# Synthetic Data Generation – OSIM 5

Kausar Mukadam, Jon Duke M.D
Georgia Tech Research Institute

Community Meeting - 4/17/2018

# Why synthetic data?

- Lack of benchmark datasets for research
- Privacy concerns for data sharing
- Data from commercial vendors is not easily accessible

# Different needs!


Less realistic data may be sufficient

Realistic, patient level data is needed

# Classification of synthetic data

## Algorithm

- Knowledge based
- Data Driven

## Data Source

- Patient Level Data
- Population Level Statistics

# Some available generators

SYNTHEA

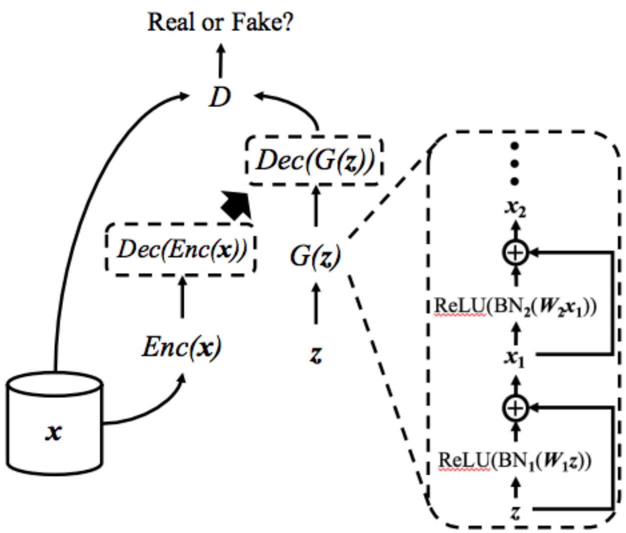
OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

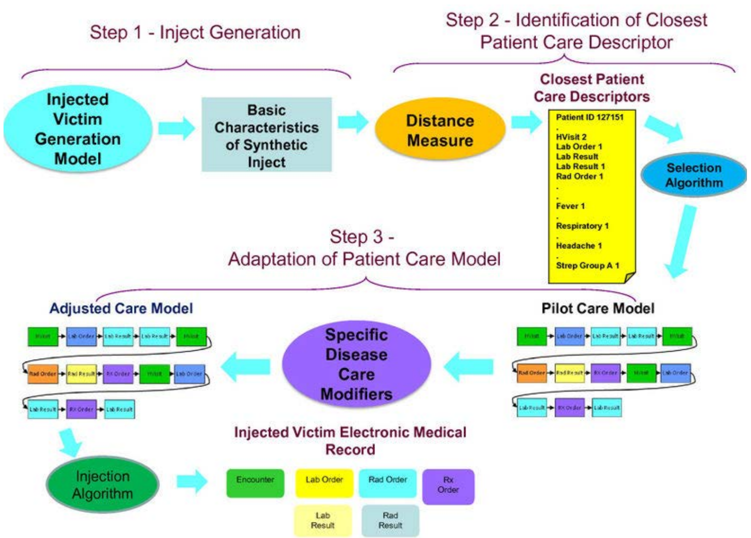
MiHIN

Shared Services

# Other research




Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks.




Anna L Buczak, Steven Babin and Linda Moniz. Data-driven approach for creating synthetic electronic medical records

# OSIM

- Simulated datasets modeled on real observational data sources
- Contain synthetic persons with condition and drug occurrences
- Based on random sampling from probability distributions
- Probability distributions defined on relationships between the actual conditions and drugs
- Originally implemented for OMOP v1 and v2[1]

[1] Richard E Murray, Patrick B Ryan, and Stephanie J Reisinger. Design and Validation of a Data Simulation Model for Longitudinal Healthcare Data. AMIA Annu Symp Proc. 2011; 2011: 1176–1185.

# Overview

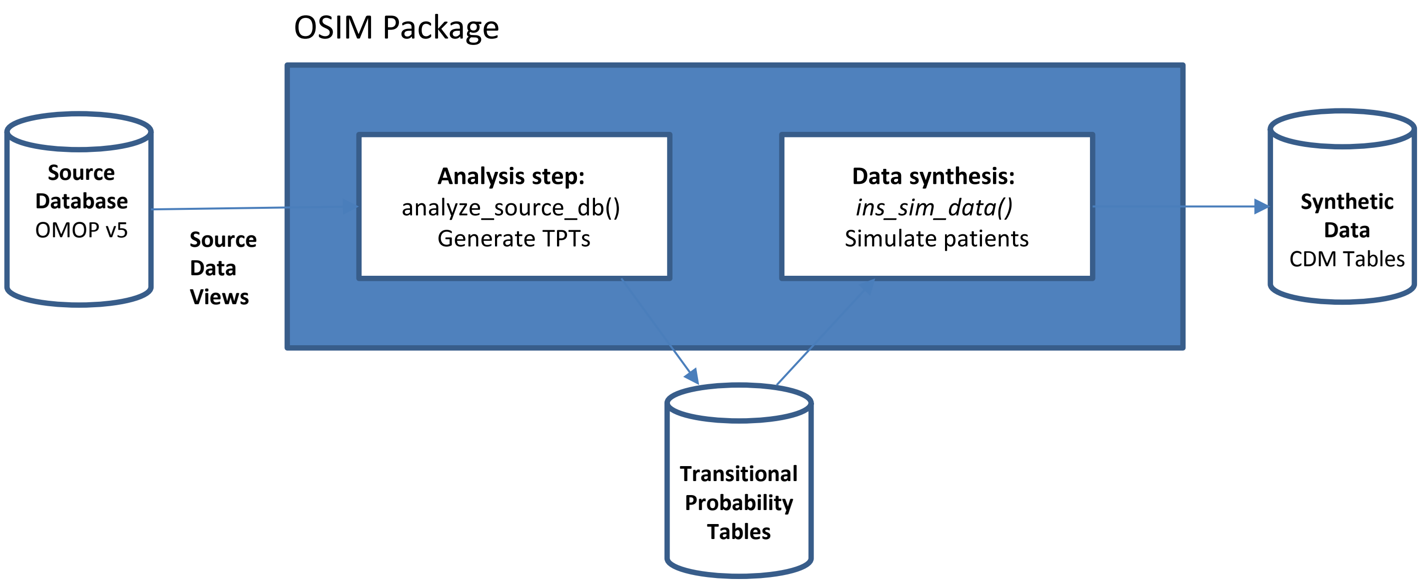



Figure 1. Process flow for OSIM

# Analysis

- Preliminary analysis of OMOP CDM data source to record characteristics
- **analyze_source_db()**: generates 18 tables that document transitional probabilities

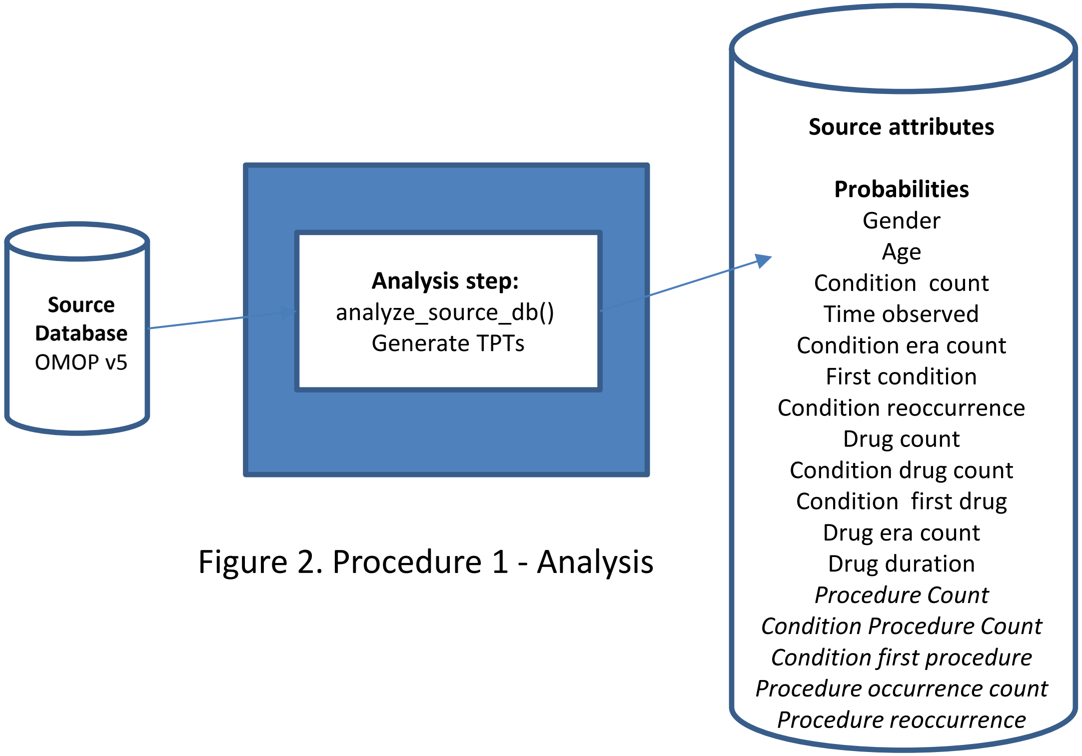



Figure 2. Procedure 1 - Analysis

# Analysis

- **src_db_attributes:** Number of persons, drug eras & condition eras, min & max dates
- **gender_prob:** P(gender concept id)
- **age_at_obs_prob:** P(age at obs | gender concept id)
- **cond_count_prob:** P(cond concept count | gender concept id, age at obs)
- **time_obs_prob:** P(time observed | gender concept id, age at obs, cond count bucket)
- **first_cond_prob:** P(condition2 concept id, delta days | gender concept id, age range, cond count bucket, time remaining, condition1 concept id)
- **cond_era_count_prob:** P(cond era count | condition concept id, cond count bucket, time remaining)

# Analysis

- **cond_reoccur_prob:** P(delta days | condition concept id, age range, time remaining)
- **drug_count_prob:** P(drug count | gender concept id, age range, cond count bucket)
- **cond_drug_count_prob:** P(drug draw count | condition concept id, interval bucket, age range, drug count bucket, cond count bucket)
- **cond_first_drug_prob:** P(drug concept id, delta days | condition concept id, interval bucket, gender concept id, age range, condition count bucket, drug count bucket, day cond count)
- **drug_era_count_prob:** P(drug era count, total exposure | drug concept id, drug count bucket, condition count bucket, age range, time remaining)
- **drug_duration_prob:** P(total duration | drug concept id, time remaining, drug era count, total exposure)

- **5 additional tables for procedure occurrence generation**

# Data Generation

**01 Person**
Simulate gender and age

**02 Observation Period**
Simulate number of conditions to be drawn and observation duration

**03 Condition Era**
Simulate conditions and reoccurrences

**04 Drug Era**
Simulate drugs for every condition

**Procedure Occurrence**
Simulate procedures for every condition

# Data Generation: Person, Observation Period

- Person: Random draw for gender & age

**Gender:**
osim_gender_prob
*P(gender concept id)*

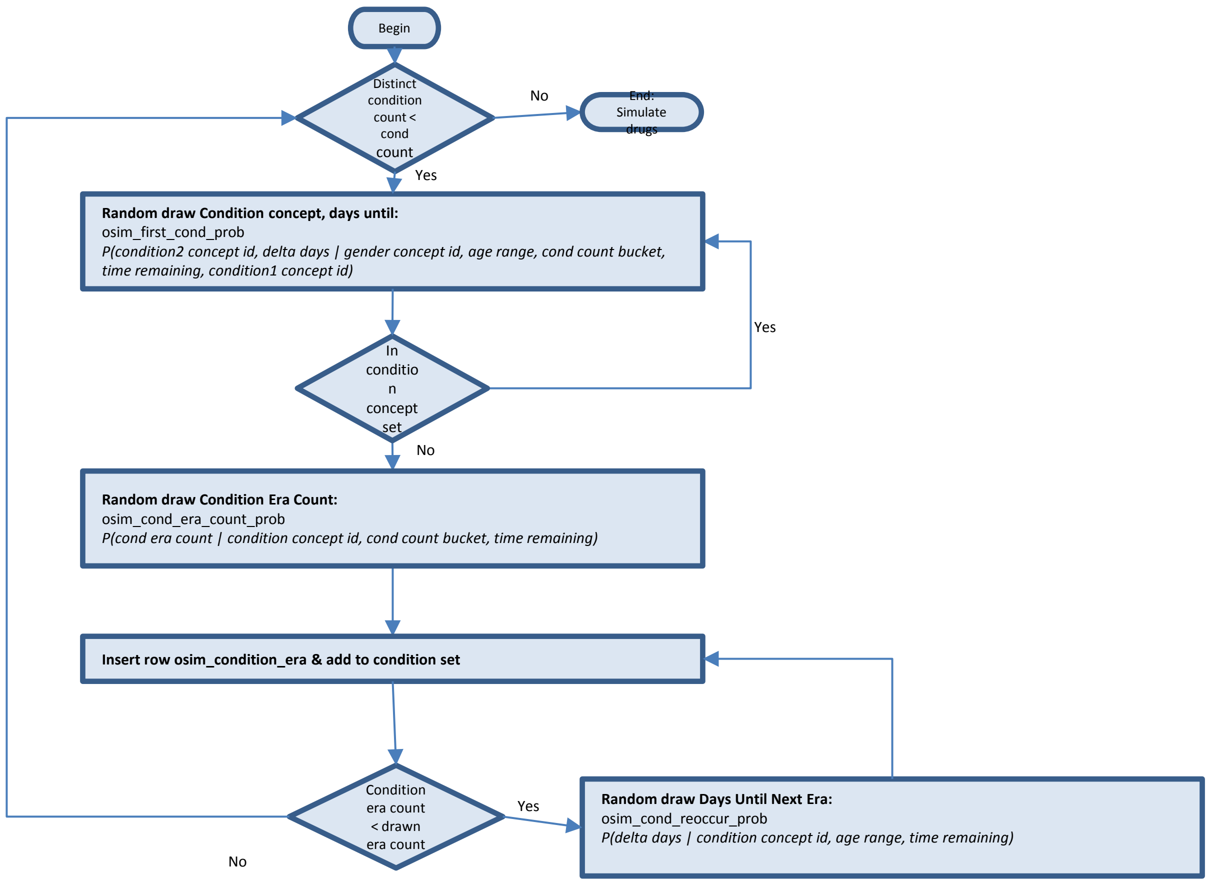
**Age:**
osim_age_at_obs_prob
*P(age at obs | gender concept id)*

- Observation Period: Random draw for condition count and observation duration

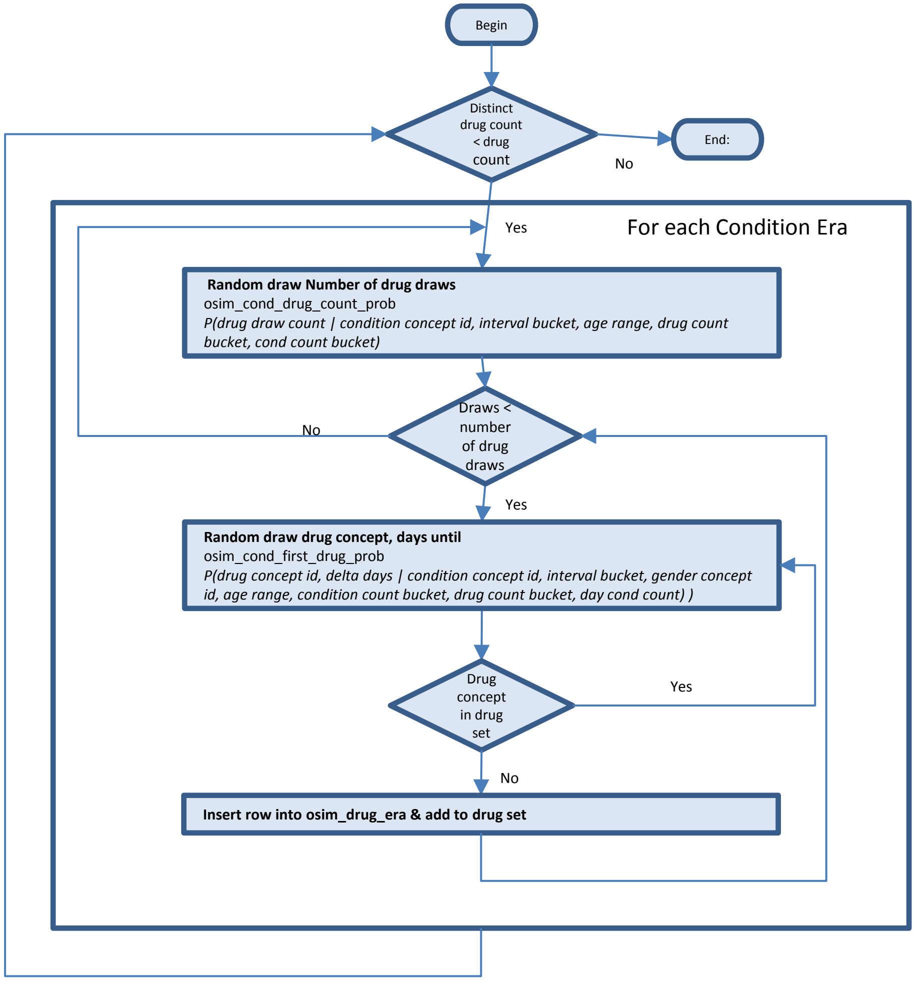
**Condition Count:**
osim_cond_count_prob
*P(cond count| gender concept id, age at obs)*

**Observation period:**
osim_time_obs_prob
*P(time observed | gender, age, cond count)*

# Data Generation: Condition Era

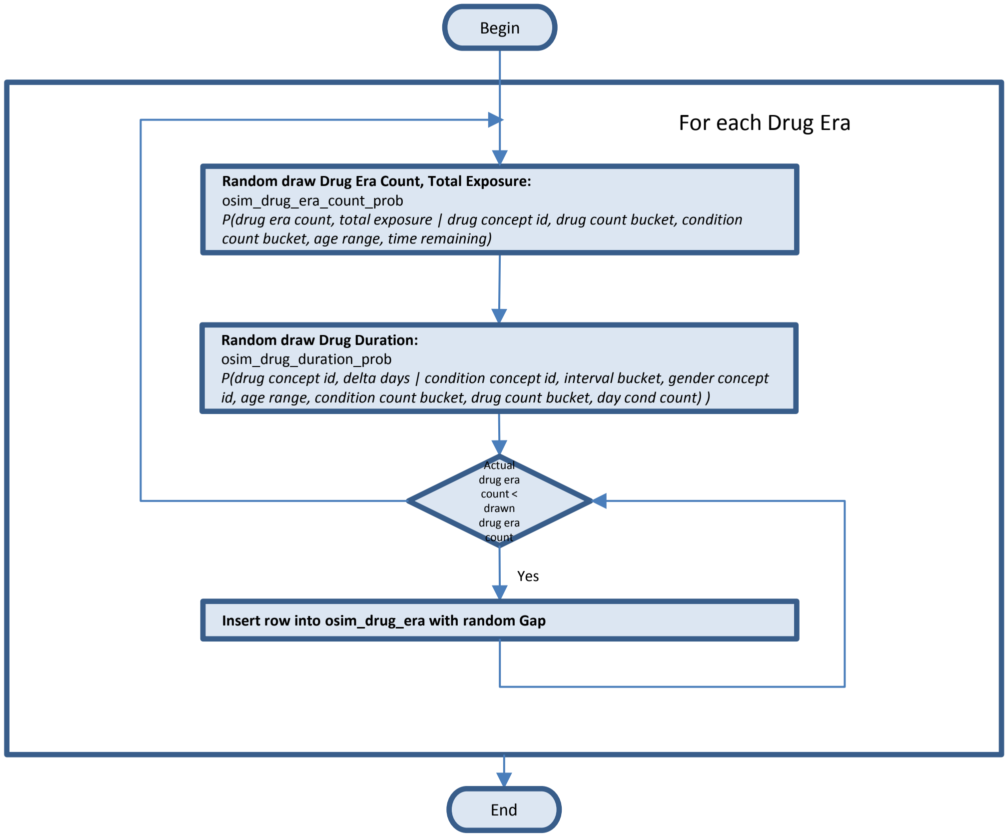


# Data Generation: Drug Era



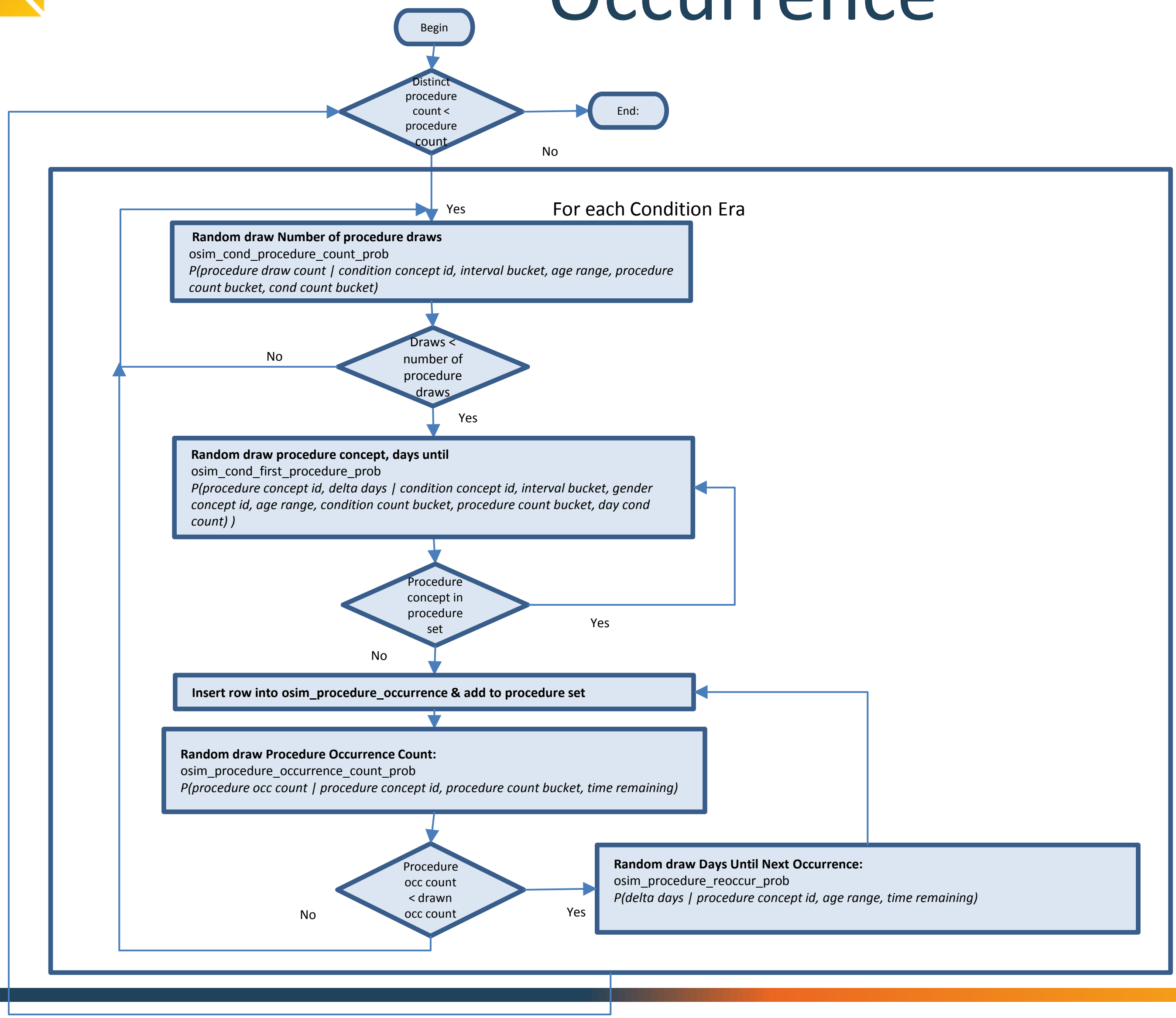

Simulate first occurrence drug eras

# Data Generation: Drug Era




Simulate subsequent drug eras

# Data Generation: Procedure Occurrence



# Data Generation: Some Potential Issues

- **Procedure Re-occurrence**
- **Explore Visits:** Generate visits first, and for each visit generate conditions, drugs, procedures
- **Drugs & Procedures Relationship:** Currently assumed that condition to procedure relationship encompasses the drugs - procedures relationship, which may not always be the case

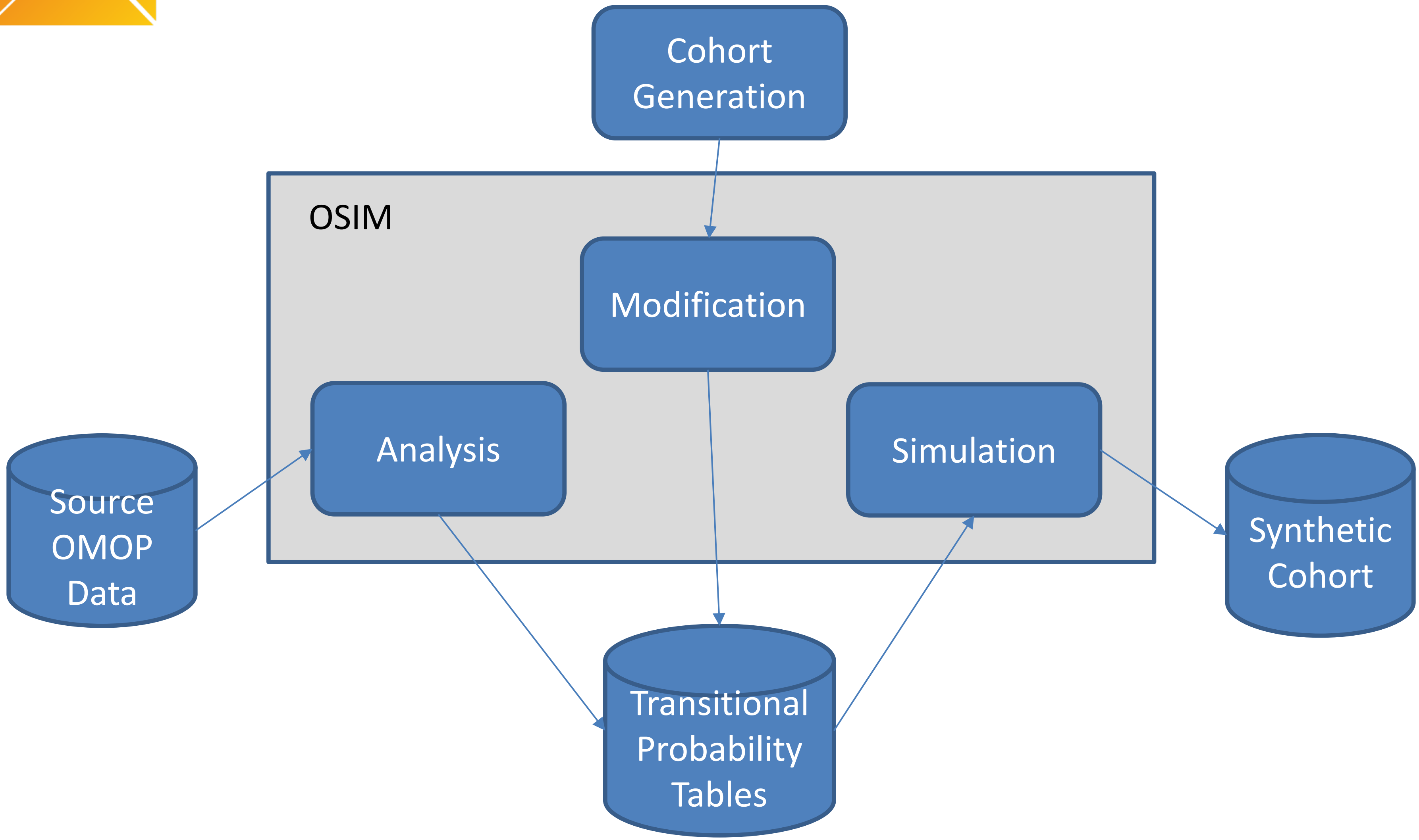
# Version 2 -> Version 5

- Uses OMOP CDM v5 format as data source
- Current only available in PostgreSQL
- Persistence
- Procedure Occurrence
- Cohort Generation

# Next Steps

- Generate data in visits
- Include Observations & Measurements in synthetic data
- Challenges
  - Condition each subsequent visit on the previous visits drugs & conditions
  - Complex relationships between condition, observation & measurement
  - Cannot be easily modeled based on existing probability framework

# Next Steps: Cohort Generation



# Datasets Available!

- Pilot v5 datasets are available for SynPUF and Mimic
- Currently without procedure occurrence – person, observation period, drug era and condition era available
- Datasets for Truven, including procedure occurrence, will be available soon

| | Mimic | Synpuf | Truven |
|---|---|---|---|
| Source Patients | 46520 | 2,326,856 | 81,826,982 |
| Synthetic Patients Available | 10,000 | 10,000 | In Progress! |

# Things to remember!

- OSIM **approximates** the true complex relationships between conditions, drugs and disease progression
- The simulated data has some **missing fields** like race, ethnicity, etc.
- Data simulation for measurements, observations, and other CDM tables is not currently implemented
- Models **population level similarities**, so analyses results on a real dataset may be worse

# Some resources

OSIM

- **OSIM v2:** ftp://ftp.ohdsi.org/osim2/
- **OSIM v5:** **https://github.com/OHDSI/OSIM-v5**
- **Design and Validation of a Data Simulation Model for Longitudinal Healthcare Data:** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243118/
- **Pilot Datasets:** https://github.gatech.edu/HDAP/synthetic-datasets

Other synthetic data generators

- **Synthea:** https://github.com/synthetichealth/synthea
- **PatientGen:** https://mihin.org/services/patient-generator/
- **Medgan:** https://github.com/mp2893/medgan

# Questions?