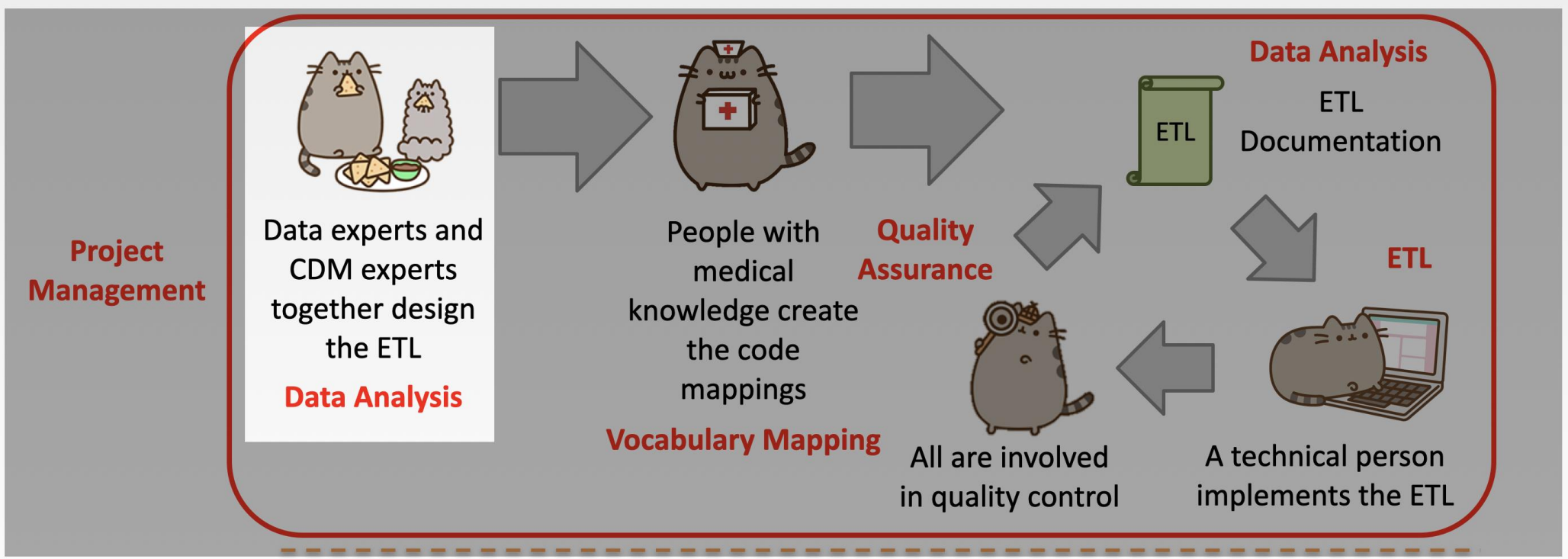
APAC Community-wide ETL Project

Sprint 1 Review

Data Analysis Team

OMOP Data Model Mapping



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

Assigned Tables to Groups

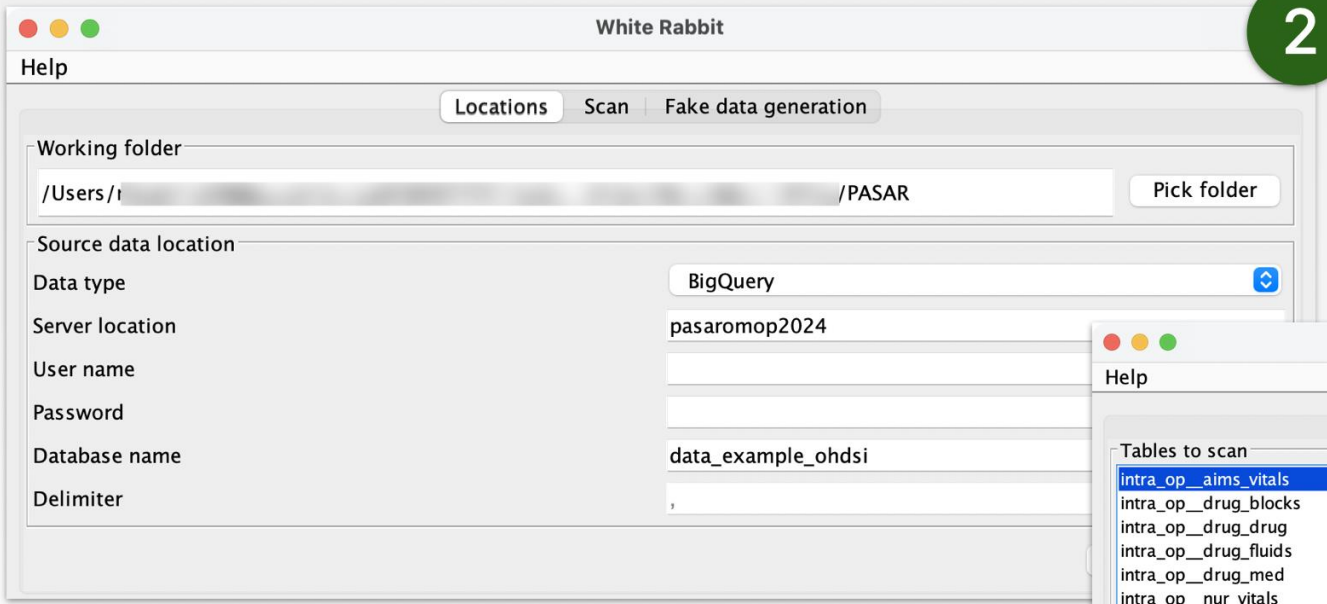
Standard Table	Types	Group	Difficulty
1. local concept master	dimension	Demo	2
2. person	dimension	Demo	1
3. location	dimension	A	1
4. care_site	dimension	A	1
5. provider	dimension	A	1
6. visit_occurrence	event	E	2
7. observation_period	event	E	2
8. death	event	B	1
9. condition_occurrence	event	B	2
10. observation	event	B	2
11. procedure_occurrence	event	B	3
12. drug_exposure	event	C	4
13. condition_era	aggregate		0
14. drug_era	aggregate		0
15. measurement	event	D	4
16. device_exposure	event	A	2
17. cost	event		0
18. payer_plan_period	event		0
19. visit_detail	event	E	2
20. specimen	event	A	1
21. note	event	A	2

	#volunteer
Group A	3
Group B	3
Group C	2
Group D	2
Group E	0

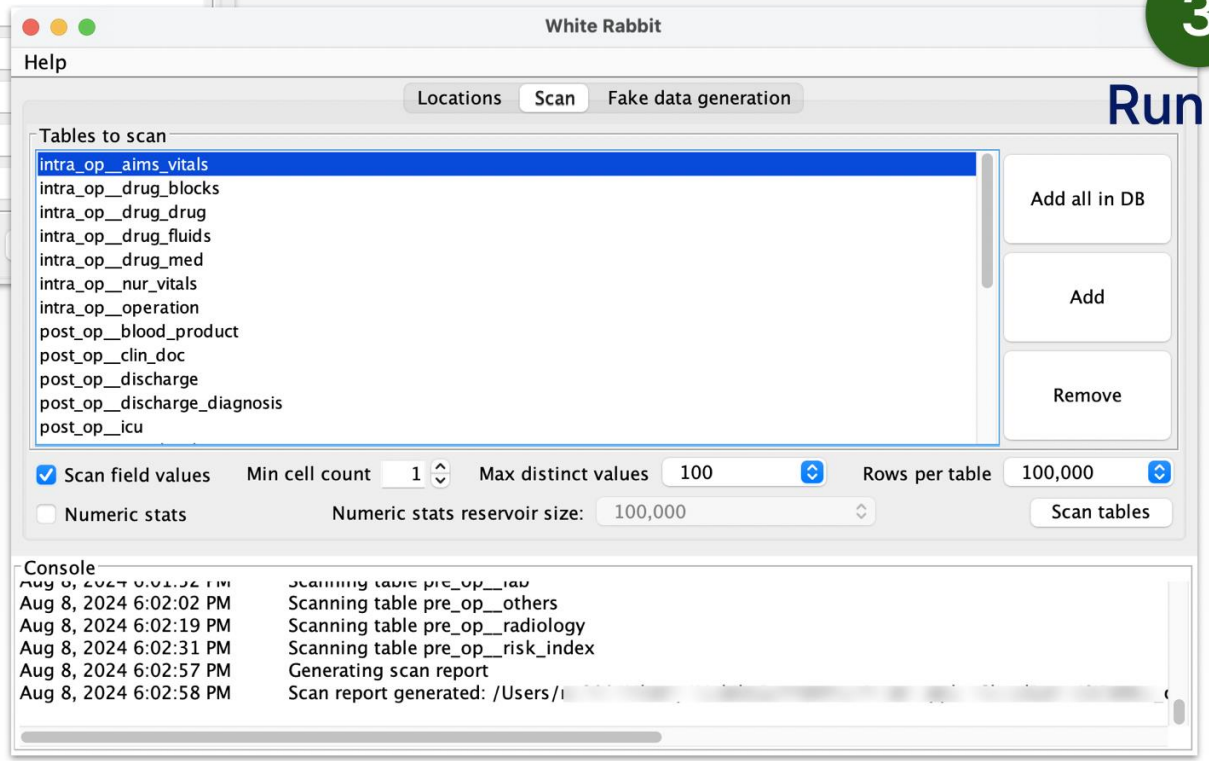
Need volunteers for group E!

WhiteRabbit: Scan Values

Tool #1: Already Done



2 Set Database Connection

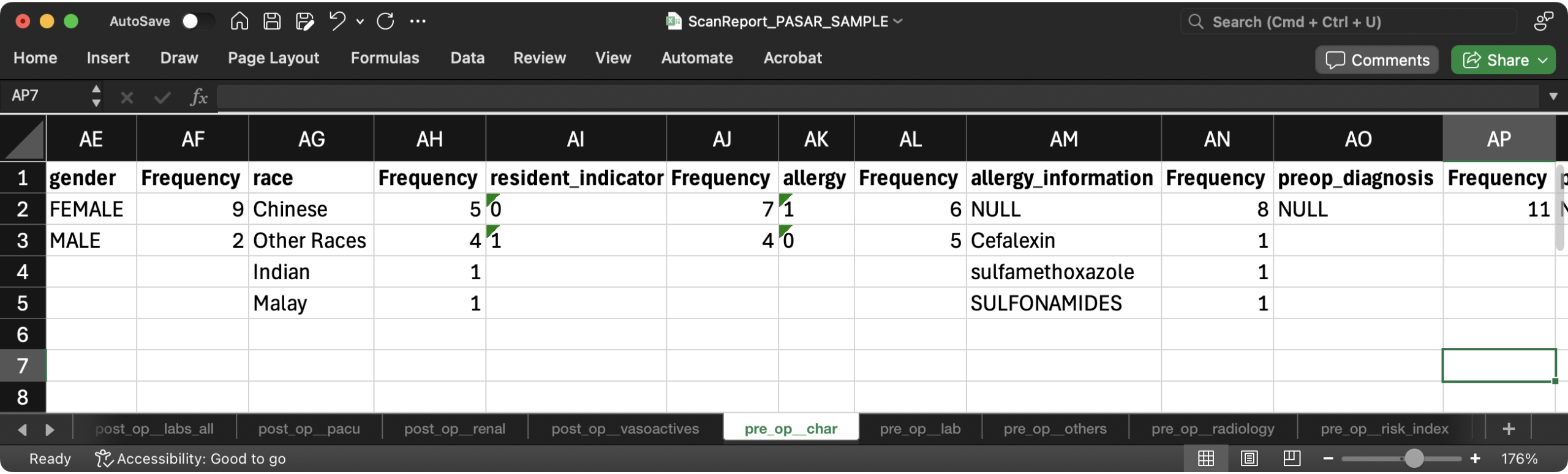


3 Run Scan

1 Install following http://ohdsi.github.io/

WhiteRabbit/WhiteRabbit.html#Installation

WhiteRabbit: ScanReport



	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP
1	gender	Frequency	race	Frequency	resident_indicator	Frequency	allergy	Frequency	allergy_information	Frequency	preop_diagnosis	Frequency
2	FEMALE	9	Chinese	5	0	7	1	6	NULL	8	NULL	11
3	MALE	2	Other Races	4	1	4	0	5	Cefalexin	1		
4			Indian	1					sulfamethoxazole	1		
5			Malay	1					SULFONAMIDES	1		
6												
7												
8												

Claude AI

Tool #2: Also Done

2

Prompt: "Provided is the table and field description of the source data for mapping to OMOP CDM v5.4. Your task is to recommend corresponding OMOP tables and fields for each source field. No explanations are required. Ensure each field is listed separately without consolidation. Deliver the results in a comprehensive CSV file."

1

Input 3 fields from PASAR data dict as CSV

	A	B	C	D	E	F
1	id	Table	Field	Explanation	Claude's Suggested Table	Claude's Suggested Field
49	47	pre_op.char	Admit_Visit_Patient_Class_Descripti	Admission class status	VISIT_OCCURRENCE	visit_concept_id
50	48	pre_op.char	Visit_Date	Date of preoperative visit	VISIT_OCCURRENCE	visit_start_date
51	49	pre_op.char	Age_Time_of_Surgery	Age	OBSERVATION	value_as_number
52	50	pre_op.char	Gender	Gender	PERSON	gender_concept_id
53	51	pre_op.char	Race	Race	PERSON	race_concept_id
54	52	pre_op.char	Resident_Indicator	Residential cluster	LOCATION	location_source_value
55	53	pre_op.char	Allergy	Drug allergy	OBSERVATION	observation_concept_id
56	54	pre_op.char	Allergy_Information	Information on drug allergy	OBSERVATION	observation_source_value
57	55	pre_op.char	Preop_Diagnosis	Preoperative dialysis	CONDITION_OCCURRENCE	condition_concept_id
58	56	pre_op.char	Proposed_Operation	Proposed operation	PROCEDURE_OCCURRENCE	procedure_concept_id
59	57	pre_op.char	Admission_Type	Type of admission	VISIT_OCCURRENCE	visit_type_concept_id
60	58	pre_op.char	Smoking_History	Smoking history	OBSERVATION	observation_concept_id
61	59	pre_op.char	Any_Steroids_Past_6_Months	Taking steroids in last 6 months	DRUG_EXPOSURE	drug_concept_id
62	60	pre_op.char	Current_Herbal_Treatment	Taking TCM in last 6 months	DRUG_EXPOSURE	drug_concept_id
63	61	pre_op.char	Rapid_Assessment	Rapid assessment score	OBSERVATION	value_as_number
64	62	pre_op.char	Any_Aspirin_Warfarin_Anti_Platelet_I	Taking any aspirin or warfarin	DRUG_EXPOSURE	drug_concept_id
65	63	pre_op.char	Alcohol_Consumption	Alcohol concumption	OBSERVATION	observation_concept_id
66	64	pre_op.char	Pregnancy_Gender	Pregnant or not	OBSERVATION	observation_concept_id
67	65	pre_op.char	Curr_TCM_herbal_Treatment	Current TCM note	DRUG_EXPOSURE	drug_concept_id
68	66	pre_op.char	Curr_TCM_herbal_Treatment_Notes	Current TCM note	DRUG_EXPOSURE	drug_source_value

Data | Data Analysis Team | **Table.Field explanation+Claude** | Mapping

3

Results not perfect but helpful

Online Excel *(see link in MS Teams)*

Tool #3: This is where we're working on together

	A	B	C	D	E	F	G	H	I	J
1	pasarTableName	pasarFieldName	mappingLogic	comments	cdmTableName	cdmFieldName	isRequired	cdmDatatype	userGuidance	etlConventions
2					person	person_id	Yes	integer	It is assumed that every person with a different unique identifier is in fact a different person and should be treated independently.	Any person linkage that needs to occur t identify Persons ought to be done prior table. This identifier can be the original i source data provided if it is an integer, o be an autogenerated number.
3					person	gender_concept_id	Yes	integer	This field is meant to capture the biological sex at birth of the Person. This field should not be used to study gender identity issues.	Use the gender or sex value present in t the assumption that it is the ogical sex a source data captures gender identity it s in the [OBSERVATION](https://ohdsi.github.io/ del/cdm531.html#observation) table. [A concepts](http://athena.ohdsi.org/searc omain=Gender&standardConcept=Stanc geSize=15&query=). Please refer to the repository](https://ohdsi.github.io/Ther oncept_id.html) for detailed convention populate this field.
4					person	year_of_birth	Yes	integer	Compute age using year_of_birth.	For data sources with date of birth, the extracted. If no year of birth is available data should be dropped from the CDM i additional information on how to popul please refer to the [THEMIS

Data | Data Analysis Team | Table.Field explanation | **Mapping**

Workbook Statistics

Give Feedback to Microsoft 100%

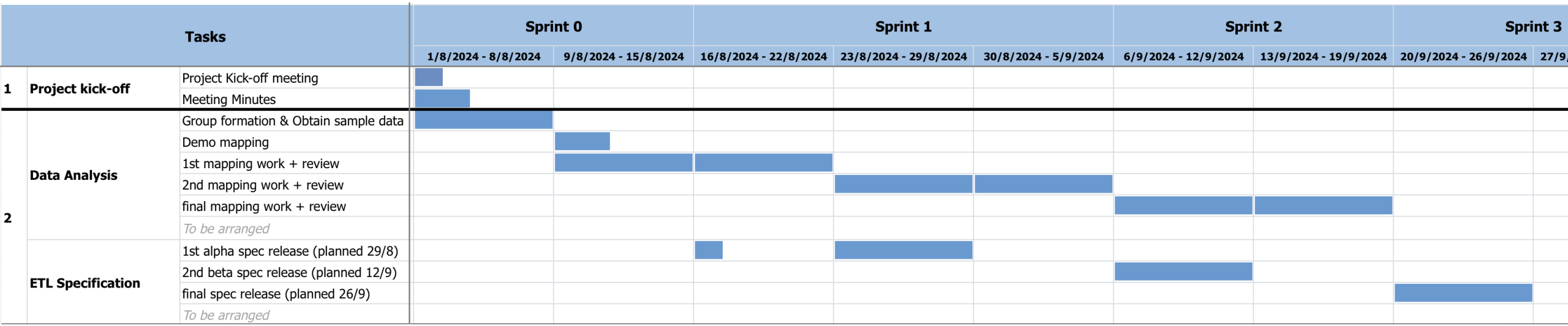
2 Each team add PASAR table and field name(s) here.

mappingLogic is for how value should be handled programmatically, e.g., MALE → 8507, FEMALE → 8532.

1 OMOP CDM Spec CSV

from https://github.com/OHDSI/CommonDataModel/blob/main/inst/csv/OMOP_CDMv5.4_Field_Level.csv

TENTATIVE Timeline



	Tasks		Sprint 0		Sprint 1			Sprint 2		Sprint 3	
			1/8/2024 - 8/8/2024	9/8/2024 - 15/8/2024	16/8/2024 - 22/8/2024	23/8/2024 - 29/8/2024	30/8/2024 - 5/9/2024	6/9/2024 - 12/9/2024	13/9/2024 - 19/9/2024	20/9/2024 - 26/9/2024	27/9
1	**Project kick-off**	Project Kick-off meeting									
		Meeting Minutes									
2	**Data Analysis**	Group formation & Obtain sample data									
		Demo mapping									
		1st mapping work + review									
		2nd mapping work + review									
		final mapping work + review									
		To be arranged									
	ETL Specification	1st alpha spec release (planned 29/8)									
		2nd beta spec release (planned 12/9)									
		final spec release (planned 26/9)									
		To be arranged									

To Vocab Team: Data Analysis Team needs your Vocab mapping (Usagi) to compile the ETL spec.
Or else, you could send the mapping to ETL team directly.

To ETL Team: We will start sending spec for some tables (e.g., PERSON) within the next week.
Then, more tables will come in the following weeks.

*Demo mapping recording on MS Teams


OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Thank you!

APAC Community-wide ETL Project

Sprint 1 Review

ETL Team

Overview

- Split into 2 teams (Python and SQL) to assess optimal pipeline
- Will see if we stick with both, or select one at the end of the next sprint

- Code tracking via Github (Afreen IC)
- Overall technical IC: Satish (will stitch stuff together)

Team organization

- Split into Python and SQL sub-teams as independent tracks
  i. Steven and Satish + members with Python as primary skillset
  ii. Afreen and Jiawei + members with SQL as primary skillset

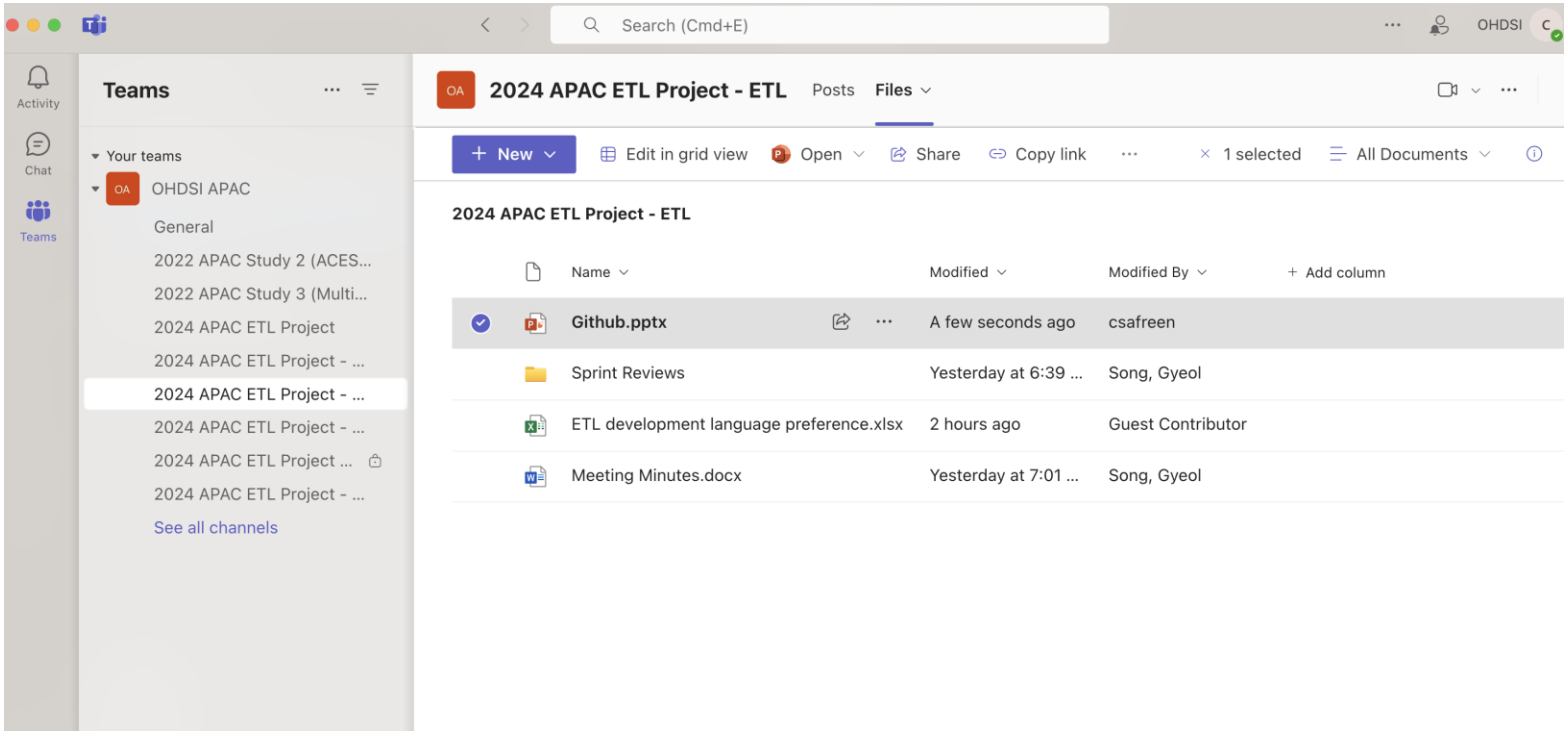
No.	Name	1st choice among SQL, Python and R	2nd choice among SQL, Python and R	Experience with Github?	Experience with Docker?
1	Evelyn Goh	SQL	Python (limited experience)	Yes	No
2	Jiawei Qian	SQL	Python (limited experience)	No	No
3	Steven Yong	Python	SQL		
4	Satish Kumar Anbazhagan	Python	Sql	Yes	Yes
5	Afreen Chitwadgi Sikandara	Python	Sql	Yes	Yes
6	Sornchai Manoson	SQL	Python (limited experience)	Yes	Yes
7	Chinapat Onprasert	SQL	Python	Yes	Yes
8	Nongnaphat Wongpiyachai	SQL	Python	Yes	Yes
9	Max Natthawut Adulyanukosol	SQL	Python	Yes	Yes
10	Ethan Lin				
11	Alicia Koh	Python	SQL	Yes	Yes
12	Pattarachai Roongsritong				
13	Hengxian Jiang				
14	Erwin Tantoso	Python	SQL	Yes	Yes
15	Brandan Tan	Python	SQL	Yes	Yes
16	Sukatat Leknimit	Python	SQL	Yes	Yes
17	Millahat Asif				

PyPASAR ETL

- Skeleton Source code https://github.com/satish-a0/pasar
- Postgres database & OMOP v5.4 Schema can be setup
- Model OMOP tables as individual python classes for isolation
- Flexible by Design. ETL Implementors can choose SQL / Python
- Wrapper execution code defined to trigger final ETL for each table in their order of dependencies. Ex: Person -> Visit Occurrence -> ...
- Final Objective: *Enable SQL constraints defined by OHDSI and run the whole ETL flow for all OMOP tables contributed by the community in a seamless manner*
- Instructions available from Readme to get familiarized on local!

GitHub Repository setup

- The instructions on how to setup the repository in local and create pull requests for new features are detailed in a PowerPoint and is uploaded in the team's channel.



Sprint Reflections

- Need to assign specific decision-makers to decide on stuff
- Team has different tech stack preferences
  - Each tech stack has their own pros and cons
- Time needs to be taken to onboard academic members of team onto industry-related software like Github


OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Thank you!