



Empowering RWD Researchers with Simulated Data

Yuichi Koretaka, Shin Kurosawa
SHIONOGI Co., Ltd.

惟高裕一, 黒澤晋
塩野義製薬株式会社

- Some RWD-related regulation had been changed in recent years
- Using RWD is inevitable for all functions (Safety, R&D, Marketing etc.)

- However, some people don't know...
 - What is RWD
 - What information is included
 - What is limitation
 - etc...
- There are some barriers, so people cannot use RWD easily



Free RWD simulator called Synthea^{*1} will help them!!



NOTE : Synthea is not suitable or appropriate for research into diseases not covered by the project or research focused on clinical discovery.

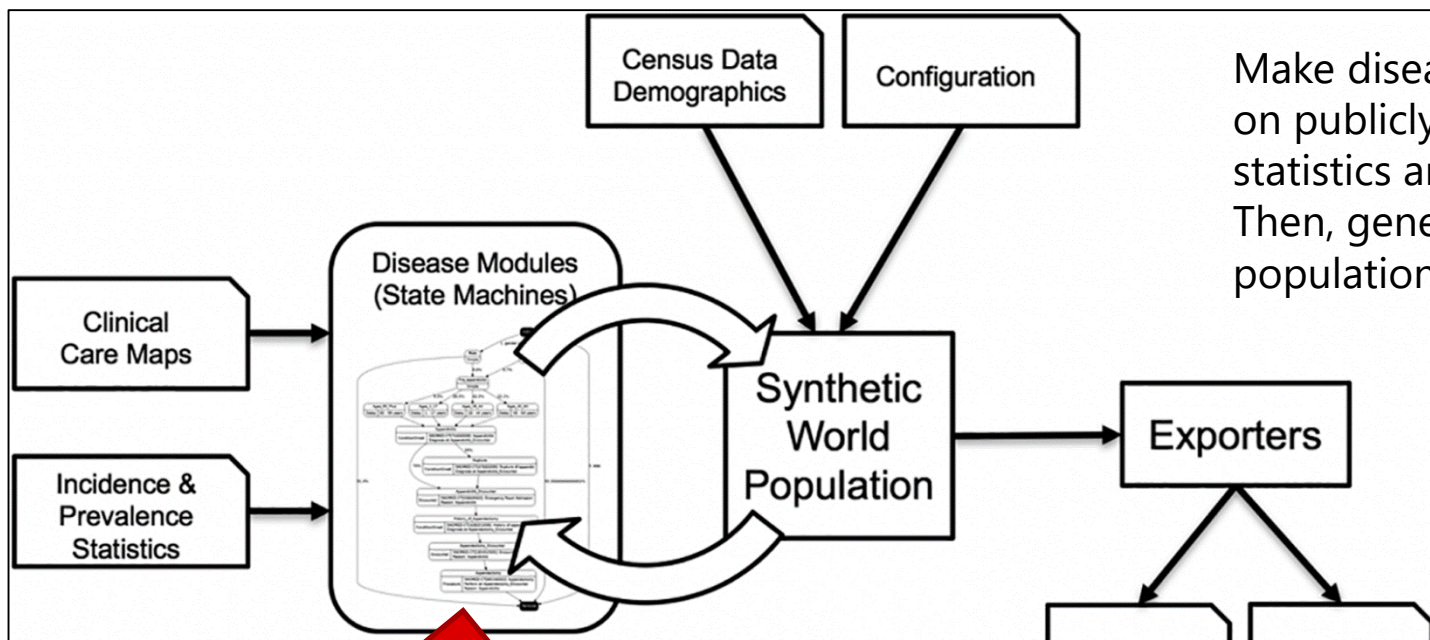
Synthea -concept-

- Synthea is a patients data and associated health records generator covering every aspect of healthcare
- To output high-quality synthetic, realistic but not real
- The resulting data is free from cost, privacy, and security restrictions.
- It can be used without restriction for a variety of secondary uses in academia, research, industry, and government
- Everyone can use through the Github

<https://github.com/synthetichealth/synthea/wiki>

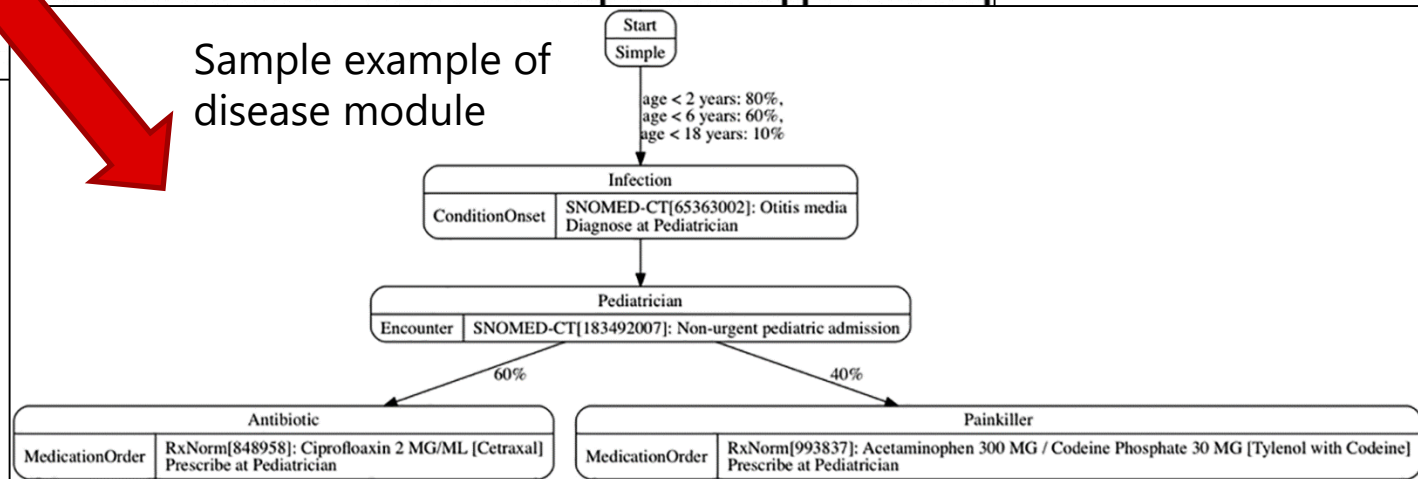


Synthea -architecture-



Make disease modules based on publicly available health statistics and clinical guidelines. Then, generate synthea population.

Sample example of disease module



Synthea -data and tables -

Generated patients is mapped to each table

Example of a synthea patient data

OBSERVATIONS:

2016-11-14	: Body Height	157.5	cm
2016-11-14	: Body Weight	104.3	kg
2016-11-14	: Body Mass Index	42.0	kg/m2
2016-11-14	: Systolic Blood Pressure	198.0	mmHg
2016-11-14	: Diastolic Blood Pressure	107.0	mmHg
2016-11-14	: Hemoglobin Alc/Hemoglobin.total in Blood	8.3	%
2016-11-14	: Glucose	133.0	mg/dL
2016-11-14	: Urea Nitrogen	13.0	mg/dL
2016-11-14	: Creatinine	1.0	mg/dL
2016-11-14	: Calcium	9.4	mg/dL
2016-11-14	: Sodium	136.0	mmol/L
2016-11-14	: Potassium	4.5	mmol/L
2016-11-14	: Chloride	102.0	mmol/L
2016-11-14	: Carbon Dioxide	27.0	mmol/L
2016-11-14	: Basic Metabolic Panel		
2016-11-14	: Total Cholesterol	243.0	mg/dL
2016-11-14	: Triglycerides	340.0	mg/dL
2016-11-14	: Low Density Lipoprotein Cholesterol	145.0	mg/dL
2016-11-14	: High Density Lipoprotein Cholesterol	30.0	mg/dL
2016-11-14	: Lipid Panel		
2016-11-14	: Microalbumin Creatine Ratio	2.0	mg/g
2016-11-14	: Estimated Glomerular Filtration Rate	>60	mL/min/

PROCEDURES:

2014-11-23	: Documentation of current medications
2011-01-02	: Documentation of current medications
2007-11-19	: Documentation of current medications

ENCOUNTERS:

2016-11-14	: Outpatient Encounter
2015-09-14	: Outpatient Encounter
2015-03-23	: Outpatient Encounter

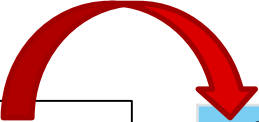


Table name	Description
Allergies	Patient allergy data.
Careplans	Patient care plan data, including goals.
Conditions	Patient conditions or diagnoses.
Encounters	Patient encounter data.
imaging_studies	Patient imaging metadata.
Immunizations	Patient immunization data.
Medications	Patient medication data.
Observations	Patient observations including vital signs and lab reports.
Organizations	Provider organizations including hospitals.
Patients	Patient demographic data.
Payer_transitions	Payer Transition data (i.e. changes in health insurance).
Payers	Payer organization data.
Procedures	Patient procedure data including surgeries.
Providers	Clinicians that provide patient care.

Synthea -execution-

- To clone the Synthea, build and run the test

```
git clone https://github.com/synthetichealth/synthea.git  
cd synthea  
./gradlew build check test
```



When the check test has finished, you will see **"BUILD SUCCESSFUL"**

- Change config settings to generate result as CSV

```
cd /users/[username]/synthea/src/main/resources  
vi synthea.properties
```



Make **"exporter.csv.export=true"**

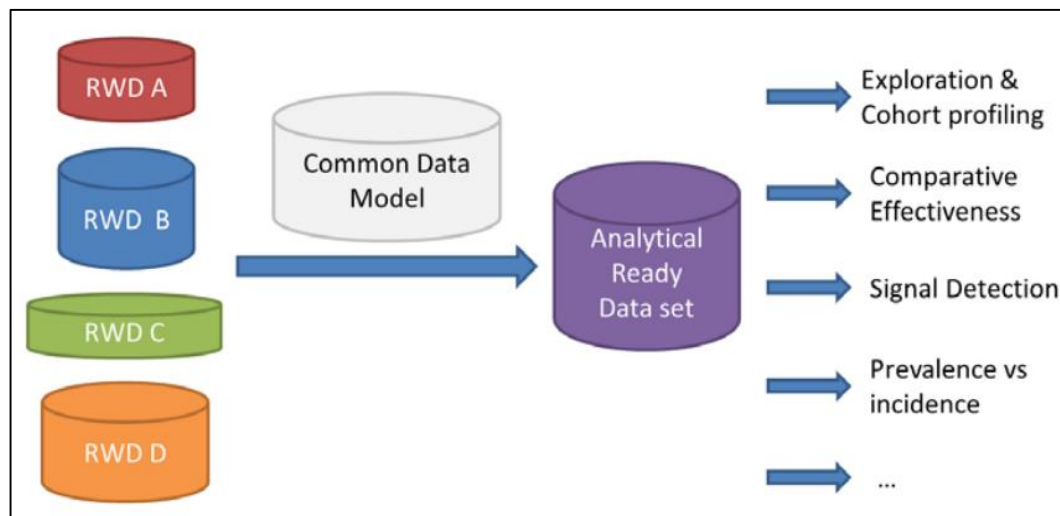
Then, all outputs will be generated in **/users/[username]/synthea/output/csv**

[EXAMPLE] Generate 100000 patients data

```
cd /users/[username]/synthea/  
./run_synthea -p 100000
```

Synthea -CDM-

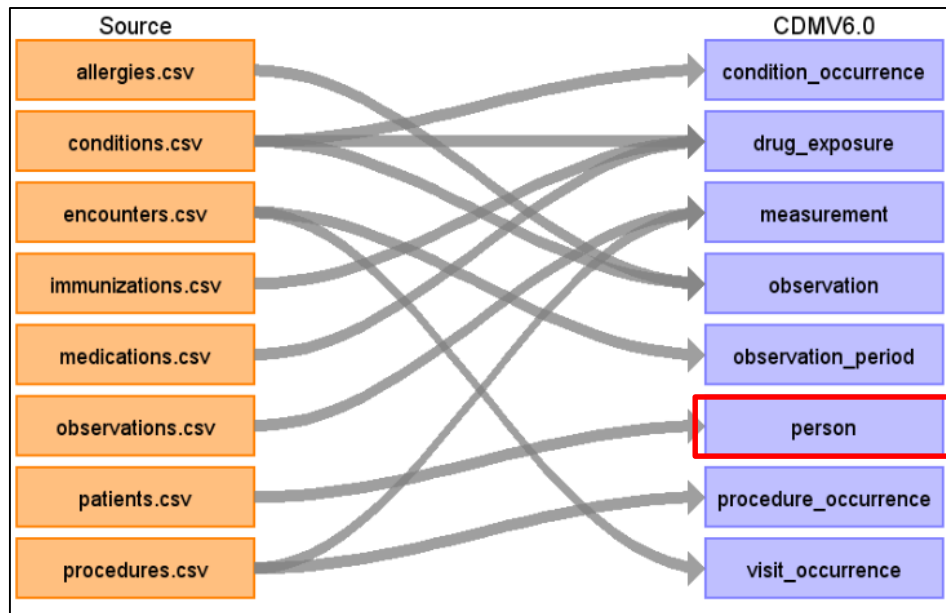
- To improve analysis productivity and speed, we can use standardized data model called common data model (CDM). CDM allows the use of common analytics and methods across multiple RWD datasets.



- Synthea can be converted to OMOP-CDM by using Synthea-ETL builder.

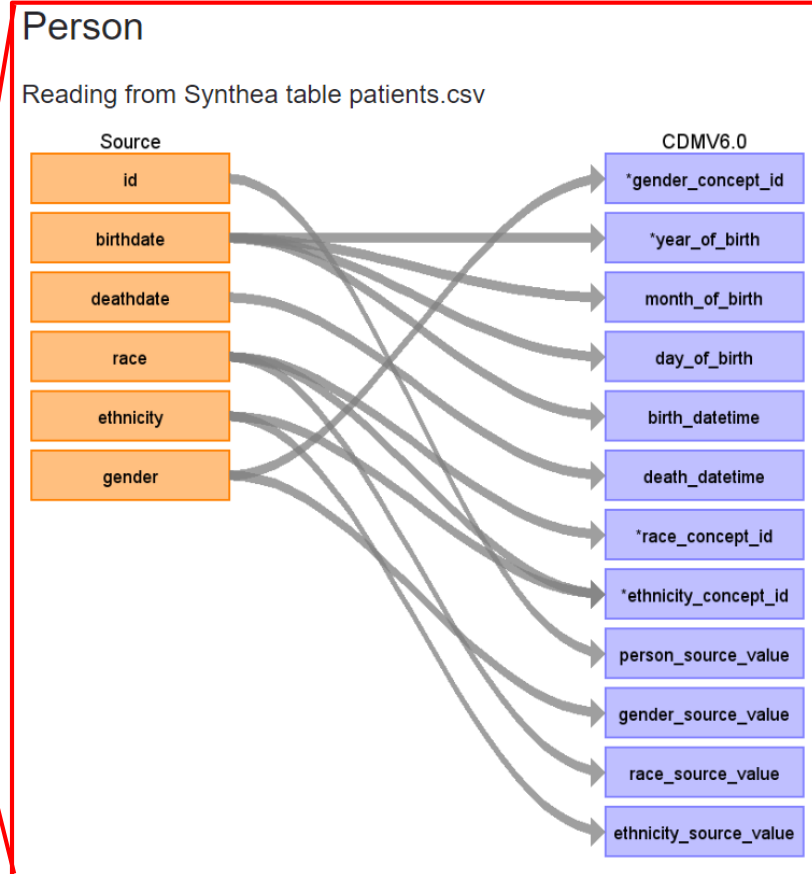
Synthea ETL Builder -concept-

- Synthea tables are mapped to its corresponding OMOP-CDM tables



OMOP-CDM are continuously developing and updating

Example of person table



Synthea ETL Builder -execution- (1/2)



[Preparation]

① Build PostgreSQL DB

- ✓ Make role as "Username: lollipop" and "PW: passwd".
- ✓ Make blank schema "cdm_synthea10" and "native".

② Download CDM vocabulary from Athena (<http://athena.ohdsi.org/>)

- ✓ Athena is standardized vocabularies for CDM



- ETL programs are developed by R, so install packages

```
devtools::install_github("OHDSI/ETL-Synthea")  
library(ETLSyntheaBuilder)
```



If you have errors for installing devtools,

remotes::install_github("OHDSI/ETL-Synthea") might work well

- Connect PostgreSQL DB

```
cd <- DatabaseConnector::createConnectionDetails( dbms = "postgresql", server =  
"localhost/synthea10", user = "postgres", password = "lollipop", port = 5432 )  
ETLSyntheaBuilder::DropVocabTables(cd,"cdm_synthea10")
```

Synthea ETL Builder -execution- (1/2)



- Run ETL programs

```
ETLSyntheaBuilder::DropEventTables(cd,"cdm_synthea10") ETLSyntheaBuilder::DropSyntheaTables(cd,"native")
ETLSyntheaBuilder::DropMapAndRollupTables (cd,"cdm_synthea10")
ETLSyntheaBuilder::CreateVocabTables(cd,"cdm_synthea10") ETLSyntheaBuilder::CreateEventTables(cd,"cdm_synthea10")
ETLSyntheaBuilder::CreateSyntheaTables(cd,"native")
ETLSyntheaBuilder::LoadSyntheaTables(cd,"native","/tmp/synthea/output/csv")
ETLSyntheaBuilder::LoadVocabFromCsv(cd,"cdm_synthea10","/tmp/Vocabulary_20181119")
ETLSyntheaBuilder::CreateVocabMapTables(cd,"cdm_synthea10")
ETLSyntheaBuilder::CreateVisitRollupTables(cd,"cdm_synthea10","native")
ETLSyntheaBuilder::LoadEventTables(cd,"cdm_synthea10","native")
```



If you have error like **"ERROR: column "provider_id" is of type integer but expression is of type text"**, you should define provider_id as 0 for each table in the SQL program because NULL seems not to be appropriate.

See example ↓

```
p.person_id,
case when srctostdvm.target_concept_id is NULL then 0 else srctostdvm.target_concept_id end a
c.start drug_exposure_start_date,
c.start drug_exposure_start_datetime,
coalesce(c.stop,c.start) drug_exposure_end_date,
coalesce(c.stop,c.start) drug_exposure_end_datetime,
c.stop verbatim_end_date,
581452 drug_type_concept_id,
cast(null as varchar) stop_reason,
0 refills,
0 quantity,
coalesce(datediff(day,c.start,c.stop),0) days_supply,
cast(null as varchar) sig,
0 route_concept_id,
0 lot_number,
0 provider_id,
#NULL provider_id,
fv.visit_occurrence_id_new visit_occurrence_id,
```

- Synthea is suitable for learning RWD because it's free from cost, privacy and security restrictions
- To grasp pros and cons of RWD is important
- CDM might be one of the useful approach for effective analysis

[Paper]

- Jason et al.(2017), download at <https://doi.org/10.1093/jamia/ocx079>

[Github]

- <https://github.com/synthetichealth/synthea>
- <https://github.com/synthetichealth/synthea/wiki>
- <https://synthetichealth.github.io/synthea/>
- <https://github.com/OHDSI/Athena>