2024 APAC ETL Project

Mui Van Zandt, Gyeol Song, Steven Yong,
Satish Kumar Anbazhagan, Santan Maddi

Agenda

- Project Overview
- Project Management Team
- Data Analysis Team
- Extract, Transform, Load (ETL) Team
- Vocabulary Mapping Team
- Quality Assurance Team

Project Overview

1 **Objectives**

- To accomplish OHDSI APAC’s 2024 OKRs
- To build ETL knowledge within the APAC community

2 **Candidate**

PASAR, a Singaporean perioperative database

3 **Method**

Remote, federated community-wide ETL

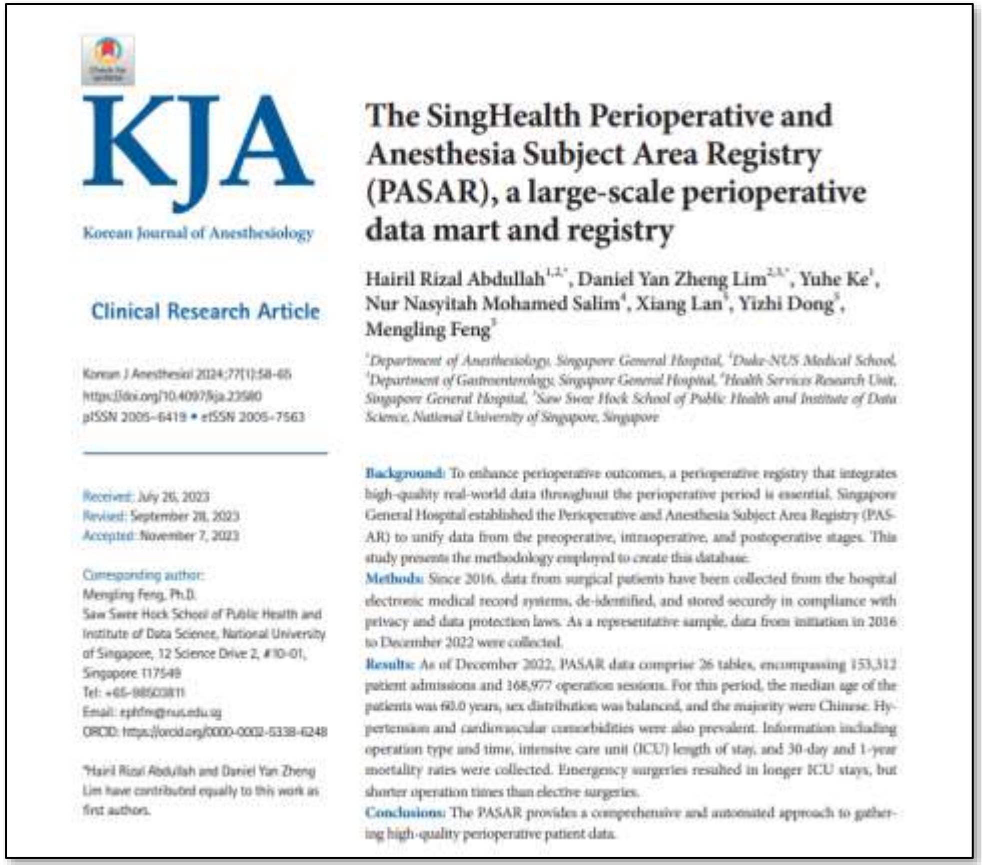
4 **Duration**

August 1 – November 7, 2024 (3.5 months, 14 weeks)

About PASAR

- Perioperative and Anesthesia Subject Area Registry (PASAR) established by Singapore General Hospital (SGH)
- Covers all patients who undergo surgery at SGH
- Consists of 153,312 admissions and 168,977 operation sessions between 2016 to December 2022

Special thanks to SGH team, especially Professor Hairil Rizal and Dr. Yuhe Ke for their support!



KJA
Korean Journal of Anesthesiology

Clinical Research Article

Korean J Anesthesiol 2024;77(1):58-65
https://doi.org/10.4097/kja.23580
pISSN 2005-6419 • eISSN 2005-7563

Received: July 26, 2023
Revised: September 28, 2023
Accepted: November 7, 2023

Corresponding author:
Mengling Feng, Ph.D.
Saw Swee Hock School of Public Health and Institute of Data Science, National University of Singapore, 12 Science Drive 2, #10-01, Singapore 117549
Tel: +65-98500811
Email: ephfm@nus.edu.sg
ORCID: https://orcid.org/0000-0002-5338-6248

*Hairil Rizal Abdullah and Daniel Yan Zheng Lim have contributed equally to this work as first authors.

The SingHealth Perioperative and Anesthesia Subject Area Registry (PASAR), a large-scale perioperative data mart and registry

Hairil Rizal Abdullah[1,2,*], Daniel Yan Zheng Lim[2,3,*], Yuhe Ke[1], Nur Nasyitah Mohamed Salim[4], Xiang Lan[5], Yizhi Dong[5], Mengling Feng[5]

[1]Department of Anesthesiology, Singapore General Hospital, [2]Duke-NUS Medical School, [3]Department of Gastroenterology, Singapore General Hospital, [4]Health Services Research Unit, Singapore General Hospital, [5]Saw Swee Hock School of Public Health and Institute of Data Science, National University of Singapore, Singapore

Background: To enhance perioperative outcomes, a perioperative registry that integrates high-quality real-world data throughout the perioperative period is essential. Singapore General Hospital established the Perioperative and Anesthesia Subject Area Registry (PASAR) to unify data from the preoperative, intraoperative, and postoperative stages. This study presents the methodology employed to create this database.

Methods: Since 2016, data from surgical patients have been collected from the hospital electronic medical record systems, de-identified, and stored securely in compliance with privacy and data protection laws. As a representative sample, data from initiation in 2016 to December 2022 were collected.

Results: As of December 2022, PASAR data comprise 26 tables, encompassing 153,312 patient admissions and 168,977 operation sessions. For this period, the median age of the patients was 60.0 years, sex distribution was balanced, and the majority were Chinese. Hypertension and cardiovascular comorbidities were also prevalent. Information including operation type and time, intensive care unit (ICU) length of stay, and 30-day and 1-year mortality rates were collected. Emergency surgeries resulted in longer ICU stays, but shorter operation times than elective surgeries.

Conclusions: The PASAR provides a comprehensive and automated approach to gathering high-quality perioperative patient data.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10834714/

Project Team

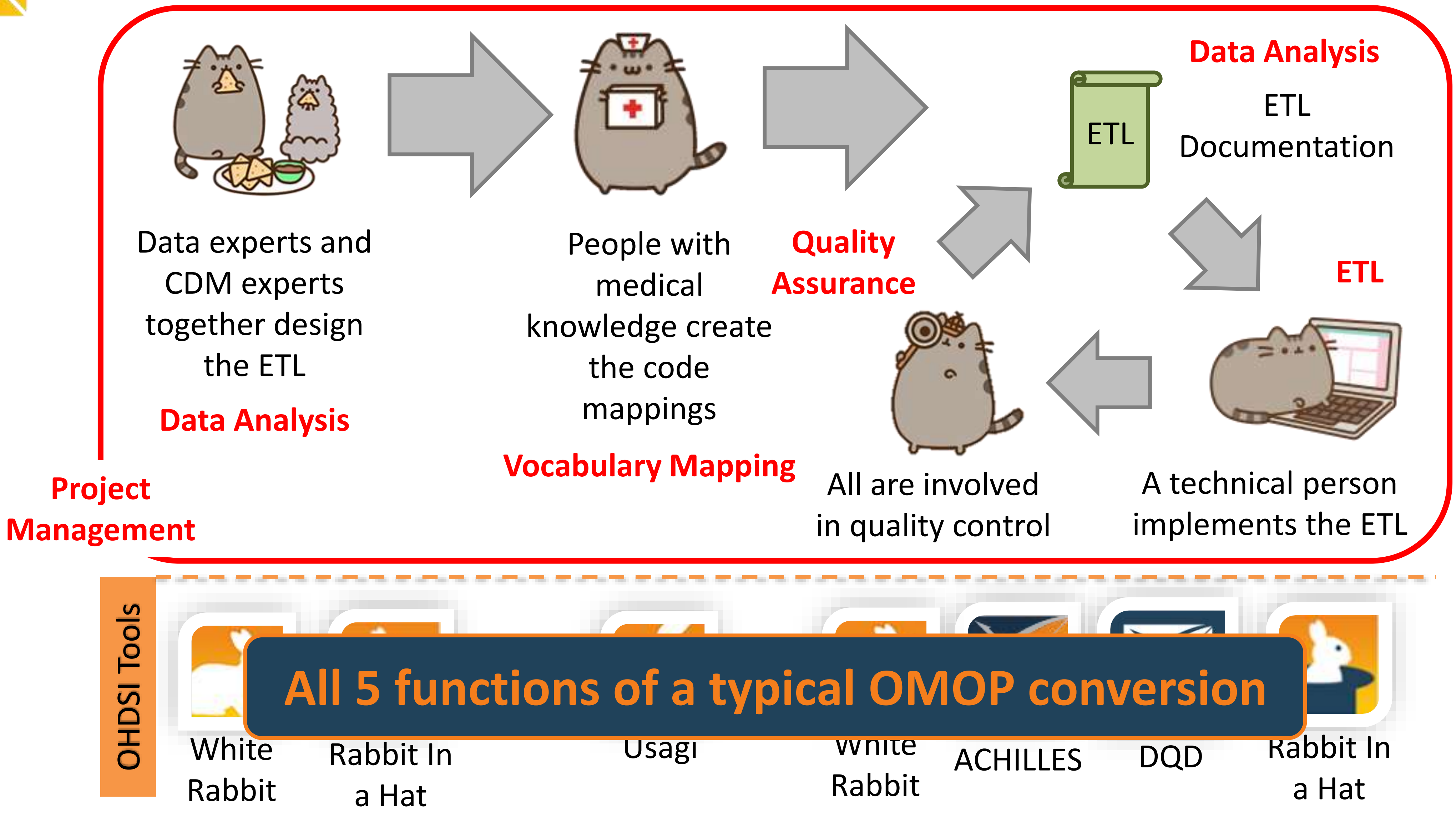
Team of 39 volunteers around the globe!

Natthawut ‘Max’ Adulyanukosol・Boon Sheng Lim・Shreema S Rao・Muhd Zulfadli Hafiz Ismail・Burin Boonwatcharapai・Naphat ‘Aut’ Permpredanun・Yoshihiro Aoyagi・Balachandran Elangovan・Yizhi Dong・Lydia Liu・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・Qi Yang・Lakshmi Kudendran・Leong Hui Wong・Kosuke Tanaka・Krittaphas Chaisutyakorn・Liying Pei・Shigemi Matsumoto・Cynthia Sung・Asif Syed・Elisabeth E. Park・Keiko Asao・Santan Maddi・Karthik Seetharaman

Special thanks to NUS team led by Professor Mengling Feng!

Project Management Team

Structure



Data experts and CDM experts together design the ETL

Data Analysis

People with medical knowledge create the code mappings

Vocabulary Mapping

Quality Assurance

Data Analysis

ETL Documentation

ETL

A technical person implements the ETL

All are involved in quality control

Project Management

OHDSI Tools

White Rabbit

Rabbit In a Hat

Usagi

White Rabbit

ACHILLES

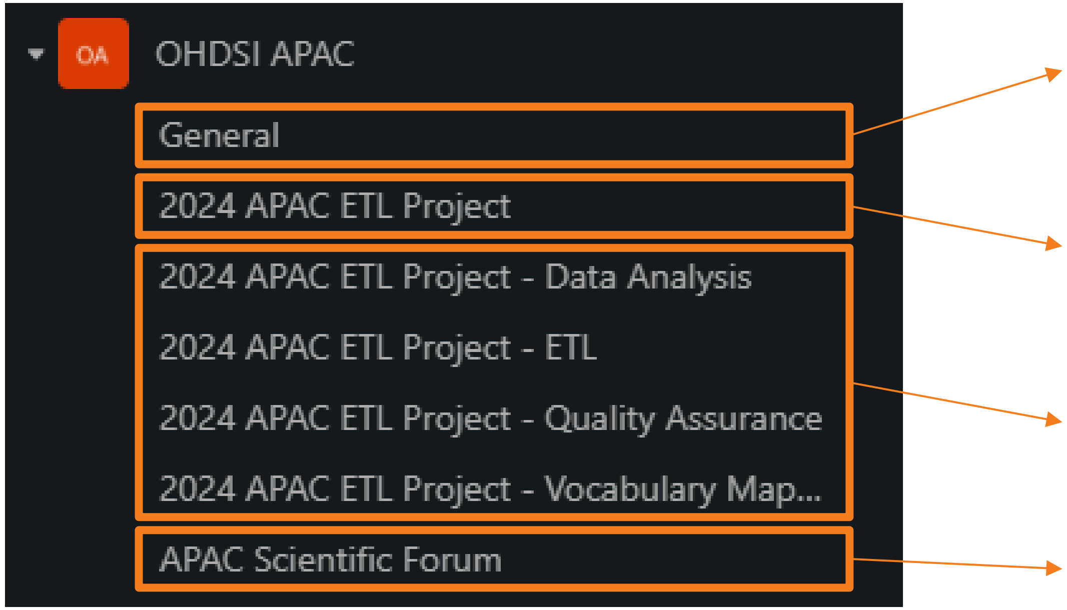
DQD

Rabbit In a Hat

All 5 functions of a typical OMOP conversion

Logistics: Environment

- Used existing/new channels under OHDSI APAC's Teams environment



OA OHDSI APAC

General

2024 APAC ETL Project

2024 APAC ETL Project - Data Analysis

2024 APAC ETL Project - ETL

2024 APAC ETL Project - Quality Assurance

2024 APAC ETL Project - Vocabulary Map...

APAC Scientific Forum

Used for general community-facing meetings (community calls)

Central channel for archiving of final project-related documentations

Team channels for archiving of working documents and hosting of individual team meetings and discussions

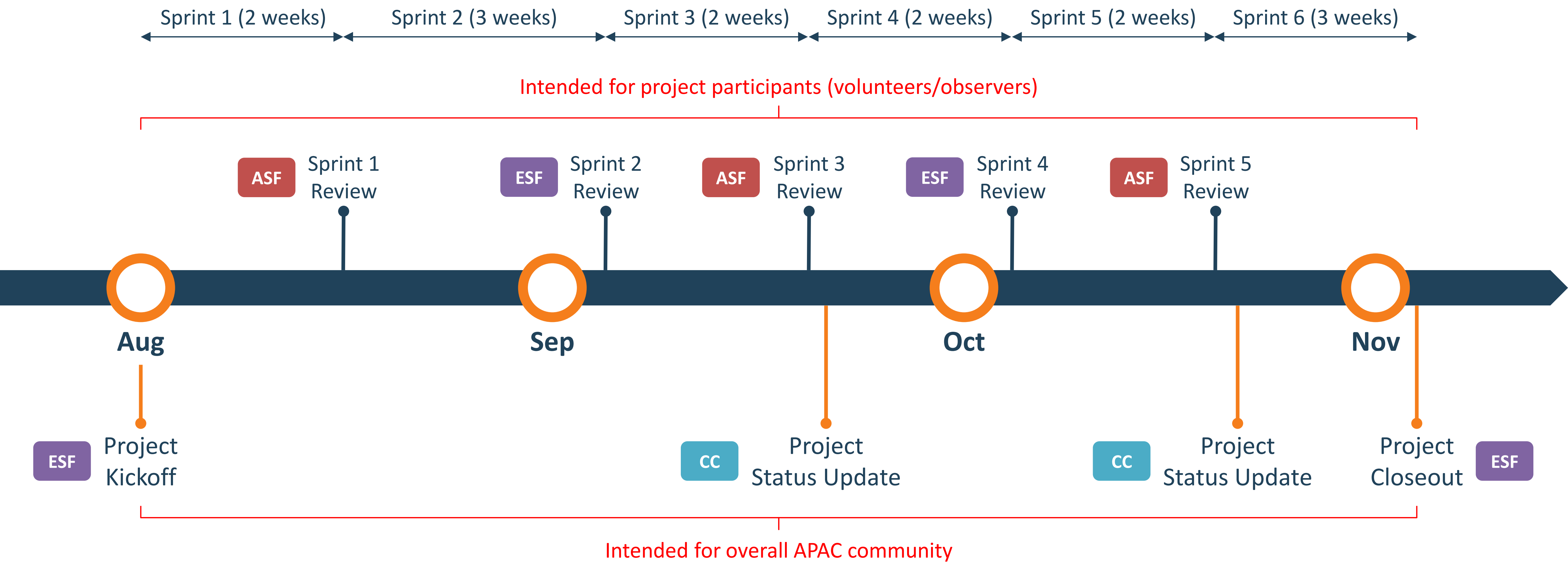
Used for sprint reviews (scientific forums)

- Public channels were used to accommodate both volunteers and observers who signed up to learn
- Also leveraged Teams chat function for direct messages and group chats

Logistics: Data

- Google Cloud Platform (GCP) was used to host data for the project
- All volunteers were requested to consent to a data use agreement before data access was granted
- 10-case sample data was initially provided due to delay in administrative processes and mainly used by the data analysis team for preparation of ETL specifications
- Unique codes from designated fields and their frequencies were extracted from the full data for the vocabulary mapping team
- ~1,000-case sample data was additional provided about a month later and mainly used by the ETL team to develop and test ETL

Timelines and Meetings



Sprint 1 (2 weeks)

Sprint 2 (3 weeks)

Sprint 3 (2 weeks)

Sprint 4 (2 weeks)

Sprint 5 (2 weeks)

Sprint 6 (3 weeks)

Intended for project participants (volunteers/observers)

ASF Sprint 1 Review

ESF Sprint 2 Review

ASF Sprint 3 Review

ESF Sprint 4 Review

ASF Sprint 5 Review

Aug

Sep

Oct

Nov

ESF Project Kickoff

CC Project Status Update

CC Project Status Update

Project Closeout ESF

Intended for overall APAC community

ESF = existing APAC Scientific Forum

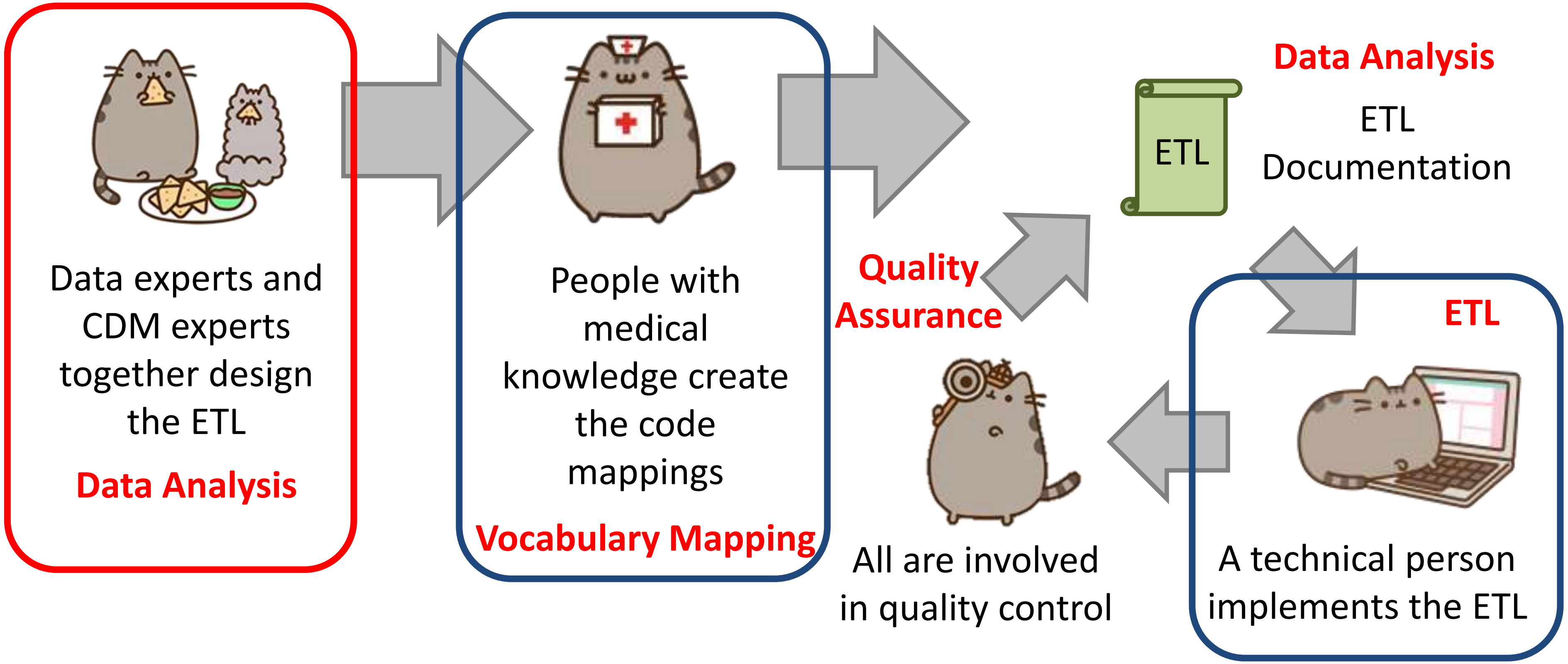
ASF = ad hoc APAC Scientific Forum

CC = APAC Community Call

Data Analysis Team

Data Analysis Process

Project Management



Data experts and CDM experts together design the ETL

Data Analysis

People with medical knowledge create the code mappings

Vocabulary Mapping

Data Analysis

ETL Documentation

ETL

A technical person implements the ETL

Quality Assurance

All are involved in quality control

OHDSI Tools

White Rabbit

Mapping Doc

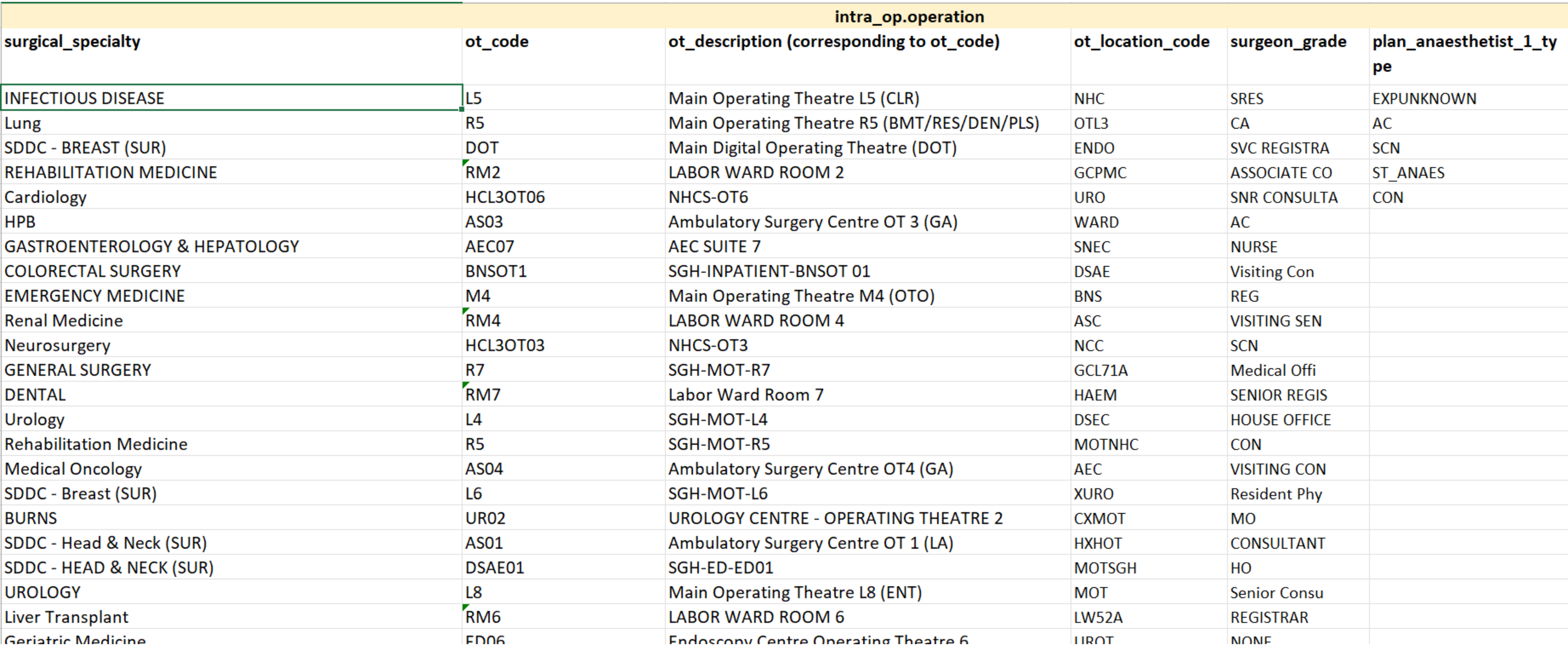
Rabbit In a Hat

Scan Output for Vocabulary Mapping

ETL Mapping Doc

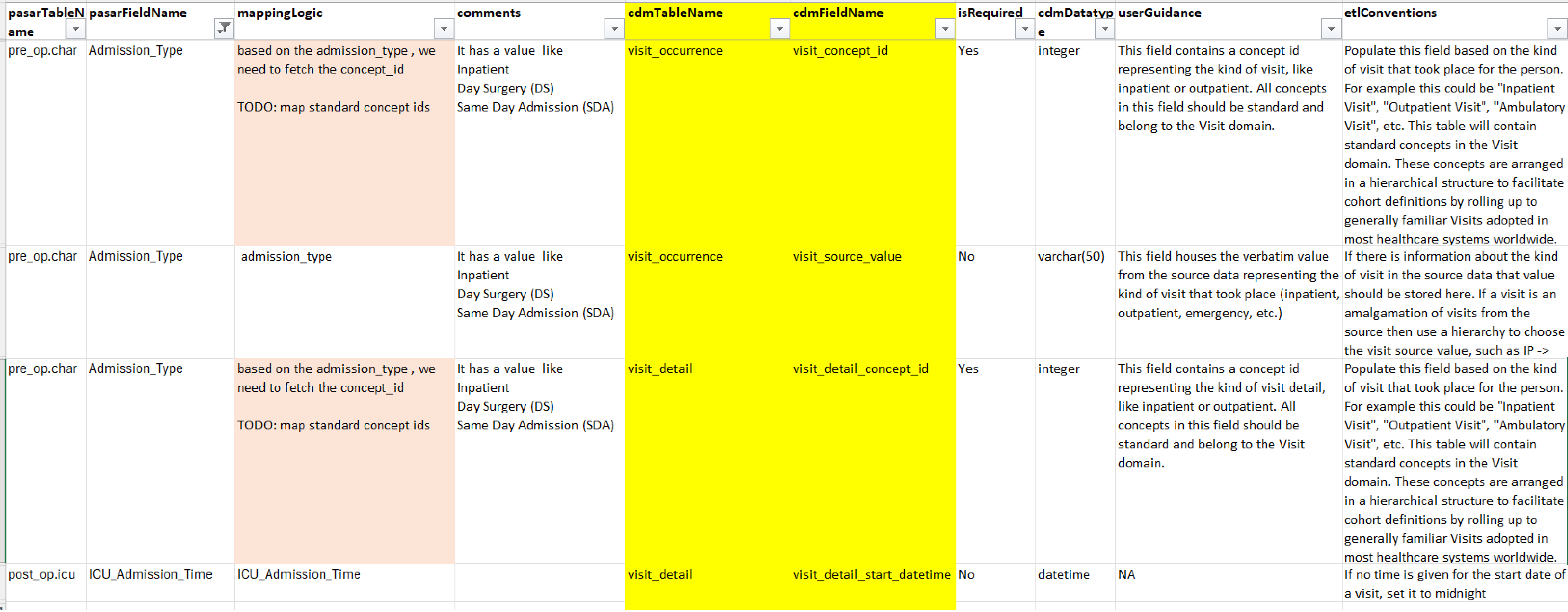
Host ETL Spec

Scan Output: Identify Distinct Values



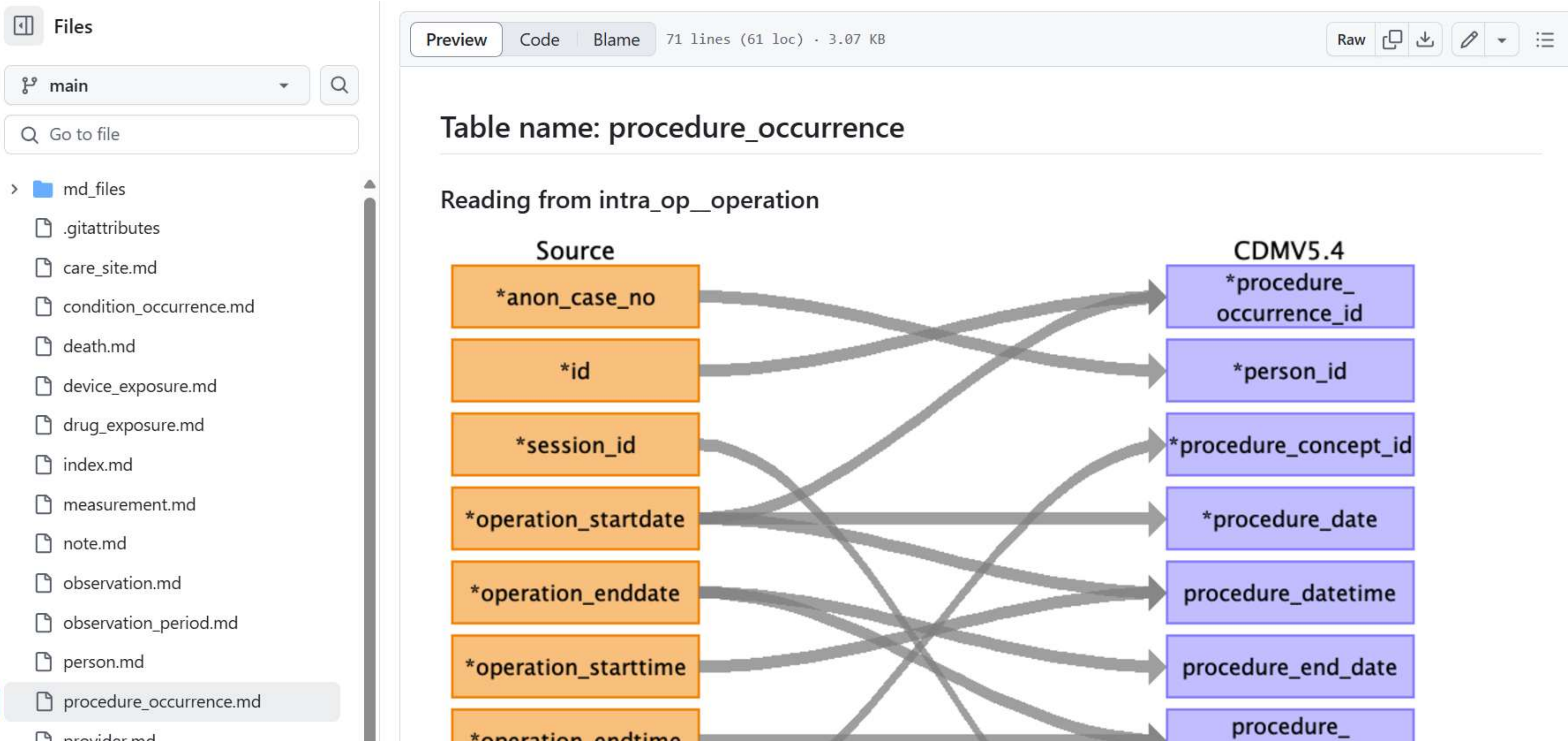
intra_op.operation					
surgical_specialty	ot_code	ot_description (corresponding to ot_code)	ot_location_code	surgeon_grade	plan_anaesthetist_1_type
INFECTIOUS DISEASE	L5	Main Operating Theatre L5 (CLR)	NHC	SRES	EXPUNKNOWN
Lung	R5	Main Operating Theatre R5 (BMT/RES/DEN/PLS)	OTL3	CA	AC
SDDC - BREAST (SUR)	DOT	Main Digital Operating Theatre (DOT)	ENDO	SVC REGISTRA	SCN
REHABILITATION MEDICINE	RM2	LABOR WARD ROOM 2	GCPMC	ASSOCIATE CO	ST_ANAES
Cardiology	HCL3OT06	NHCS-OT6	URO	SNR CONSULTA	CON
HPB	AS03	Ambulatory Surgery Centre OT 3 (GA)	WARD	AC	
GASTROENTEROLOGY & HEPATOLOGY	AEC07	AEC SUITE 7	SNEC	NURSE	
COLORECTAL SURGERY	BNSOT1	SGH-INPATIENT-BNSOT 01	DSAE	Visiting Con	
EMERGENCY MEDICINE	M4	Main Operating Theatre M4 (OTO)	BNS	REG	
Renal Medicine	RM4	LABOR WARD ROOM 4	ASC	VISITING SEN	
Neurosurgery	HCL3OT03	NHCS-OT3	NCC	SCN	
GENERAL SURGERY	R7	SGH-MOT-R7	GCL71A	Medical Offi	
DENTAL	RM7	Labor Ward Room 7	HAEM	SENIOR REGIS	
Urology	L4	SGH-MOT-L4	DSEC	HOUSE OFFICE	
Rehabilitation Medicine	R5	SGH-MOT-R5	MOTNHC	CON	
Medical Oncology	AS04	Ambulatory Surgery Centre OT4 (GA)	AEC	VISITING CON	
SDDC - Breast (SUR)	L6	SGH-MOT-L6	XURO	Resident Phy	
BURNS	UR02	UROLOGY CENTRE - OPERATING THEATRE 2	CXMOT	MO	
SDDC - Head & Neck (SUR)	AS01	Ambulatory Surgery Centre OT 1 (LA)	HXHOT	CONSULTANT	
SDDC - HEAD & NECK (SUR)	DSAE01	SGH-ED-ED01	MOTSGH	HO	
UROLOGY	L8	Main Operating Theatre L8 (ENT)	MOT	Senior Consu	
Liver Transplant	RM6	LABOR WARD ROOM 6	LW52A	REGISTRAR	
Geriatric Medicine	ED06	Endoscopy Centre Operating Theatre 6	UROT	NONE	

ETL Mapping Document



pasarTableName	pasarFieldName	mappingLogic	comments	cdmTableName	cdmFieldName	isRequired	cdmDatatype	userGuidance	etlConventions
pre_op.char	Admission_Type	based on the admission_type , we need to fetch the concept_id TODO: map standard concept ids	It has a value like Inpatient Day Surgery (DS) Same Day Admission (SDA)	visit_occurrence	visit_concept_id	Yes	integer	This field contains a concept id representing the kind of visit, like inpatient or outpatient. All concepts in this field should be standard and belong to the Visit domain.	Populate this field based on the kind of visit that took place for the person. For example this could be "Inpatient Visit", "Outpatient Visit", "Ambulatory Visit", etc. This table will contain standard concepts in the Visit domain. These concepts are arranged in a hierarchical structure to facilitate cohort definitions by rolling up to generally familiar Visits adopted in most healthcare systems worldwide.
pre_op.char	Admission_Type	admission_type	It has a value like Inpatient Day Surgery (DS) Same Day Admission (SDA)	visit_occurrence	visit_source_value	No	varchar(50)	This field houses the verbatim value from the source data representing the kind of visit that took place (inpatient, outpatient, emergency, etc.)	If there is information about the kind of visit in the source data that value should be stored here. If a visit is an amalgamation of visits from the source then use a hierarchy to choose the visit source value, such as IP ->
pre_op.char	Admission_Type	based on the admission_type , we need to fetch the concept_id TODO: map standard concept ids	It has a value like Inpatient Day Surgery (DS) Same Day Admission (SDA)	visit_detail	visit_detail_concept_id	Yes	integer	This field contains a concept id representing the kind of visit detail, like inpatient or outpatient. All concepts in this field should be standard and belong to the Visit domain.	Populate this field based on the kind of visit that took place for the person. For example this could be "Inpatient Visit", "Outpatient Visit", "Ambulatory Visit", etc. This table will contain standard concepts in the Visit domain. These concepts are arranged in a hierarchical structure to facilitate cohort definitions by rolling up to generally familiar Visits adopted in most healthcare systems worldwide.
post_op.icu	ICU_Admission_Time	ICU_Admission_Time		visit_detail	visit_detail_start_datetime	No	datetime	NA	If no time is given for the start date of a visit, set it to midnight

ETL Specification @ Github



md_files
.gitattributes
care_site.md
condition_occurrence.md
death.md
device_exposure.md
drug_exposure.md
index.md
measurement.md
note.md
observation.md
observation_period.md
person.md
procedure_occurrence.md
provider.md

71 lines (61 loc) · 3.07 KB

Table name: procedure_occurrence

Reading from intra_op__operation

Source
*anon_case_no
*id
*session_id
*operation_startdate
*operation_enddate
*operation_starttime
*operation_endtime

CDMV5.4
*procedure_occurrence_id
*person_id
*procedure_concept_id
*procedure_date
procedure_datetime
procedure_end_date
procedure_

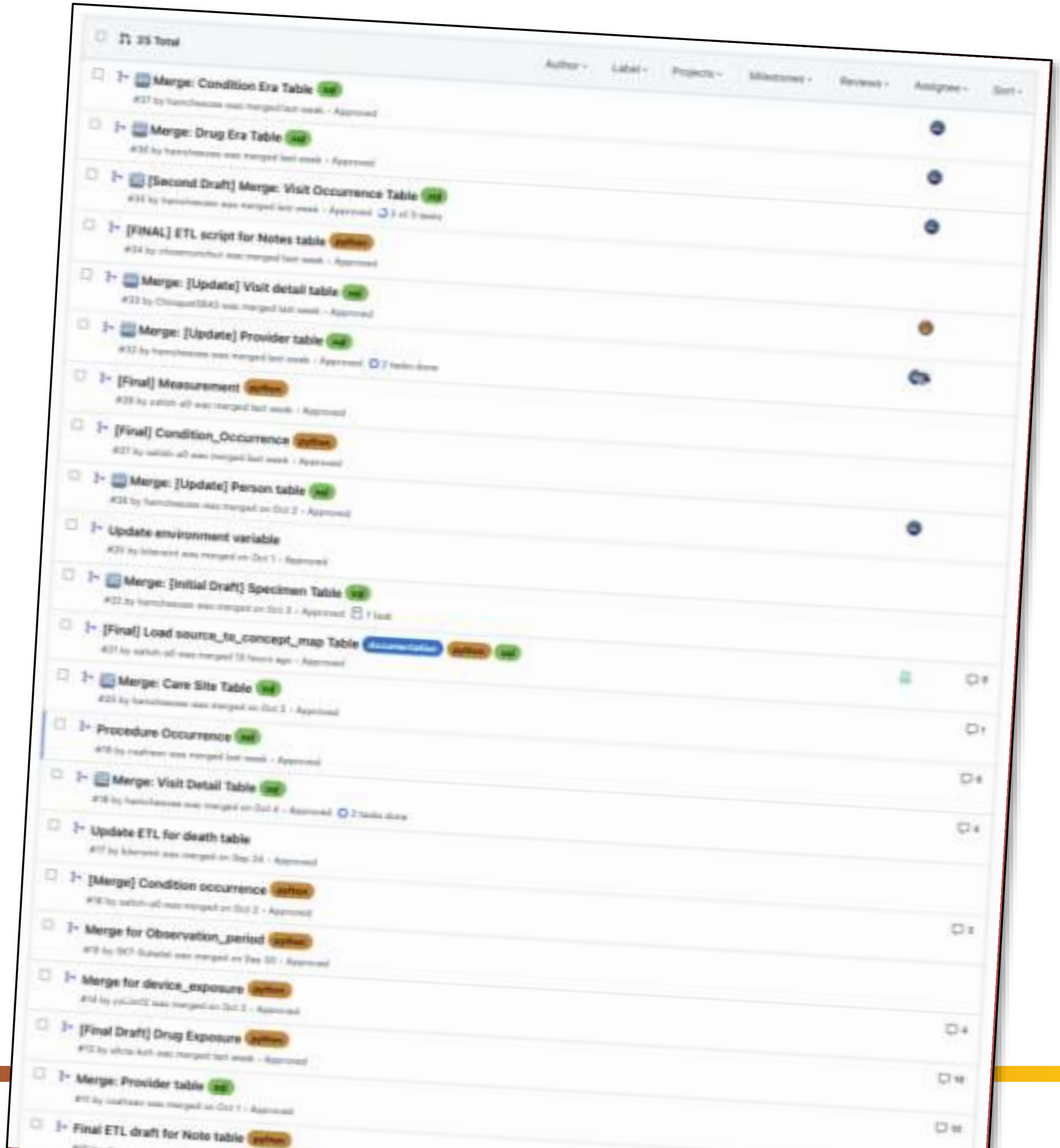
ETL Team

Setup

- SQL & Python sub teams created within ETL group
- 2 GCP VMs for each sub team
- 1 common Postgres Cloudsql Instance - With 3 schemas for Pasar 1% data and OMOP schemas created
- Separation of concerns with ssh access, VS Code, python virtual environment and local git for each user – Compliant with DUA
- Common Python ETL Framework developed
  - for a structured coordinated development among ETL members
  - Preload dependencies such as Athena Vocabulary data and source_to_concept_map
  - Run the whole pipeline for all the tables together with a single command

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 Occurrence tables – CPT4 Codes
- ~40 DQD QA issues resolved



□ 25 Total

□ Merge: Condition Era Table sql
#37 by [illegible] was merged last week · Approved

□ Merge: Drug Era Table sql
#36 by [illegible] was merged last week · Approved

□ [Second Draft] Merge: Visit Occurrence Table sql
#35 by [illegible] was merged last week · Approved · 3 of 3 tasks

□ [FINAL] ETL script for Notes table python
#34 by [illegible] was merged last week · Approved

□ Merge: [Update] Visit detail table sql
#33 by [illegible] was merged last week · Approved

□ Merge: [Update] Provider table sql
#32 by [illegible] was merged last week · Approved · 7 tasks done

□ [Final] Measurement python
#30 by [illegible] was merged last week · Approved

□ [Final] Condition_Occurrence python
#27 by [illegible] was merged last week · Approved

□ Merge: [Update] Person table sql
#26 by [illegible] was merged on Oct 2 · Approved

□ Update environment variable
#25 by [illegible] was merged on Oct 1 · Approved

□ Merge: [Initial Draft] Specimen Table sql
#22 by [illegible] was merged on Oct 3 · Approved · 1 task

□ [Final] Load source_to_concept_map Table documentation python sql
#21 by [illegible] was merged 13 hours ago · Approved

□ Merge: Care Site Table sql
#20 by [illegible] was merged on Oct 3 · Approved · 9

□ Procedure Occurrence sql
#19 by [illegible] was merged last week · Approved · 1

□ Merge: Visit Detail Table sql
#18 by [illegible] was merged on Oct 4 · Approved · 2 tasks done · 6

□ Update ETL for death table
#17 by [illegible] was merged on Sep 24 · Approved · 6

□ [Merge] Condition occurrence python
#16 by [illegible] was merged on Oct 2 · Approved

□ Merge for Observation_period python
#15 by [illegible] was merged on Sep 30 · Approved · 2

□ Merge for device_exposure python
#14 by [illegible] was merged on Oct 3 · Approved

□ [Final Draft] Drug Exposure python
#12 by [illegible] was merged last week · Approved · 4

□ Merge: Provider table sql
#11 by [illegible] was merged on Oct 1 · Approved · 10

□ Final ETL draft for Note table python

Ingestion Statistics

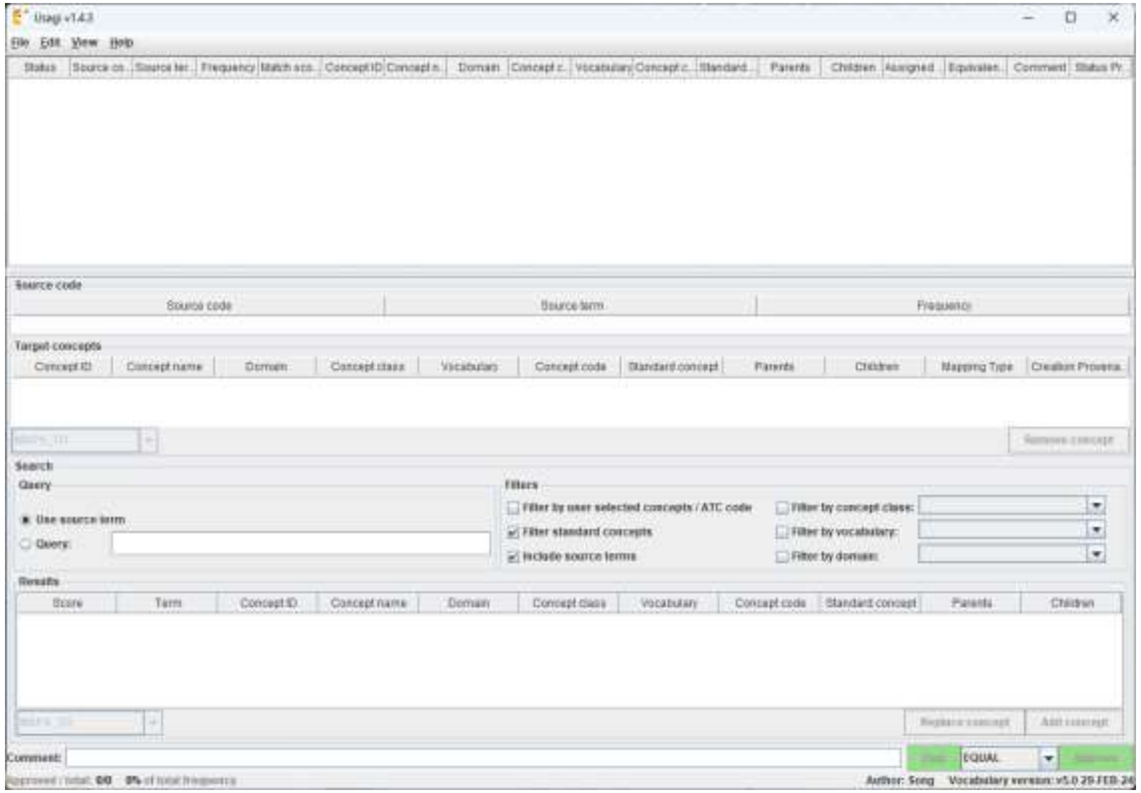
Clinical Table	Records count	Time Taken
cdm_source	1	.08s
care_site	111	.46s
provider	872	.1s
person	999	.29s
observation_period	999	.42s
death	143	.4s
visit_occurrence	1,600	.18s
visit_detail	268	3.17s
condition_occurrence	26,464	39.9s
condition_era	4,193	1.78s
drug_exposure	26,065	2.65s
drug_era	6,793	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

Vocabulary Mapping Team

Tool: USAGI


Usagi v1.4.3
File Edit View Help
Status | Source co.. | Source ter.. | Frequency | Match sco.. | Concept ID | Concept n.. | Domain | Concept c.. | Vocabulary | Concept c.. | Standard.. | Parents | Children | Assigned.. | Equivalen.. | Comment | Status Pr..
Source code
Source code | Source term | Frequency
Target concepts
Concept ID | Concept name | Domain | Concept class | Vocabulary | Concept code | Standard concept | Parents | Children | Mapping Type | Creation Provena..
Remove concept
Search
Query
Filters
Results
Score | Term | Concept ID | Concept name | Domain | Concept class | Vocabulary | Concept code | Standard concept | Parents | Children
Replace concept
Add concept
Comment:

● Use source term
○ Query:

□ Filter by user selected concepts / ATC code
☑ Filter standard concepts
☑ Include source terms
□ Filter by concept class:
□ Filter by vocabulary:
□ Filter by domain:

EQUAL

Approved / total: 0/0 0% of total frequency

Author: Song Vocabulary version: v5.0 29-FEB-24

- Little knowledge of OMOP standardized vocabularies and mapping process
- Training session to familiarize team on how to use USAGI
- Utilized USAGI to map source codes to OMOP concept_ids
- Additional peer review

Mapping Results

Name of the file	Assigned to	Status
intraop_aimsvitals_vitalcode	Liying Pei	Complete
intraop_drugdrug_group1	Leong Hui Wong	Complete
intraop_drugmed_group1	Leong Hui Wong	Complete
intraop_nurvitals_group1	Lakshmi Kubendran	Complete
intraop_operation_group1	Liying Pei, Shigemi Matsumoto, Kosuke Tanaka, Qi Yang, Asif Syed, Elizabeth E. Park, Keiko Asao	Complete
postop_clindoc_group1	Qi Yang	Complete
postop_info_group1	Leong Hui Wong	Complete
postop_lab_testdesc	Lakshmi Kubendran	Complete
postop_labmicro_antibioticname	Not Mapped	
postop_labmicro_microresulted proceduredescription		
postop_labmicro_organismdescription		
postop_labsall_group1	Lakshmi Kubendran	Complete
postop_pacu_group1	Kosuke Tanaka	Complete
postop_renal_group1	Asif Syed	Complete
preop_char_allergyinformation	Lakshmi Kubendran	Complete
preop_char_gender	Qi Yang	Complete
preop_char_race	Qi Yang	Complete
preop_lab_prepoplabtestdescription	Lakshmi Kubendran	Complete
preop_radiology_procedurename	Qi Yang	Complete
Surgical specialty	Lakshmi Kubendran	Complete

- Prioritized mapping by sorting codes by descending order of frequency
- Targeted 95% mapping rate by code frequency
- Categorized mapping files based on difficulty level (e.g., requirement of pharmaceutical/clinical knowledge)
- Assigned mappings to volunteers' expertise/capabilities
- For surgery codes, assigned multiple resources due to importance of data

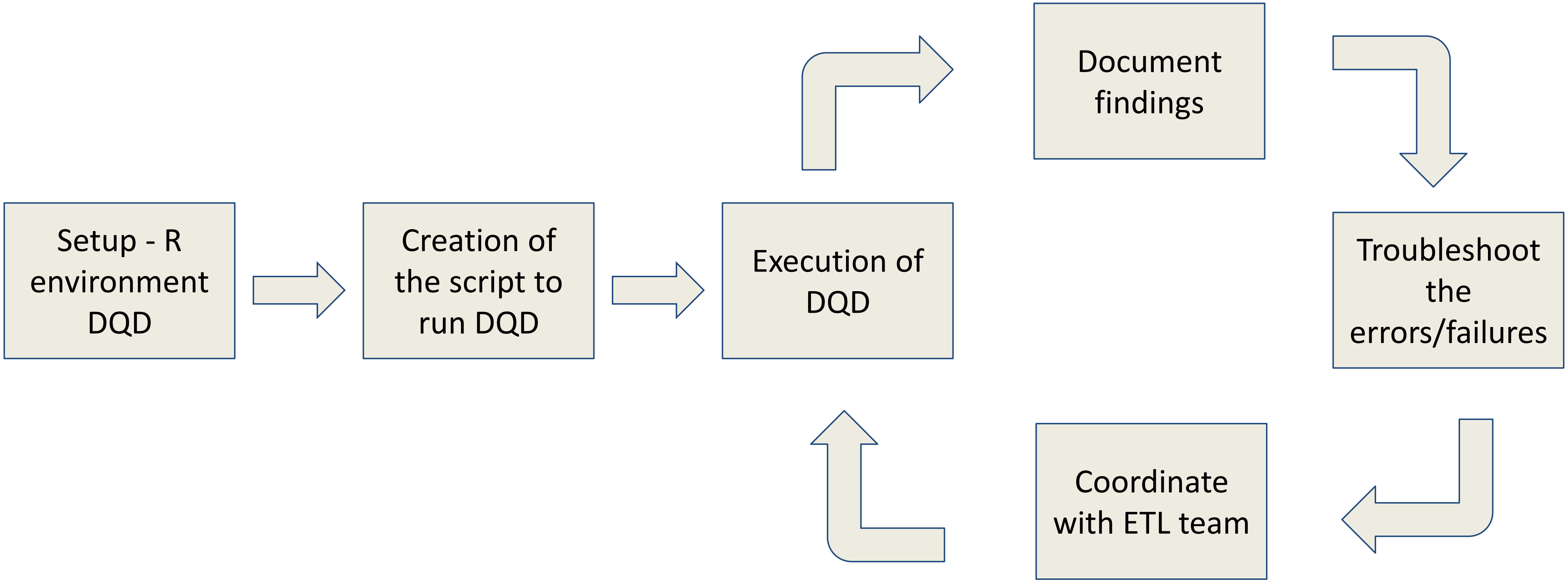
Quality Assurance Team

Goals

- Evaluate the quality of the data mapped to OMOP CDM in the ETL process
- Ensure the quality of the data meets certain % threshold

QA Journey

QA Journey



Setup - R environment DQD

Creation of the script to run DQD

Execution of DQD

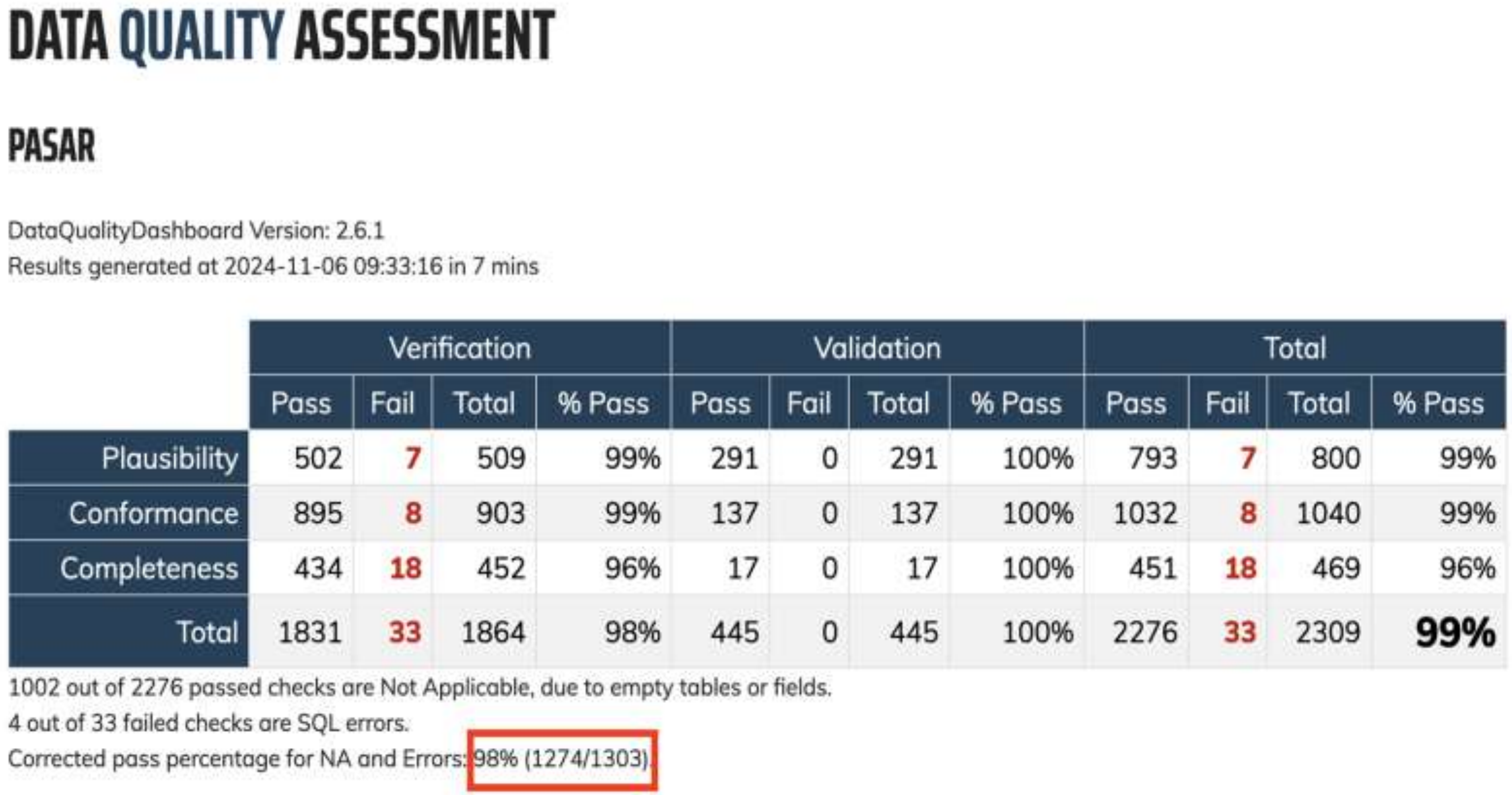
Document findings

Troubleshoot the errors/failures

Coordinate with ETL team

Data Quality

PASAR OMOP CDM Data Quality is 98%



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).

Data Quality Assets

- Scripts and results are available on github

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Thank you!