Simulacra and Simulation
How simulated data can enable OHDSI application development, methods research, and user adoption.

BACKGROUND

- Patient level data is not readily available to all members of the OHDSI community
- Limitations on distribution of data due to licensing or privacy restrictions impedes the community's ability to share research methods and tools
- Existing synthetic data have lacked adequate flexibility and update frequency for the purposes of community members

METHODS

- Synthea is software capable of producing a "source of synthetic electronic health records that is readily available; suited to industrial, innovation, research, and educational uses; and free of legal, privacy, security, and intellectual property restrictions."
- We leverage Synthea to produce a clinically reasonable, customizable patient level data set
- We leveraged standard OHDSI ETL tools during the development including WhiteRabbit to produce documentation on the conversion logic (figure 1)
- We developed the ETL Synthea Builder R package to convert data from Synthea CSV to OMOP CDM

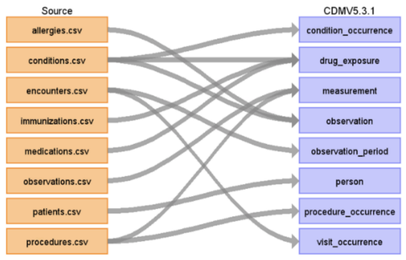


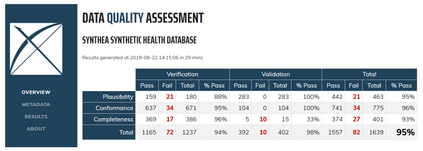
Figure 1

We created a process and produced realistic simulated data to support training, testing, and development across the OHDSI community.

Scan me

RESULTS

- Leveraging Synthea provides unique functionality previously lacking from any other simulated data approach including on-demand updates, scaling, and custom modules
- Data sets were tested and shown to support use in WebAPI, ATLAS, and R Methods Packages such as CohortMethod and Data Quality Dashboard (figure 2)



	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	159	21	180	88%	283	0	283	100%	442	21	463	95%
Conformance	637	34	671	95%	104	0	104	100%	741	34	775	96%
Completeness	369	17	386	96%	5	10	15	33%	374	27	401	93%
Total	1165	72	1237	94%	392	10	402	98%	1557	82	1639	**95%**

Figure 2

SCALABILITY

- The scalability of the ETL Synthea Builder solution was tested by generating a 100K patient data set and a 2.7M patient data set using Synthea and transforming that data to the OMOP CDM format (figure 3, table 1)

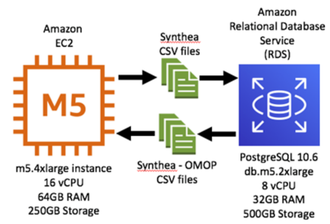


Figure 3

Synthea Data Set Size	Generation Time	ETL Conversion Time	Total AWS Cost
100,000 patients	25 minutes	18 minutes	~$0.66
2.7M patients	8 Hours	6 Hours	~$12.00

Table 1

Frank J. DeFalco[1], Clair Blacketer1, Anthony Molinaro[1], James Wiggins[2]
[1]Janssen Research & Development, Raritan, NJ; [2]Amazon, Seattle, WA