Common Data Model
v4 to v5 Conversion Script

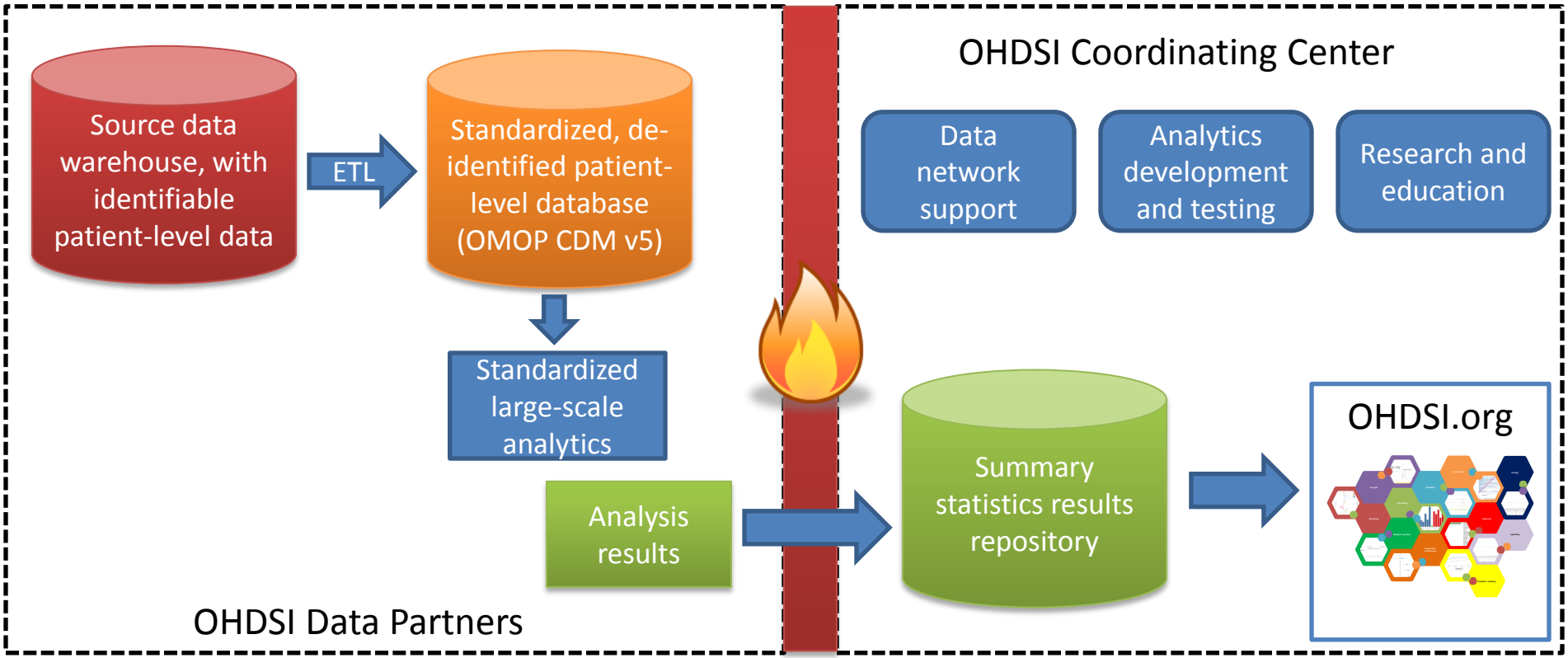
Anthony Sena

11 August 2015

Agenda

- OMOP CDM & conversion from CDM V4 to V5
- Where can I find the conversion script?
- Using the conversion script
- Questions/Comments

How OHDSI Works



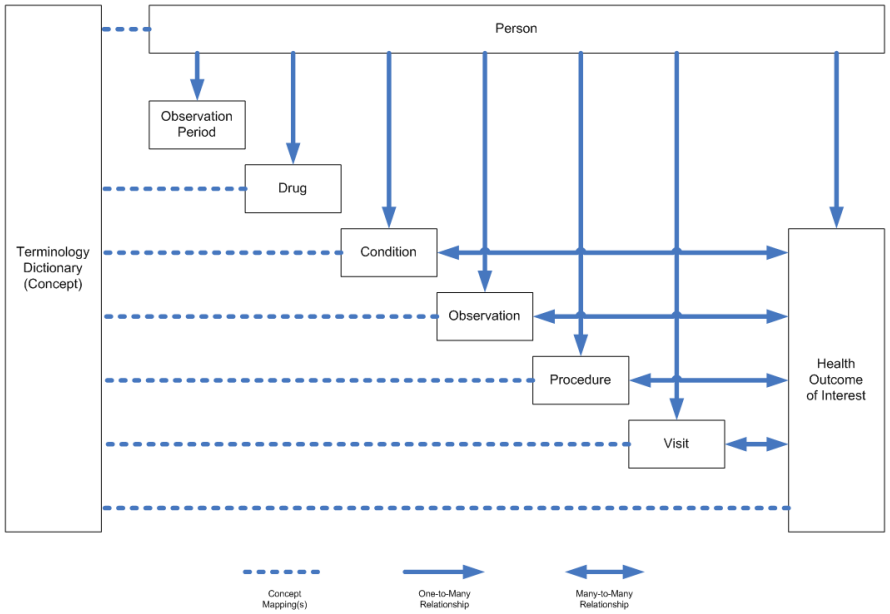
Objectives in OMOP Common Data Model development

- One model to accommodate both administrative claims and electronic health records
  - Claims from private and public payers, and captured at point-of-care
  - EHRs from both inpatient and outpatient settings
  - Also used to support registries and longitudinal surveys
- One model to support collaborative research across data sources both within and outside of US
- One model that can be manageable for data owners and useful for data users (efficient to put data IN and get data OUT)
- Enable standardization of structure, content, and analytics focused on specific use cases

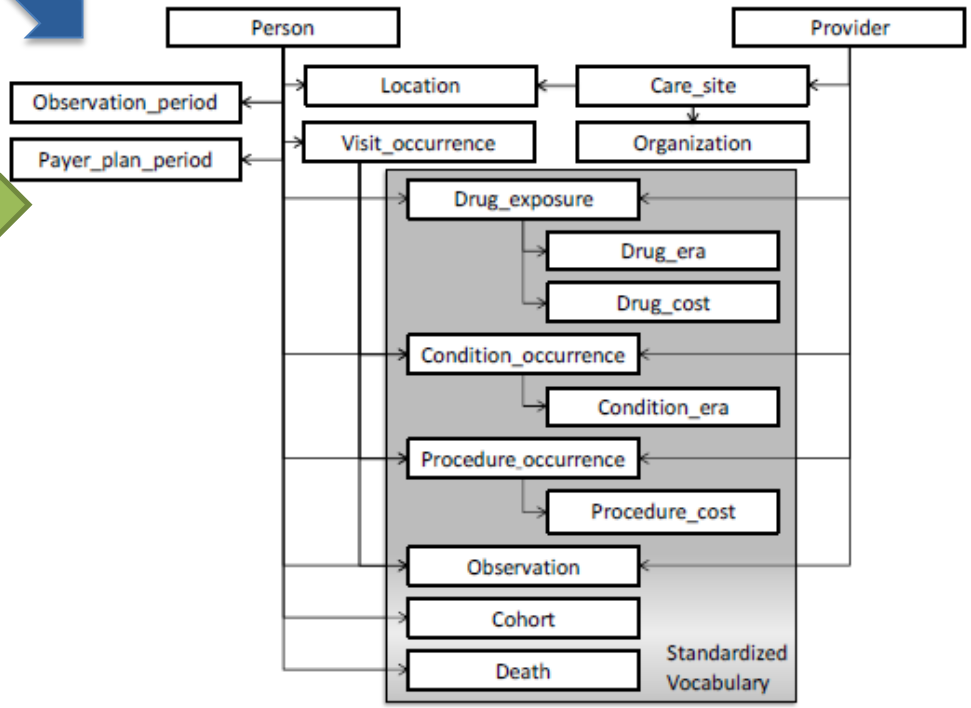
Evolution of the OMOP Common data model

OMOP CDM now Version 5, following multiple iterations of implementation, testing, modifications, and expansion based on the experiences of the OMOP community who bring on a growing landscape of research use cases.

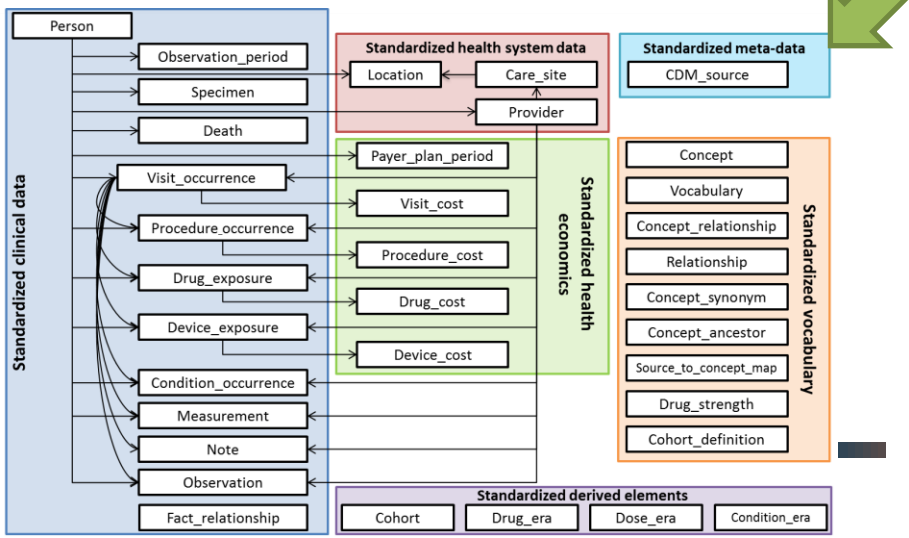
OMOP CDMv2



OMOP CDMv4

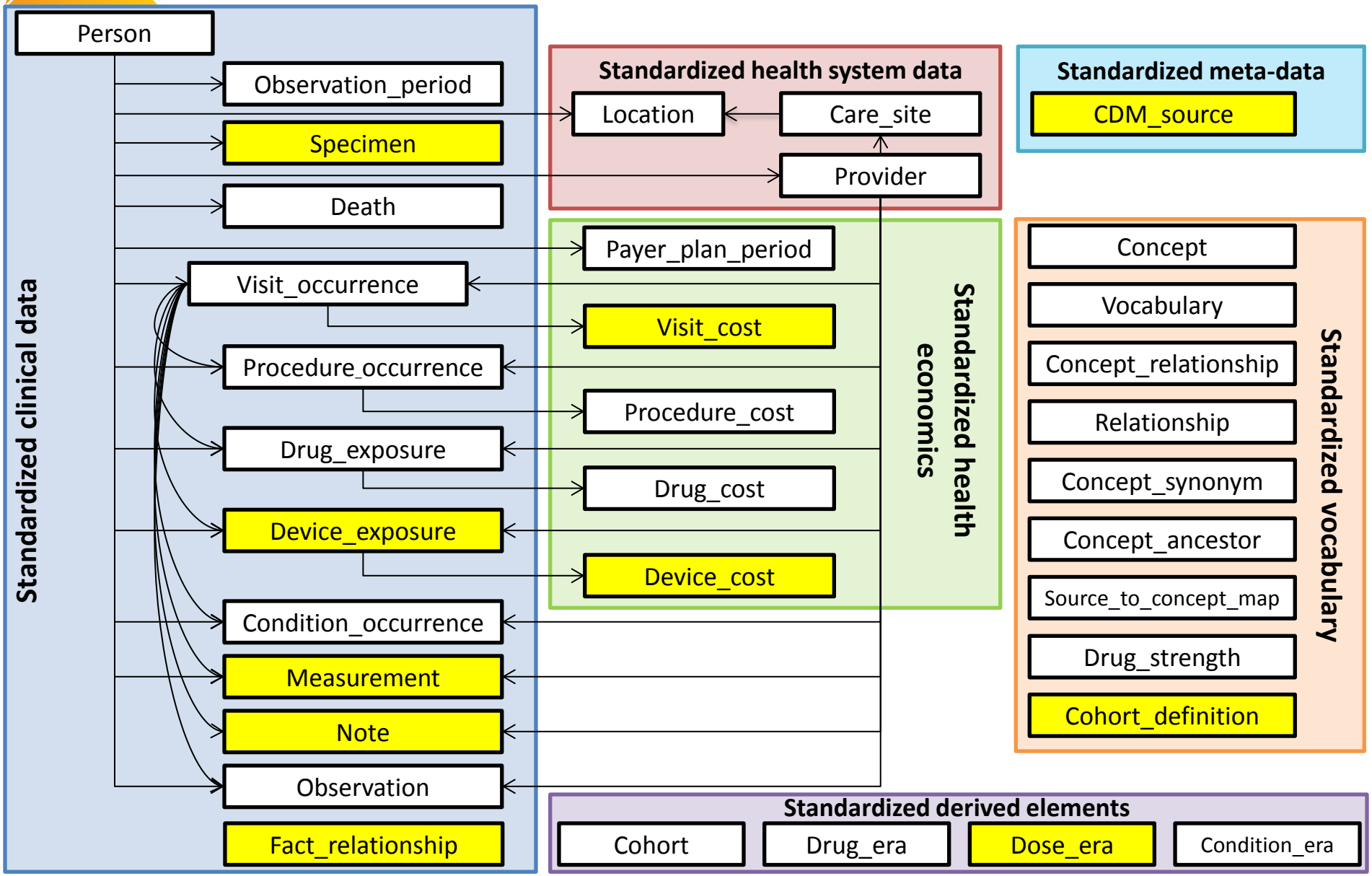


OMOP CDMv5



http://omop.org/CDM

New tables in CDMv5



Vocabulary-driven ETL

Codes in one coding system can belong to different domains. For example:

ICD-9 V80.1 (Screening for glaucoma)

is a diagnose code, but indicates that a procedure was performed. In the new vocabulary it is mapped to

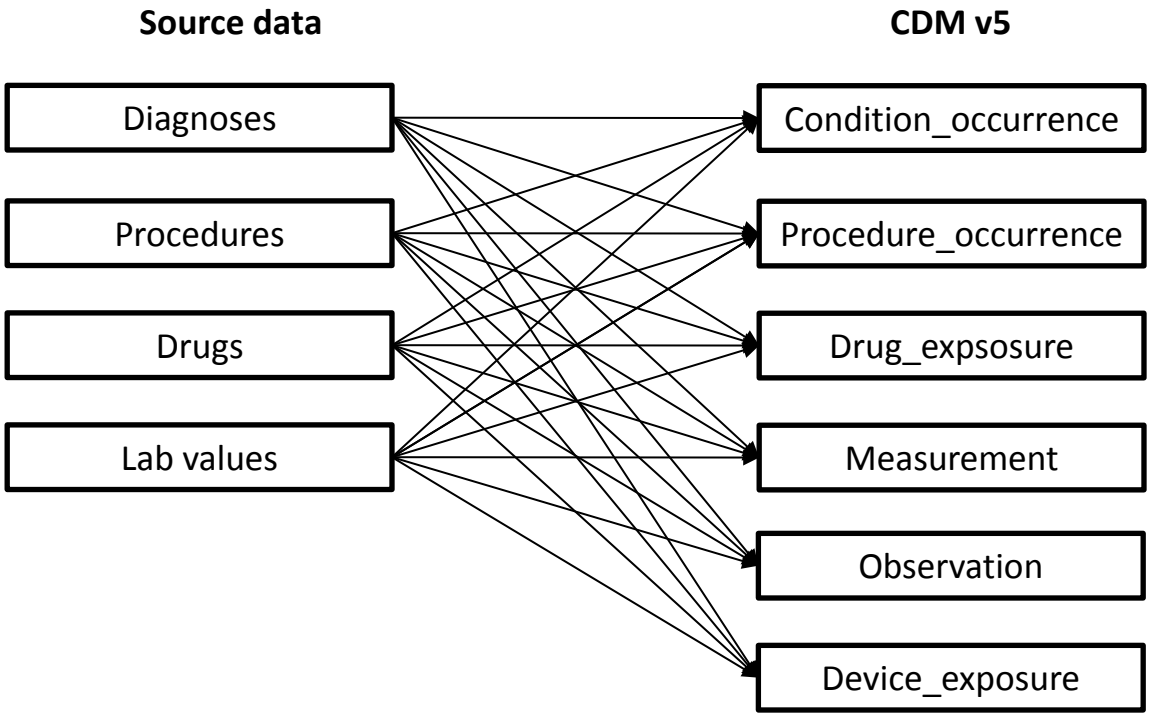
- SNOMED 171215009 (Glaucoma screening)

This concept belongs to the Procedure domain, and should go into the Procedure_occurrence table

Vocabulary-driven ETL

- Only Standard concepts are allowed in the CDM concept_id fields (see *standard_concept* field in Concept)
- Each Standard concept has one (and only one) domain (see *domain_id* field in Concept), which defines where the concept should go

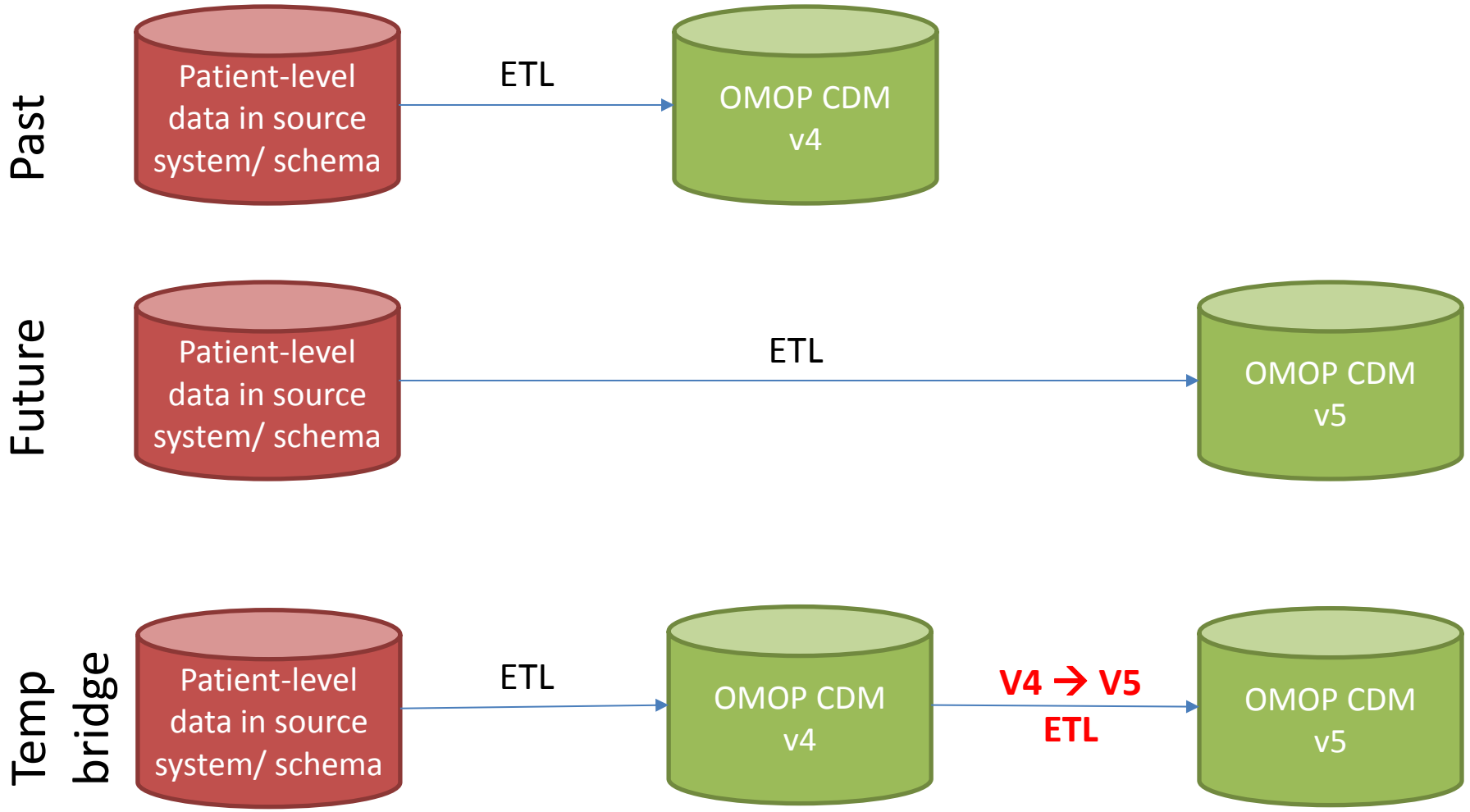
Many-to-many mappings



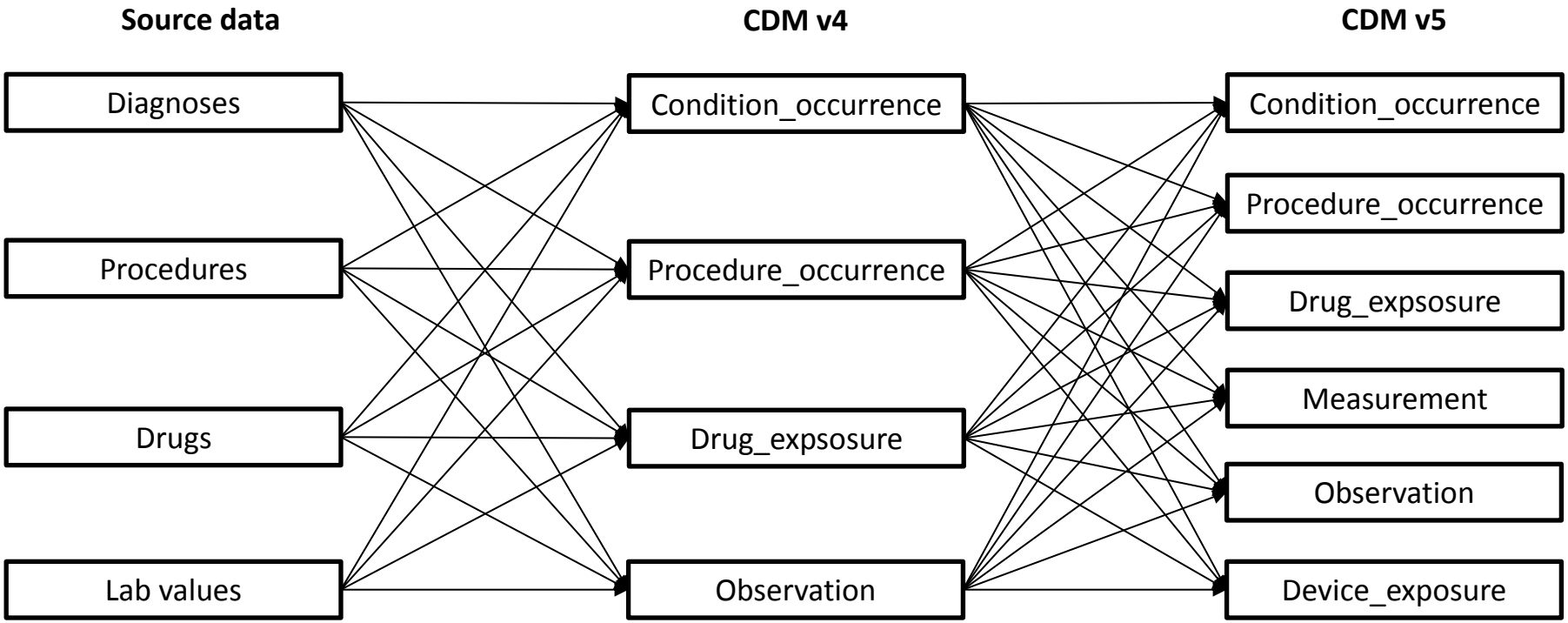
Typical:

- Diagnoses
  - 90% conditions
  - 8% procedures
  - 2% other
- Procedures
  - 70% procedures
  - 20% measurements
  - 10% devices
- Drugs
  - 90% drugs
  - 10% devices
- Lab values
  - 100% measurements

Progression of OMOP CDM ETLs



Conversion script mappings



Data carries forward from V4 to V5

Where can I find the script?

- GitHub -> CommonDataModel -> Version 4 To Version 5 Conversion
- What you'll find on GitHub:
  - An "OHDSI-SQL" (aka templateSQL) script that you can modify and run through SqlRender
  - Subfolders that contain rendered SQL files for the different RDBMS (SQL Server, PostgreSQL, Oracle, etc)
- Script has 2 sections:
  - Conversion from V4 to V5
  - Queries that will provide some basic stats to perform some quality checks of the conversion

How do I use it?

- Prerequisites
  - Create a target CDM v5 database on the same database server as your CDM v4 database
  - Download the V5 vocabulary from Athena and load it into the CDM v5 database
  - Ensure that you have enough disk space to accommodate the creation of the new. We estimate you will need 6x the storage of your current V4 database.
  - If you made any changes in your V4 database, be sure to carry those over to V5 and to check the conversion script for potential impact. For example, if you converted a field to BIGINT in V4, you will need to ensure this is carried forward.

How do I use it?

- Running the script
  - The script requires read rights from the V4 database and will need owner privileges to the V5 database
  - Perform a find & replace on the script for the search terms:
    - [SOURCE_CDMV4].[SCHEMA] - Your V4 database name + schema
    - [SOURCE_CDMV4] - Your V4 database name
    - [TARGET_CDMV5].[SCHEMA] - Your V5 database name + schema
    - [TARGET_CDMV5] - Your V5 database name
  - Run the script & review the output to see how your data was moved from V4 to V5.

Checking the results

- 3 Queries provide results:
  1. Row count comparisons between V4 and V5 tables
  2. Mapping of V4 tables to target V5 tables using the vocabulary
  3. Total of the V4 tables mapped to their V5 domains versus the V5 row counts to tie out the mapping.

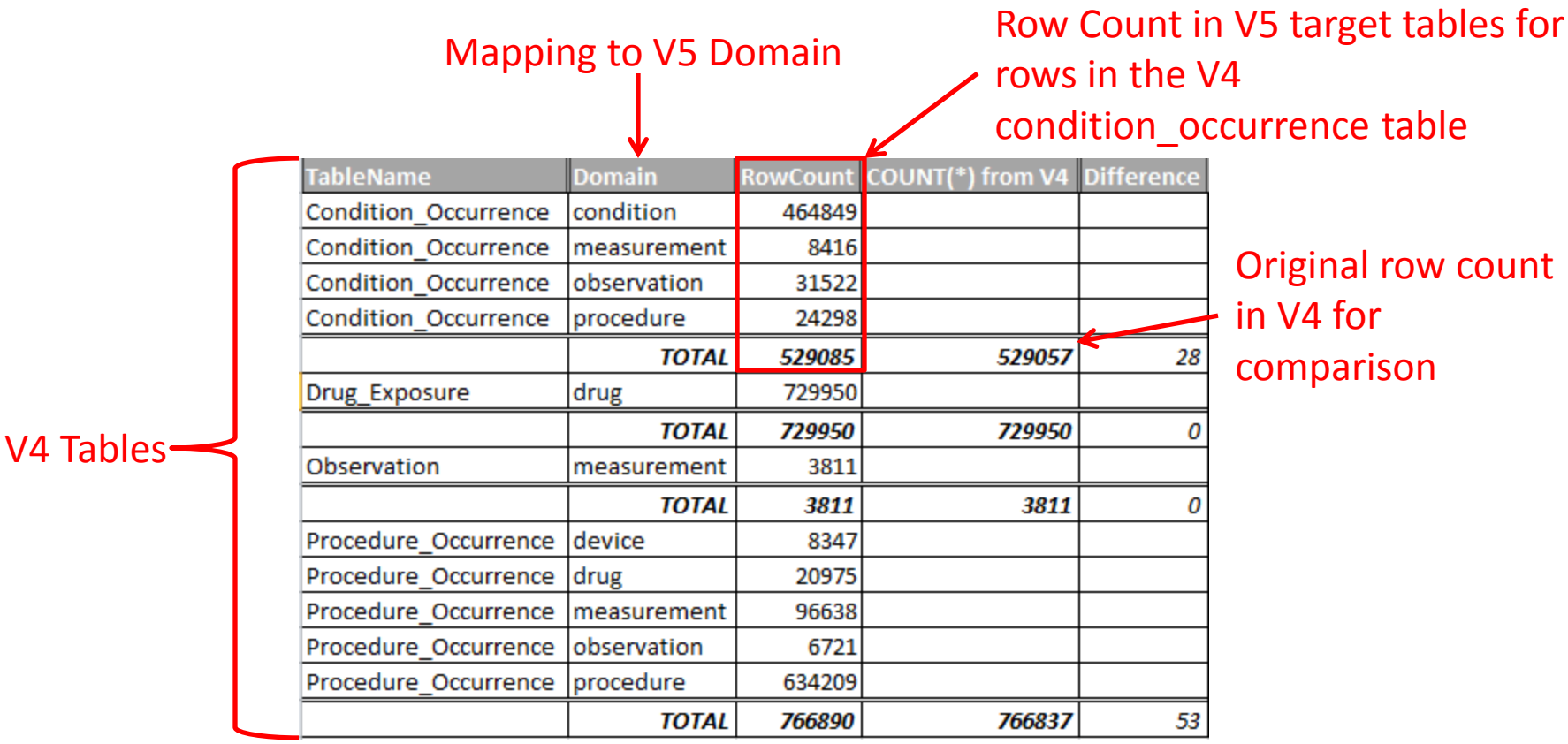
Example Output

Reviewing the results

Database Name	TableName	Rows	Database Name	TableName	Rows	Row Count Chg
[CDMV4]	CARE_SITE	1	[CDMV5]	care_site	1	0
[CDMV4]	CONDITION_ERA	349425	[CDMV5]	condition_era	304903	-44522
[CDMV4]	CONDITION_OCCURRENCE	529057	[CDMV5]	condition_occurrence	464849	-64208
[CDMV4]	DEATH	3	[CDMV5]	death	3	0
None	NULL	NULL	[CDMV5]	device_exposure	8347	8347
[CDMV4]	DRUG_COST	714085	[CDMV5]	drug_cost	733711	19626
[CDMV4]	DRUG_ERA	255640	[CDMV5]	drug_era	272178	16538
[CDMV4]	DRUG_EXPOSURE	729950	[CDMV5]	drug_exposure	750925	20975
[CDMV4]	LOCATION	54	[CDMV5]	location	54	0
None	NULL	NULL	[CDMV5]	measurement	108865	108865
[CDMV4]	OBSERVATION	3811	[CDMV5]	observation	38243	34432
[CDMV4]	OBSERVATION_PERIOD	6968	[CDMV5]	observation_period	6968	0
[CDMV4]	PAYER_PLAN_PERIOD	22459	[CDMV5]	payer_plan_period	22459	0
[CDMV4]	PERSON	6000	[CDMV5]	person	6000	0
[CDMV4]	PROCEDURE_COST	777007	[CDMV5]	procedure_cost	777007	0
[CDMV4]	PROCEDURE_OCCURRENCE	766837	[CDMV5]	procedure_occurrence	658507	-108330
[CDMV4]	PROVIDER	7170225	[CDMV5]	provider	7170225	0
[CDMV4]	VISIT_OCCURRENCE	376185	[CDMV5]	visit_occurrence	376185	0

Row counts changes due to vocabulary driven conversion process

Reviewing the results



TableName	Domain	RowCount	COUNT(*) from V4	Difference
Condition_Occurrence	condition	464849		
Condition_Occurrence	measurement	8416		
Condition_Occurrence	observation	31522		
Condition_Occurrence	procedure	24298		
	TOTAL	***529085***	***529057***	*28*
Drug_Exposure	drug	729950		
	TOTAL	***729950***	***729950***	*0*
Observation	measurement	3811		
	TOTAL	***3811***	***3811***	*0*
Procedure_Occurrence	device	8347		
Procedure_Occurrence	drug	20975		
Procedure_Occurrence	measurement	96638		
Procedure_Occurrence	observation	6721		
Procedure_Occurrence	procedure	634209		
	TOTAL	***766890***	***766837***	*53*

Reviewing the results

TableName	Domain	RowCount	COUNT(*) from V4	Difference
Condition_Occurrence	condition	464849		
Condition_Occurrence	measurement	8416		
Condition_Occurrence	observation	31522		
Condition_Occurrence	procedure	24298		
	TOTAL	***529085***	***529057***	*28*
Drug_Exposure	drug	729950		
	TOTAL	***729950***	***729950***	*0*
Observation	measurement	3811		
	TOTAL	***3811***	***3811***	*0*
Procedure_Occurrence	device	8347		
Procedure_Occurrence	drug	20975		
Procedure_Occurrence	measurement	96638		
Procedure_Occurrence	observation	6721		
Procedure_Occurrence	procedure	634209		
	TOTAL	***766890***	***766837***	*53*

Domain (V5 Target)	Rowcount Sum	V5 Rowcount	Difference
condition	464849	464849	0
device	8347	8347	0
drug	750925	750925	0
measurement	108865	108865	0
observation	38243	38243	0
procedure	658507	= 658507	0

Rows from V4 that map to target tables in V5 with row counts (634209 + 24298 = 605507)

Special Thanks

- Patrick Ryan & Chris Knoll – script authors
- Lee Evans – Testing Environment

Questions or comments

Questions or comments?

Thanks!