Community-wide ETL Project

Project Closeout

Announcements

- Revised data use agreement will be distributed to all project volunteers for consent
- Please make sure to archive all project-related documentation into the project Teams channels and any other relevant repositories
- December scientific forum will be canceled due to overlap with the 2024 OHDSI APAC Symposium, new cadence will be set up in 2025

Data Analysis Team

Team **Leads** and Members

1. **Lydia Liu (PM)**
2. **Natthawut 'Max' Adulyanukosol (Lead)**
3. **Lim Boon Sheng (Lead)**
4. Yizhi Dong
5. Balachandran E.
6. Naphat 'Aut' Permpredanun
7. Shreema S Rao
8. Muhd Zulfadli Hafiz Ismail
9. Yoshihiro Aoyagi
10. Burin Boonwatcharapal
11. Satish Kumar Anbazhagan

Recap

Project Management

Data experts and CDM experts together design the ETL

Data Analysis

People with medical knowledge create the code mappings

Vocabulary Mapping

Quality Assurance

All are involved in quality control

Data Analysis

ETL Documentation

ETL

A technical person implements the ETL

White Rabbit

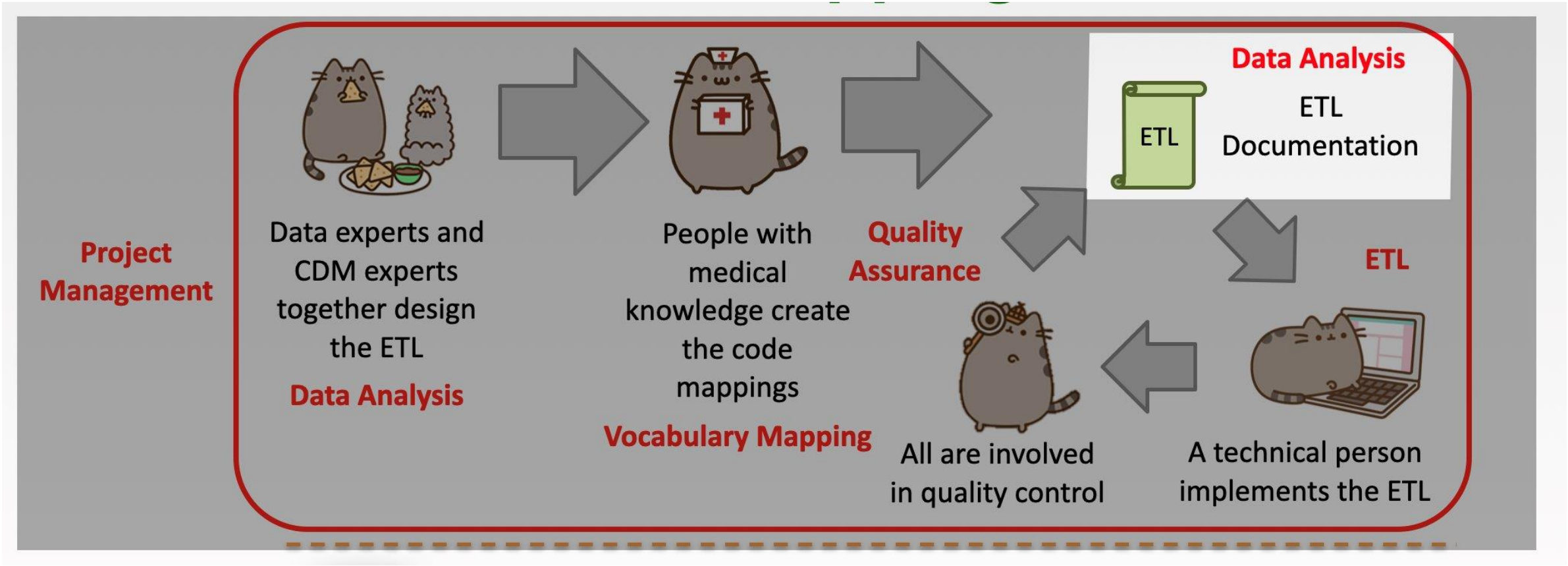
Similar function to Rabbit-in-a-Hat, but easier to use and collaborate

Claude

Suggest likely matches PASAR → OMOP;
Submitted only already public info to the AI

Tried ChatGPT-4o, but performed poorly

Recap

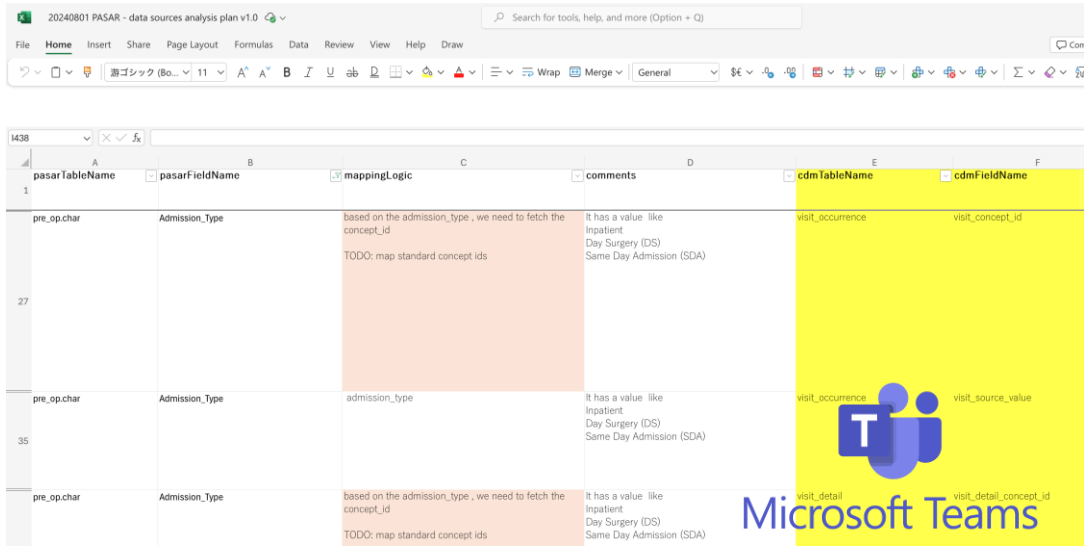


RABBIT IN A HAT

GitHub

Deliverables

- Data Analysis

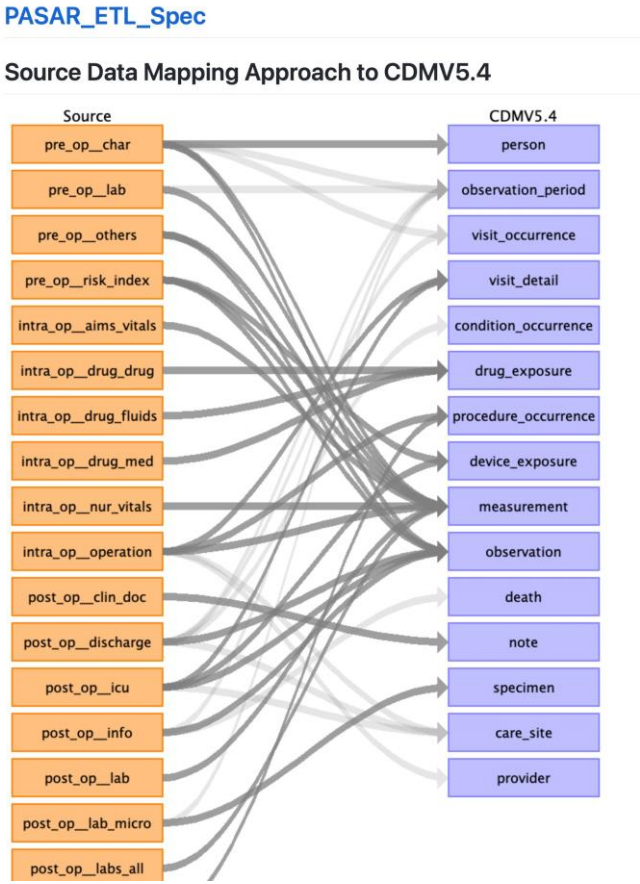


Mapping tutorial

https://www.loom.com/share/3ee567014830493b889583411a96e450?sid=e4268da5-7d22-47f4-adf6-c500093251fb

- ETL Spec

https://github.com/sidataplus/PASAR_ETL_Spec



Reflections

#	Issue	Impact	Mitigation
1.	Unavailability of certain team members due to their day work	Delay in work delivery and response to the other team members.	Close check-in with team members, and weekly cadence to get the status update.
2.	Availability and responsiveness of data owner and clinician who understand the source data to assist in data analysis	Unable to confirm the purpose of the source information to map to destination, and unable to finalize the mapping to OMOP CDM structure.	Project manager and leads close follow up with the data owner to seek and confirm the logic and rules. Suggest to include data owner as co-DA team lead in future.
3	Unavailability of the full set of the data at the beginning	Unable to perform the comprehensive data profiling, resulting in delay of conducting the data mapping and terminology.	Confirm the range of data for analysis at the beginning stage and set the expectation. Optimize the resource to prepare and loading the required data.
4	The development infrastructure readiness delay	Delay in commencing the data analysis work as per plan.	Confirm and get ready the development platform at the planning stage prior kick-start the project.
5	The approach of information sharing is not optimized as using document sharing by multiple parties, which lack of version control and notification	As the information being update, the team has to review and identify the changes and update to the ETL spec.	The clear waterfall approach with proper gate to avoid changing requirement for the downstream activities. Need better change history and version tracking control.

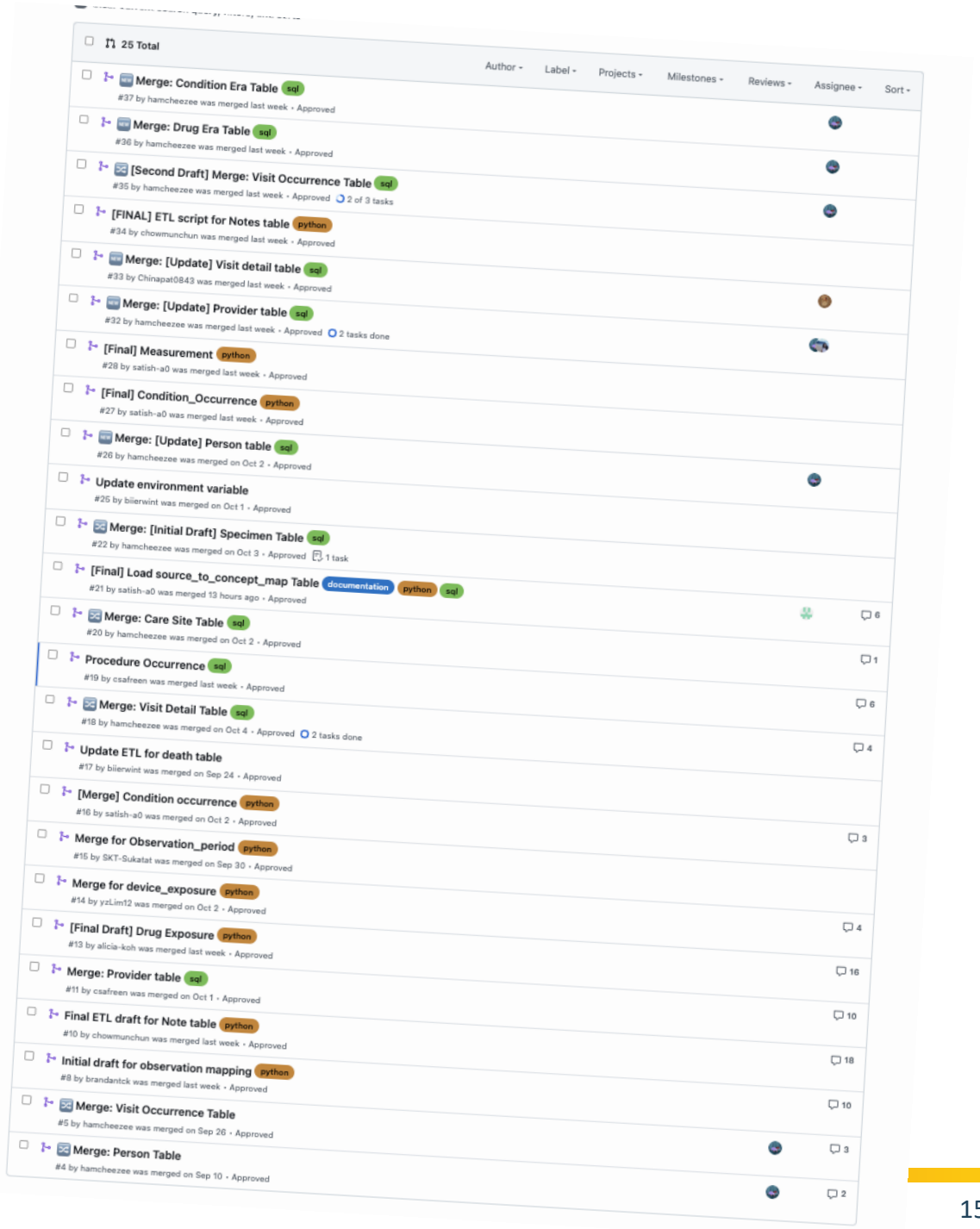
ETL Team

Team **Leads** and Members

- **Evelyn Goh**
- **Satish Kumar Anbazhagan**
- **Steven Yong**
- **Jiawei Qian**
- **Afreen Chitwadgi Sikandara**
- **Nongnaphat Wongpiyachai**
- Sornchai Manoson
- Chinapat Onprasert
- Alicia Koh
- Hengxian Jiang
- Erwin Tantoso
- Brandan Tan
- Sukatat Leknimit
- Yong Zhe Lim
- Mun Chun Chow
- Gyeol Song

Recap - Highlights

- 19 OMOP clinical tables ETL pipeline implemented
- Additionally support ingestion of 4 Vocabulary tables
- Total records ~14 million
- Total time taken ~1 hour
- 25 Pull requests merged, 1 final fixes pending merge
- 12 contributors, ~150 commits, 75 files added
- Constraints enabled except for Concept & Procedure Occurre tables – CPT4 Codes
- ~40 DQD QA issues resolved
- Final **omop_etl_schema** available for browsing



Recap – Ingestion statistics

Clinical Table	Records count	Time Taken
cdm_source	1	.08s
care_site	111	.46s
provider	872	.1s
person	999	.29s
observation_period	999	.42s
death	143	.4s
visit_occurrence	1,600	.18s
visit_detail	268	3.17s
condition_occurrence	26,464	39.9s
condition_era	4,193	1.78s
drug_exposure	26,065	2.65s
drug_era	6,793	28.5s
procedure_occurrence	10,074	.89s
device_exposure	132	1s
observation	1,161,982	105.2s
note	3,869	.99s
specimen	52,690	2.85s
measurement	12,659,065	3905s

Total Clinical Records	Total Time Taken
13,956,320	~1 hour 8 minutes

Vocabulary table	Records count	Time Taken
source_to_concept_map	3053	20s
concept	6,372,686	182.5s
concept_ancestor	77,386,059	100.4s
concept_relationship	40,883,488	115s

Deliverables

Deliverables	Artifacts
• Source code • Instructions to setup	https://github.com/satish-a0/pasar Github.pptx
Source_to_concept_map data from Vocabulary Team	https://github.com/satish-a0/pasar/pull/21
GCP Raw Source PASAR data	ohdsi_omop_2024/onepercent_data
PASAR ETL Scripts – GCP -> Cloudsql Postgres	gs-to-psql.zip
GCP Athena Vocabulary Data	ohdsi_omop_2024/vocab_2024Nov03_v5
Clarifications with Data Analysis Team	ETL Development.xlsx
• ETL Assumptions • Clarifications with Data owner	ETL Development.xlsx ETL Development.xlsx
Clarifications with Vocabulary Team	ETL Development.xlsx
QA Issues status	ETL Development.xlsx
PASAR Postgres data	Cloudsql Postgres – preop, intraop, postop schemas
Ingested OMOP v5.4 Data	Cloudsql Postgres - omop_etl_schema

Reflections

- Achieved what we wanted - Structured, flexible, collaborative yet sufficient isolation in development and a reproducible ETL setup (linux/mac os)
- Near complete Automation - From schema creation -> Concepts loading -> OMOP clinical tables ingestion in the order of dependency
- Challenging to make assumptions for ETL mapping Spec. No single right approach.
  - 1 to many
    - Maintain linkages
    - Non standard - Standard concept mapping
  - Clinical & Visit Dates correlation
  - Domain knowledge / people with experience – invaluable
- Communication with other teams critical
- Leveraging existing OMOP Resources – DDL scripts, Vocabularies, Tools

Reflections

- Scope for improvement
  - 1 to many mappings – PASAR to OMOP
    - Across different domains
    - Concepts: History of event and Heart Disorder
  - PASAR to OMOP Mappings based on source concepts
  - Revisit ETL Assumptions – Time constraints
  - Pending DQD Issues
  - Scalability
- Undoubtedly good experience gained!

Vocabulary Mapping Team

Team Leads and Members

- **Qi Yang (Lead)**
- Leong Hui Wong
- Dr. Kosuke Tanaka
- Dr. Krittaphas Chaisutyakorn
- Dr. Liying Pei
- Dr. Shigemi Matsumoto
- Dr. Cynthia Sung
- Dr. Asif Syed
- Dr. Elisabeth E. Park
- Dr. Keiko Asao
- **Lakshmi Kubendran (PM)**

Recap

- Initially, the project had four volunteers, which expanded to ten over time, though there were frequent dropouts.
- Despite being a new concept, the project brought unique learning experiences, with a core team remaining actively engaged in mapping tasks.
- Training on the USAGI tool was foundational, equipping volunteers with essential mapping skills and problem-solving approaches.
- Addressing various volunteer queries was interesting and challenging, involving communication and guiding them through the mapping process.
- Although some volunteers were less responsive, teamwork flourished when others stepped in to ensure the project was completed on time.
- Few files were identified and added in the later phase of the project but with dynamic team support, completed the mappings on time.

Deliverables

Name of the file	Assigned to	Status
intraop_aimsvitals_vitalcode	Pei Liying	Completed
intraop_drugdrug_group1	wong.LeongHui	Completed
intraop_drugmed_group1	wong.LeongHui	Completed
intraop_nurvitals_group1	Lakshmi	Completed
intraop_operation_group1.xlsx	Pei Liying, Matsumoto Tanaka, Qi, Elizabeth, Asif Syed Keiko	Completed
postop_clindoc_group1.xlsx	Qi	Completed
postop_info_group1.xlsx	wong.LeongHui	Completed
postop_lab_testdesc.xlsx	Lakshmi	Completed
postop_labmicro_antibioticname	Do not Map	
postop_labmicro_microresultedproceduredescription		
postop_labmicro_organismdescription		
postop_labsall_group1	Lakshmi	Completed
postop_pacu_group1	Kosuke Tanaka	Completed
postop_renal_group1	Asif Syed	Completed
preop_char_allergyinformation	Lakshmi	Completed
preop_char_gender	Qi	Completed
preop_char_race	Qi	Completed
preop_lab_prepoplabtestdescription	Lakshmi	Completed
preop_radiology_procedurename	Qi	Completed
Surgical Specialty	Lakshmi	Completed

Reflections

- Working with Singapore-specific surgical codes highlighted regional variations, and it is a good exercise to understand both local codes and OMOP standards.
- Some procedures lacked direct OMOP matches, requiring careful consideration and approximate mappings to bridge gaps in clinical practice differences.
- The project strengthened skills in OHDSI tools, particularly ATHENA and USAGI which are essential for effective terminology standardization and mapping.
- Mapping patient data to OMOP emphasized OMOP's value in enabling Singapore data interoperability for global research and benchmarking.


A BIG
Thank
You

Quality Assurance Team

Team **Leads** and Members

- **Santan Maddi (Lead)**
- Karthik Seetharaman (PM)

Recap

Recap

- Deployed DQD on two instances
- Provided R scripts for running DQD
- Ran and shared DQD results with different teams
- Initial troubleshooting and suggesting solutions based on results
- Co-ordinated with ETL team to fix issues

Recap

DATA QUALITY ASSESSMENT

PASAR

DataQualityDashboard Version: 2.6.1
Results generated at 2024-11-06 09:33:16 in 7 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	502	7	509	99%	291	0	291	100%	793	7	800	99%
Conformance	895	8	903	99%	137	0	137	100%	1032	8	1040	99%
Completeness	434	18	452	96%	17	0	17	100%	451	18	469	96%
Total	1831	33	1864	98%	445	0	445	100%	2276	33	2309	**99%**

1002 out of 2276 passed checks are Not Applicable, due to empty tables or fields.
4 out of 33 failed checks are SQL errors.
Corrected pass percentage for NA and Errors: 98% (1274/1303).

Deliverables

- Scripts and results – Github repo
- Sprint reviews and presentations – Sharepoint slides

Reflections

- Quality work needs a lot of co-ordination and teamwork with other teams to succeed.
- Data mapping quality is quite high (98%).
- Future work - to understand why there are a high number of non-applicable checks and how this number can be reduced.

PS: It was a great experience to be part of this well organised OHDSI APAC project.


OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Thank you!