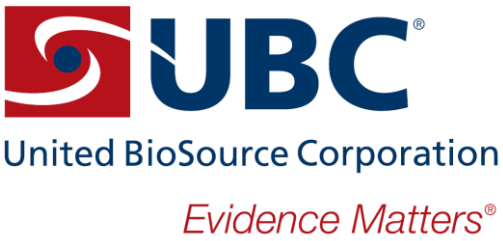
Data Dictionary for the Enhanced Observational Medical Dataset Simulator (OSIM 2) v1.5.005



OBSERVATIONAL MEDICAL OUTCOMES PARTNERSHIP
A PUBLIC PRIVATE RESEARCH PARTNERSHIP OF THE FNIH

Last Revised: 20 January 2011

Table of Contents

Introduction to the Enhanced Observational Medical Dataset Simulator Project 4
Overview of OSIM 2 Package 5
Overview of OSIM 2 Package 5
OSIM 2 Oracle Package Header 5
Overview of OSIM 2 Components 7
OSIM 2 CDM Source Data Synonyms and Views 8
OSIM 2 Source Data Access Layers 8
OSIM 2 Source Data Synonyms 8
OSIM 2 Source Data Views 9
VIEW v_src_person 9
VIEW v_src_person_strata 9
VIEW v_src_observation_period 11
VIEW v_src_condition_era1_ids 11
VIEW v_src_condition_era1 12
VIEW v_src_first_conditions 12
VIEW v_src_all_conditions 13
VIEW v_src_drug_era1 13
VIEW v_src_drug_era1_ids 13
VIEW v_src_first_drugs 14
Execution Parameters 15
Program Execution Parameters 15
OSIM Execution Parameters: 15
User-Modifiable Range Functions 16
OSIM 2 Process Log 18
Table osim_log 18
General Source Database Attributes Table 19
Table osim_src_db_attributes 19
Transition Probability Tables 20
Table osim_gender_probability 21
Table osim_age_at_obs_probability 22
Table osim_cond_count_probability 23
Table osim_time_obs_probability 24
Table osim_first_cond_probability 25
Table osim_cond_era_count_prob 26

Table osim_cond_reoccur_probability 27
Table osim_drug_count_prob 28
Table osim_cond_drug_count_prob 29
Table osim_cond_first_drug_prob 30
Table osim_drug_era_count_prob 31
Table osim_drug_duration_probability 32
Simulated CDM Tables 33
OSIM_PERSON 34
OSIM_OBSERVATION_PERIOD 35
OSIM_CONDITION_ERA 36
OSIM_DRUG_ERA 37
Outcomes Table (Optional) 38
TABLE OSIM_DRUG_OUTCOME 38
Revision History 39

Introduction to the Enhanced Observational Medical Dataset Simulator Project

There is great interest in developing new methods for active drug safety surveillance, but to date, little work has empirically evaluated the performance of alternative approaches. Methodological research requires assessment of methods by comparing predictions to a known reference set. In the context of active surveillance, the primary challenge in using real observational data is that the true relationships between drugs and conditions may be unknown. The promise of simulation is to produce data that sufficiently models real-world phenomenon, but through defined relationships that can be fully characterized. Simulation models nearly always require simplification from the real world, but can be used as a foundation for studying real-world effects in a controlled environment.

The successful development of the Observational Medical Dataset Simulator (OSIM; http://omop.fnih.org/OSIM) has enabled methods development across a broad research community. OSIM can construct many millions of hypothetical patients, with drug exposure, background conditions, and known adverse events that can be used to benchmark method performance. OSIM has provided access to large-scale data to methodologists, and facilitated the establishment of the OMOP Cup Competition. It has also advanced the OMOP Research Team's learnings about the complex interdependencies between clinical observations in real data, and how those relationships may influence a method's behavior in identifying true associations and discerning from false positive findings. Based on these learnings, the Principal Investigators have recommended that OMOP continue research and development into a second-generation simulated dataset procedure. OSIM 2 establishes a complementary model to the original OSIM module, applying an alternative design to accommodate additional complexities observed in real-world data, including advanced modeling of the correlations between drugs and conditions. OSIM 2 should allow for more direct comparisons between simulated data and real observational databases, and should enable greater method evaluation by allowing assessment of how methods accommodate these complex interrelationships.

This document provides a Data Dictionary describing the files and tables required by the OSIM 2 package, and the output produced by the program.

Overview of OSIM 2 Package

OSIM 2 is an Oracle package containing Types, Functions, and Procedures that are accessed externally by users and internally by the package itself. The package can be used by executing just a few of the procedures. Several of the functions can be modified by the user to customize strata ranges.

OSIM 2 Oracle Package Header

```
CREATE OR REPLACE PACKAGE osim2 AS
  TYPE COND_TRANSITION IS RECORD (
    gender_concept_id      osim_first_cond_probability.gender_concept_id%TYPE,
    age_range              osim_first_cond_probability.age_range%TYPE,
    cond_count_bucket      osim_first_cond_probability.cond_count_bucket%TYPE,
    time_remaining         osim_time_obs_probability.time_observed%TYPE,
    condition1_concept_id  osim_first_cond_probability.condition1_concept_id%TYPE,
    condition2_concept_id  osim_first_cond_probability.condition2_concept_id%TYPE,
    delta_days             osim_cond_reoccur_probability.delta_days%TYPE);

  TYPE TAB_COND_TRANSITIONS IS TABLE OF COND_TRANSITION;

  TYPE TAB_DRUG_OUTCOME IS TABLE OF osim_drug_era%ROWTYPE;

  TYPE DRUG_COND_OUTCOME IS RECORD (
    person_id              osim_drug_era.person_id%TYPE,
    drug_era_id            osim_drug_era.drug_era_id%TYPE,
    condition_era_id       osim_condition_era.condition_era_id%TYPE);

  TYPE TAB_OUTCOME IS TABLE OF DRUG_COND_OUTCOME;

  TYPE CONCEPT_VECTOR IS TABLE OF NUMBER(1) INDEX BY BINARY_INTEGER;

  FUNCTION condition_count_bucket (
    condition_count NUMBER) RETURN NUMBER;

  FUNCTION drug_count_bucket (
    drug_count NUMBER) RETURN NUMBER;

  FUNCTION age_bucket (
    age NUMBER) RETURN NUMBER;

  FUNCTION time_observed_bucket (
    days NUMBER) RETURN NUMBER;

  FUNCTION round_days (
    days NUMBER) RETURN NUMBER;

  FUNCTION randomize_days (
    days NUMBER) RETURN NUMBER;

  FUNCTION duration_days_bucket (
    days NUMBER) RETURN NUMBER;

  FUNCTION min_num (
    value1 NUMBER,
    value2 NUMBER) RETURN NUMBER;

  FUNCTION get_first_cond_transitions
    RETURN TAB_COND_TRANSITIONS PIPELINED;

  FUNCTION get_outcome_drug_eras (
    drug_concept_id            NUMBER,
    condition_concept_id       NUMBER,
    outcome_risk_type          VARCHAR2,
    outcome_onset_days_min     NUMBER,
    outcome_onset_days_max     NUMBER) RETURN TAB_DRUG_OUTCOME PIPELINED;
```

```
FUNCTION get_outcome_eras (
    drug_concept_id            NUMBER,
    condition_concept_id       NUMBER,
    outcome_risk_type          VARCHAR2,
    outcome_onset_days_min     NUMBER,
    outcome_onset_days_max     NUMBER) RETURN TAB_OUTCOME PIPELINED;

  PROCEDURE insert_log (
    MESSAGE               VARCHAR2,
    stored_procedure_name VARCHAR2 := '');

  PROCEDURE ins_src_db_attributes;

  PROCEDURE ins_gender_probability;

  PROCEDURE ins_age_at_obs_probability;

  PROCEDURE ins_cond_count_probability;

  PROCEDURE ins_time_obs_probability;

  PROCEDURE ins_first_cond_probability;

  PROCEDURE ins_cond_era_count_prob;

  PROCEDURE ins_cond_days_before_prob;

  PROCEDURE ins_drug_count_prob;

  PROCEDURE ins_cond_drug_count_prob;

  PROCEDURE ins_cond_first_drug_prob;

  PROCEDURE ins_drug_era_count_prob;

  PROCEDURE ins_drug_duration_probability;

  PROCEDURE drop_osim_indexes;

  PROCEDURE create_osim_indexes;

  PROCEDURE copy_final_data(
    destination_schema VARCHAR2 DEFAULT 'osim_dev',
    total_person_count NUMBER   DEFAULT 5000);

  PROCEDURE analyze_source_db;

  PROCEDURE delete_all_sim_data;

  PROCEDURE ins_sim_data(
    person_count    NUMBER DEFAULT 5000,
    person_start_id NUMBER DEFAULT 0);

  PROCEDURE ins_outcomes;

END osim2;
```

Overview of OSIM 2 Components

The input and output data and components used and produced by the **OSIM 2** can be divided into eight types:

- OSIM 2 CDM Synonyms and Source Data Views
- Execution Parameters
- User Modifiable Range Functions
- Process Log
- General Source Database Attributes Table
- Transition Probability Tables
- Simulated CDM Tables
- Outcomes Table (optional)

Figure 1 below describes the high-level process flow of the OSIM 2, illustrating how each type of data is used by the system.

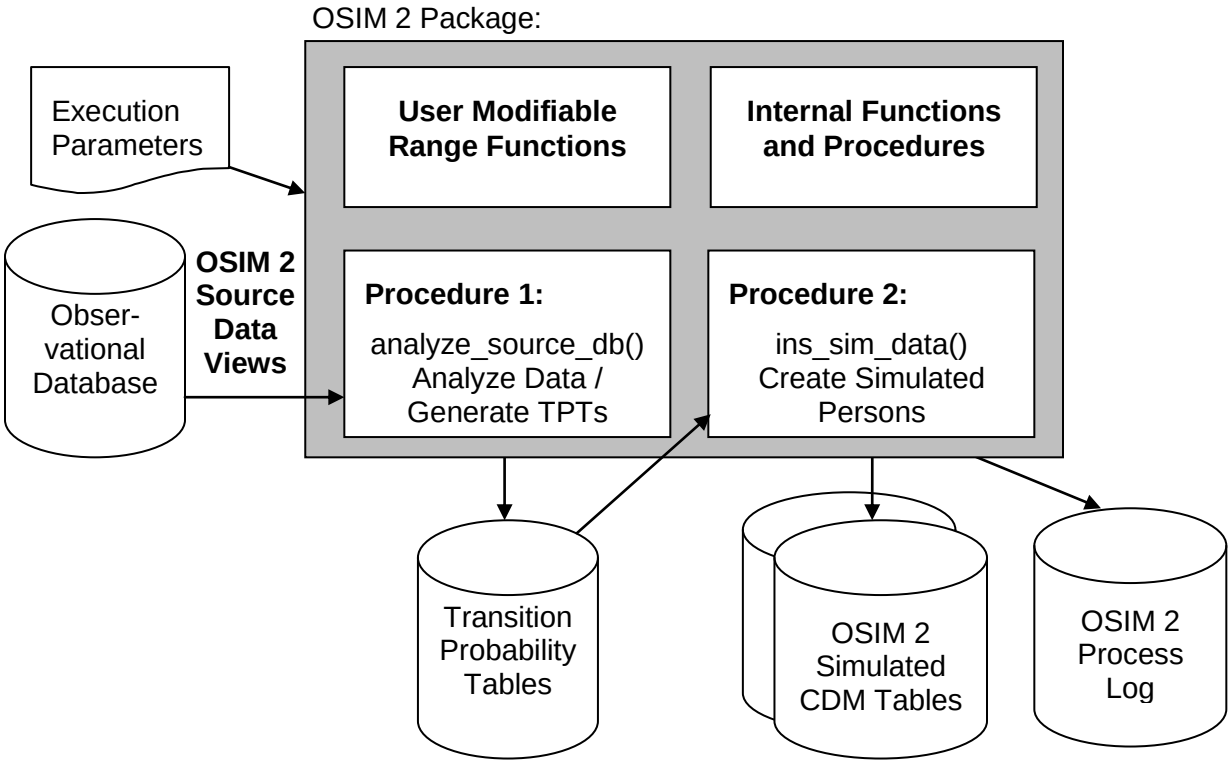


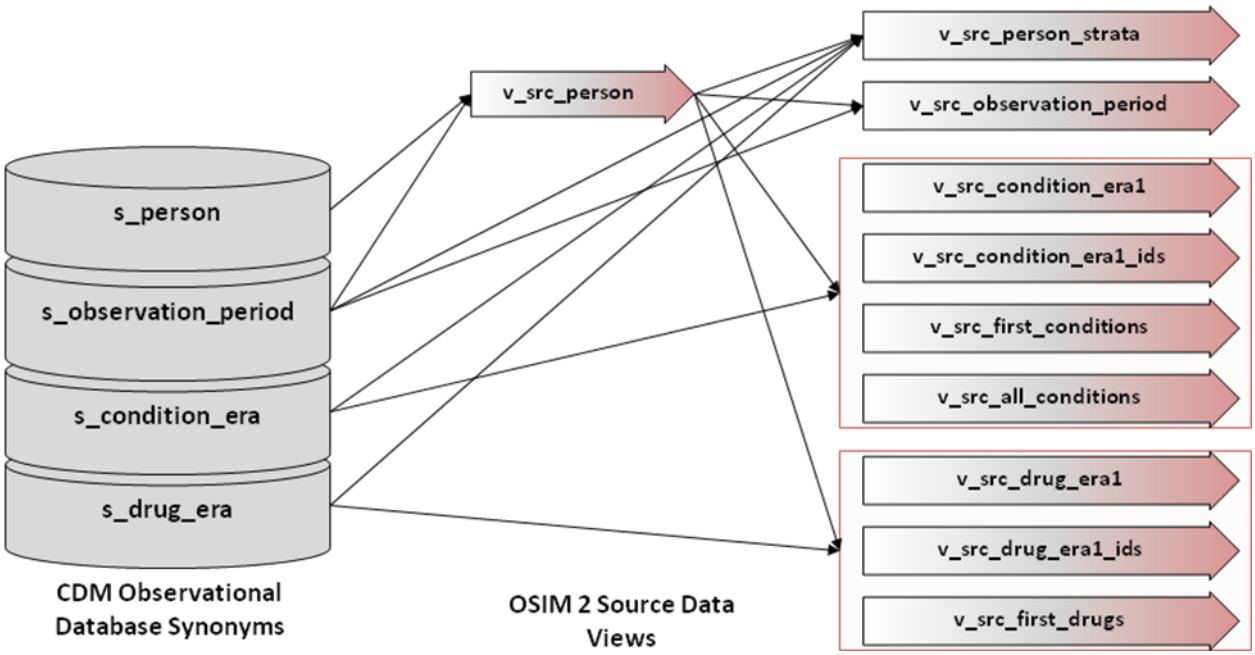
Figure 1: OSIM 2 High Level Process Flow

Each type of data is described in detail throughout the rest of the document.

OSIM 2 CDM Source Data Synonyms and Views

The OSIM 2 database analysis uses a set of read-only views to access the data in the source CDM database. These views can be modified to limit or refine the analysis, in particular for the persistence window eras. The views access the source CDM database through synonyms to the person, observation period, drug era, and condition era tables.

OSIM 2 Source Data Access Layers



OSIM 2 Source Data Synonyms

The synonyms must be created in the target schema before running the OSIM 2 analysis.

```
--
-- Example OSIM 2 CDM Synonyms
--
CREATE OR REPLACE SYNONYM s_person FOR mslr_cdm.person;
/
CREATE OR REPLACE SYNONYM s_condition_era FOR mslr_cdm.omop_condition_era;
/
CREATE OR REPLACE SYNONYM s_observation_period FOR mslr_cdm.observation_period;
/
CREATE OR REPLACE SYNONYM s_drug_era FOR mslr_cdm.omop_drug_era;
/
COMMIT;
/
```

OSIM 2 Source Data Views

The views must be created in the target schema before running the OSIM 2 analysis. The views must be created in order because of inter-dependency.

VIEW v_src_person

```
--=============================================================================
-- VIEW v_src_person
--=============================================================================
CREATE OR REPLACE VIEW v_src_person
 (person_id, year_of_birth, gender_concept_id, race_concept_id,
  location_concept_id, source_person_key, source_gender_code,
  source_location_code, source_race_code) AS
SELECT /*+ NO_PARALLEL(person) */DISTINCT
  person.person_id,person.year_of_birth, person.gender_concept_id,
  person.race_concept_id, person.location_concept_id, person.source_person_key,
  person.source_gender_code, person.source_location_code, person.source_race_code
FROM s_person person
INNER JOIN s_observation_period period on person.person_id = period.person_id
WHERE person.year_of_birth IS NOT NULL
  AND period.observation_period_start_date IS NOT NULL
WITH READ ONLY;
```

VIEW v_src_person_strata

```
--=============================================================================
-- VIEW v_src_person_strata
--=============================================================================
CREATE OR REPLACE VIEW v_src_person_strata
 (person_id, year_of_birth, gender_concept_id, race_concept_id,
  location_concept_id, source_person_key, source_gender_code,
  source_location_code, source_race_code, observation_period_start_date,
  observation_period_end_date, age, obs_duration_days, condition_concepts,
  drug_concepts) AS
WITH drug_counts AS
 (SELECT
    person.person_id,
    COUNT (DISTINCT drug_concept_id) AS drugs
  FROM v_src_person person
  INNER JOIN s_drug_era drug_era ON person.person_id = drug_era.person_id
  WHERE drug_exposure_type = '7' -- SET PERSISTENCE WINDOW
  GROUP BY person.person_id),
```

```
cond_counts AS
 (SELECT
    person.person_id,
    COUNT (DISTINCT condition_concept_id) AS conditions
  FROM v_src_person person
  INNER JOIN s_condition_era condition_era
    ON person.person_id = condition_era.person_id
  WHERE condition_occurrence_type = '65' -- SET PERSISTENCE WINDOW
  GROUP BY person.person_id),
person_strata AS
 (SELECT /*+ NO_PARALLEL (person) */
    person.person_id, person.year_of_birth, person.gender_concept_id,
    person.race_concept_id, person.location_concept_id, person.source_person_key,
    person.source_gender_code, person.source_location_code, person.source_race_code,
    MIN(period.observation_period_start_date) AS observation_period_start_date,
    MAX(period.observation_period_end_date) AS observation_period_end_date,
    TO_NUMBER(TO_CHAR(MIN(period.observation_period_start_date),'yyyy'))
        - person.year_of_birth as age,
    MAX(period.observation_period_end_date)
        - MIN(period.observation_period_start_date) AS obs_duration_days
  FROM v_src_person person
  INNER JOIN s_observation_period period on person.person_id = period.person_id
  GROUP BY   person.person_id, person.year_of_birth, person.gender_concept_id,
    person.race_concept_id, person.location_concept_id, person.source_person_key,
    person.source_gender_code, person.source_location_code, person.source_race_code)
SELECT strata.person_id, strata.year_of_birth, strata.gender_concept_id,
  strata.race_concept_id, strata.location_concept_id, strata.source_person_key,
  strata.source_gender_code, strata.source_location_code, strata.source_race_code,
  strata.observation_period_start_date, strata.observation_period_end_date,
  strata.age, strata.obs_duration_days,
  NVL(cond.conditions,0) AS condition_concepts,
  NVL(drug.drugs,0) AS drug_concepts
FROM person_strata strata
  LEFT JOIN cond_counts cond ON strata.person_id = cond.person_id
  LEFT JOIN drug_counts drug ON strata.person_id = drug.person_id
WITH READ ONLY;
```

VIEW v_src_observation_period

```
--================================================================================
-- VIEW v_observation_period
--================================================================================
CREATE OR REPLACE VIEW v_src_observation_period
 (observation_period_id,   observation_period_start_date, observation_period_end_date,
  person_id, person_status_concept_id, rx_data_availability, dx_data_availability,
  hospital_data_availability, confidence) AS
SELECT /*+ NO_PARALLEL(obs) */ obs.observation_period_id,
  obs.observation_period_start_date,
  obs.observation_period_end_date,
  obs.person_id,
  obs.person_status_concept_id,
  obs.rx_data_availability,
  obs.dx_data_availability,
  obs.hospital_data_availability,
  obs.confidence
FROM s_observation_period obs
INNER JOIN v_src_person person ON obs.person_id = person.person_id
WITH READ ONLY;
```

VIEW v_src_condition_era1_ids

```
CREATE OR REPLACE VIEW v_src_condition_era1_ids
 (condition_occurrence_id, condition_occurrence_count) AS
SELECT /*+ NO_PARALLEL(cond) */ DISTINCT
  condition_era_id, cond.condition_occurrence_count
FROM s_condition_era cond
INNER JOIN v_src_person person ON cond.person_id = person.person_id
WHERE condition_occurrence_type = '65' -- SET PERSISTENCE WINDOW
WITH READ ONLY;
```

VIEW v_src_condition_era1

```
--=============================================================================
-- VIEW v_src_condition_era1
--=============================================================================
CREATE OR REPLACE VIEW v_src_condition_era1
  (condition_era_id, condition_era_start_date, person_id, condition_era_end_date,
   condition_concept_id, condition_occurrence_type, condition_occurrence_count) AS
SELECT /*+ NO_PARALLEL(cond) */
  condition_era_id, condition_era_start_date, cond.person_id, condition_era_end_date,
  condition_concept_id, condition_occurrence_type, condition_occurrence_count
FROM s_condition_era cond
INNER JOIN v_src_person person ON cond.person_id = person.person_id
WHERE condition_occurrence_type = '65' -- SET PERSISTENCE WINDOW
WITH READ ONLY;
```

VIEW v_src_first_conditions

```
CREATE OR REPLACE VIEW v_src_first_conditions
(person_id, condition_era_start_date, condition_concept_id) AS
SELECT DISTINCT
  cond.person_id,
  FIRST_VALUE(condition_era_start_date)
    OVER
    (PARTITION BY cond.person_id, cond.condition_concept_id
      ORDER BY cond.condition_era_start_date) AS condition_era_start_date,
  condition_concept_id
FROM v_src_condition_era1 cond
INNER JOIN s_person person  ON cond.person_id = person.person_id
GROUP BY cond.person_id, condition_era_start_date,  condition_concept_id
WITH READ ONLY;
```

VIEW v_src_all_conditions

```
CREATE OR REPLACE VIEW v_src_all_conditions
(person_id, gender_concept_id, age, condition_era_start_date, condition_concept_id) AS
  SELECT DISTINCT
    person.person_id,
    gender_concept_id,
    TO_NUMBER(TO_CHAR(condition_era_start_date,'yyyy')) - year_of_birth AS age,
    condition_era_start_date,
    condition_concept_id
  FROM v_src_condition_era1 cond
  INNER JOIN v_src_person person ON cond.person_id = person.person_id
WITH READ ONLY;
```

VIEW v_src_drug_era1

```
CREATE OR REPLACE VIEW v_src_drug_era1
  (drug_era_id, drug_era_start_date, drug_era_end_date, person_id,
   drug_exposure_type, drug_concept_id, drug_exposure_count) AS
SELECT /*+ NO_PARALLEL(drug) */
  drug_era_id, drug_era_start_date, drug_era_end_date, drug.person_id,
  drug_exposure_type, drug_concept_id, drug_exposure_count
FROM s_drug_era drug
INNER JOIN v_src_person person ON drug.person_id = person.person_id
WHERE drug_exposure_type = '7' -- SET PERSISTENCE WINDOW
WITH READ ONLY;
```

VIEW v_src_drug_era1_ids

```
CREATE OR REPLACE VIEW v_src_drug_era1_ids
 (drug_exposure_id, drug_exposure_count) AS
SELECT /*+ NO_PARALLEL(cond) */ DISTINCT
  drug_era_id, drug.drug_exposure_count
FROM s_drug_era drug
INNER JOIN v_src_person person ON drug.person_id = person.person_id
WHERE drug_exposure_type = '7' -- SET PERSISTENCE WINDOW
WITH READ ONLY;
```

VIEW v_src_first_drugs

```
CREATE OR REPLACE VIEW v_src_first_drugs
(person_id, drug_era_start_date, drug_era_end_date, drug_concept_id) AS
SELECT DISTINCT
  drug.person_id,
  FIRST_VALUE(drug_era_start_date)
    OVER
    (PARTITION BY drug.person_id, drug.drug_concept_id
      ORDER BY drug.drug_era_start_date) AS drug_era_start_date,
  FIRST_VALUE(drug_era_end_date)
    OVER
    (PARTITION BY drug.person_id, drug.drug_concept_id
      ORDER BY drug.drug_era_start_date) AS drug_era_end_date,
  drug_concept_id
FROM v_src_drug_era1 drug
INNER JOIN s_person person ON drug.person_id = person.person_id
GROUP BY drug.person_id, drug_era_start_date,  drug_era_end_date, drug_concept_id
WITH READ ONLY;
```

Execution Parameters

The Execution Parameters are parameters set by the analyst prior to executing the OSIM 2. All parameters have default values which can be used or modified. Execution Parameters are set on specific OSIM 2 Package Procedure calls. The procedure calls can be saved as scripted SQL text files.

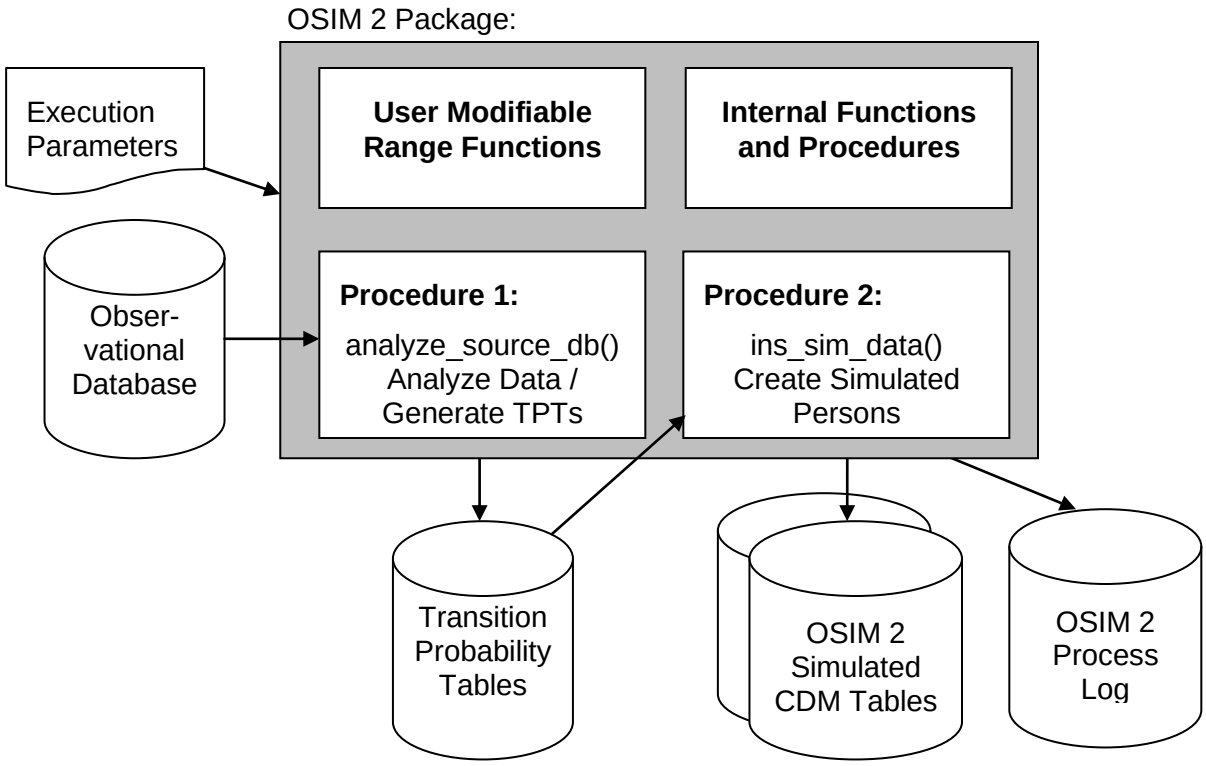


Figure 2: OSIM 2 High-Level Process Flow

Program Execution Parameters

Program Execution Parameters control how the program will be executed. They are set on the stored procedure call. A description of each parameter is contained in Table 1 below.

OSIM Execution Parameters:

Parameter Name	Procedure	Description	Format	Valid Range
person_count	ins_sim_data()	Number of simulated Persons that will be generated during the program execution	Number	1 – 500000000 Default: 5000
person_start_id	ins_sim_data()	Starting person_id value for the simulated set of persons. If set to 0 (default), then the simulation will add 1 to the current maximum person_id in the osim_person table.	Number	1 – 1000000000 Default: 0

User-Modifiable Range Functions

The user-modifiable range functions are used by both the database analysis and simulation phases of OSIM 2. They specify the ranges of various strata to be bucketed together for the transition probabilities. The user has control over the categories and related ranges associated with some of the key distributions, via a set of functions in the OSIM2 Oracle package described below. The simulation must run with the same range functions that were used to create the transition tables.

FUNCTION age_bucket

The following CASE statement can be modified to change the default age categories for analysis. The default values are shown in the example below; range values can be changed based on desired age categories.

```
CASE TRUE
  WHEN age IS NULL THEN RETURN NULL;
  WHEN age < 6 THEN RETURN 6;
  WHEN age < 14 THEN RETURN 14;
  WHEN age < 20 THEN RETURN 20;
  WHEN age < 55 THEN RETURN 55;
  WHEN age < 70 THEN RETURN 70;
  ELSE RETURN 120;
END CASE;
```

FUNCTION condition_count_bucket

The following CASE statement can be modified to change the default categories which place the count of conditions into pre-defined bins. The default values are shown in the example below; the range values can be changed based on desired age categories.

```
CASE TRUE
  WHEN condition_count <= 2 THEN RETURN 2;
  WHEN condition_count <= 7 THEN RETURN 7;
  WHEN condition_count <= 25 THEN RETURN 25;
  ELSE RETURN 2000;
END CASE;
```

FUNCTION drug_count_bucket

The following CASE statement can be modified to change the default categories which place the count of drugs into pre-defined bins. The default values are shown in the example below; the range values can be changed based on desired age categories.

```
CASE TRUE
  WHEN drug_count <= 2 THEN RETURN 2;
  WHEN drug_count <= 7 THEN RETURN 7;
  WHEN drug_count <= 25 THEN RETURN 25;
  ELSE RETURN 2000;
END CASE;
```

FUNCTION time_observed_bucket

The following CASE statement can be modified to change the default categories which place the person's observation period time remaining into pre-defined bins. The default values are shown in the example below, which are simply full semi-annual periods.

```
CASE TRUE
  WHEN days > 0 THEN RETURN FLOOR((1+days) / 182.625);
  ELSE RETURN 0;
END CASE;
```

FUNCTION round_days

The following CASE statement can be modified to change the default categories which place era duration and separating intervals (in days) into pre-defined bins. The default values are shown in the example below.

```
CASE TRUE
  WHEN days <= 75 THEN RETURN ROUND(days);
  ELSE RETURN ROUND(days - 15 + DBMS_RANDOM.VALUE * 30);
END CASE;
```

FUNCTION duration_days_bucket

The following CASE statement can be modified to change the default categories which place drug era duration intervals and time between condition days for drug simulation (in days) into pre-defined bins. The default values are shown in the example below.

```
CASE TRUE
  WHEN days <= 7 THEN RETURN 7;
  ELSE RETURN 8;
END CASE;
```

OSIM 2 Process Log

The OSIM 2 Process Log contains various messages posted during analysis and simulation.

Procedure: ins_src_db_attributes()

Table osim_log

Field Name	Format	Description / Comment
log_date	Date	Auto timestamp of log message.
stored_procedure_name	Text	OSIM 2 stored procedure posting the log message.
message	Text	Log message.

General Source Database Attributes Table

The General Source Database Attributes Table contains several key attributes of the CDM database being analyzed for simulation.

Procedure: ins_src_db_attributes()

Table osim_src_db_attributes

Field Name	Format	Description / Comment
db_min_date	Date	The minimum era or observation period start date in the CDM database being analyzed for simulation.
db_max_date	Date	The maximum era or observation period end date in the CDM database being analyzed for simulation.
persons_count	Number	The number of persons in the CDM database being analyzed for simulation.
condition_eras_count	Number	The number of 0 day persistence window condition eras in the CDM database being analyzed for simulation.
drug_eras_count	Number	The number of 30 day persistence window drug eras in the CDM database being analyzed for simulation.
condition_occurrence_type	Text	The condition_occurrence_type of the analyzed condition eras.
drug_exposure_type	Text	The drug_exposure_type of the analyzed drug eras.

Transition Probability Tables

The Transition Probability Tables contain the all the transition probabilities that are generated by Procedure 1 of the OSIM 2, using real observational data as input. These tables are produced separately from the Simulated CDM Files, and the values contained within them are used to produce all the Simulated CDM data.

There are 12 Transition Probability Tables populated by the database analysis and then subsequently used for person simulation.

The OSIM 2 package contains a procedure to purge and populate each table.

Probability Transition Tables consist of four column types:

Strata: Strata are the classification columns to calculate probability for each distinct result.

Results: Result values are the one or two columns that are the randomly drawn for during simulation based on already-known strata.

Count(n): The column n always contains the count represented by the distinct set strata and results values. The count is not used by the simulation, but is maintained for validation purposes.

Accumulated Probability: The accumulated probability column is the accumulating probability based on the count for each return value with the same strata. The final accumulated probability column for each stratum should have a value of 1.0.

The probability transition tables are commonly expressed in formula form:

Fn(Result(s) | Strata)

Table osim_gender_probability

Fn(gender_concept_id)

The osim_gender_probability table contains a row for each distinct gender_concept_id in the analyzed source CDM database. The final two columns are common to all the transition tables; n is the count for all the strata and results represented by the row (in this case there are no additional strata) and accumulated_probability, which is the accumulating probability for all results with common strata in descending n sort order.

Procedure: ins_gender_probability()

Field Name	Format	Description / Comment
gender_concept_id	Number	(Result) Distinct gender concept ID from the analyzed source CDM database.
n	Number	Number of persons in the source CDM database with the specified gender concept ID.
accumulated_probability	Float	Accumulating probability of gender concept ID.

Table osim_age_at_obs_probability

Fn(age_at_obs | gender_concept_id)

The osim_age_at_obs_probability table contains a row for each distinct gender and person age at observation start in the analyzed source CDM database. The final two columns are common to all the transition tables; n is the count for all the strata and results represented by the row (in this case gender_concept_id is the only additional stratum) and accumulated_probability, which is the accumulating probability for all results with common strata in descending n sort order.

Procedure: ins_age_at_obs_probability()

Field Name	Format	Description / Comment
gender_concept_id	Number	(Stratum) Distinct gender concept ID from the analyzed source CDM database.
age_at_obs	Number	(Result) Age in whole years of person at beginning of observation period.
n	Number	Number of persons in the source CDM database with the specified gender concept ID and age at observation.
accumulated_probability	Float	Accumulating probability of age at observation for each gender concept ID.

Table osim_cond_count_probability

Fn(cond_concept_count | gender_concept_id, age_at_obs)

The osim_cond_count_probability table contains a row for each distinct gender, age at observation start, and number of distinct condition concepts in the analyzed source CDM database. The final two columns are common to all the transition tables; n is the count for all the strata represented by the row (in this case gender_concept_id and age_at_obs) and accumulated_probability, which is the accumulating probability for all results with common strata in descending n sort order.

Procedure: ins_cond_count_probability()

Field Name	Format	Description / Comment
gender_concept_id	Number	(Stratum) Distinct gender concept ID from the analyzed source CDM database.
age_at_obs	Number	(Stratum) Age in whole years of person at beginning of observation period.
cond_concept_count	Number	(Result) Count of distinct condition concepts.
n	Number	Number of persons in the source CDM database with the specified gender concept ID age at observation, and number of distinct conditions.
accumulated_probability	Float	Accumulating probability for the strata.

Table osim_time_obs_probability

Fn(time_observed | gender_concept_id, age_at_obs, cond_count_bucket)

The osim_time_obs_probability table contains a row for each distinct gender_concept_id, age at observation start, bucketed number of distinct condition concepts, and full semi-years observed in the analyzed source CDM database. The final two columns are common to all the transition tables; n is the count for all the strata represented by the row (in this case gender_concept_id, age_at_obs, and cond_count_bucket) and accumulated_probability, which is the accumulating probability for all results with common strata in descending n sort order.

Procedure: ins_time_obs_probability()

Field Name	Format	Description / Comment
gender_concept_id	Number	(Stratum) Distinct gender concept ID from the analyzed source CDM database.
age_at_obs	Number	(Stratum) Age in whole years of person at beginning of observation period.
cond_count_bucket	Number	(Stratum) Bucketed count of distinct condition concepts.
time_observed	Number	(Result) Person observation duration in full semi-years. A value of zero represents less than 6 months observed.
n	Number	Number of persons in the source CDM database with the specified gender concept ID age at observation, and number of distinct conditions.
accumulated_probability	Float	Accumulating probability for the strata.

Table osim_first_cond_probability

Fn(condiiton2_concept_id, delta_days | gender_concept_id, age_range, cond_count_bucket, time_remaining, condition1_concept_id)

The osim_time_obs_probability table contains the probability for transitioning from one prior initial occurrence condition concept to the next initial occurrence. Only the very first condition eras for a condition concept are analyzed for each person. The transitions are stratified by gender, age bucket (at time of the prior condition era), bucketed number of distinct conditions, and time remaining in the observation period (in full semi-years). The result values are the next condition concept and the number of days until it occurs from the prior condition (see the round_days bucket function). A pseudo condition concept of -1 is used to represent a person's observation start and prime the initial condition concept. The age range is bucketed except for the initial -1 condition concept, where the actual age in whole years is used.

Procedure: ins_first_cond_probability()

Field Name	Format	Description / Comment
gender_concept_id	Number	(Stratum) Distinct gender concept ID from the analyzed source CDM database.
age_range	Number	(Stratum) Age in whole years of person at beginning of observation period for initial condition (condition1_concept_id = -1); bucketed age at beginning of condition era for actual condition concepts.
cond_count_bucket	Number	(Stratum) Bucketed count of distinct condition concepts.
time_remaining	Number	(Stratum) Person observation duration in full semi-years. A value of zero represents less than 6 months observed.
condition1_concept_id	Number	(Stratum) Prior condition concept, -1 for observation period start.
condition2_concept_id	Number	(Result) Next condition concept.
delta_days	Number	(Result) Days from prior to next condition (see the round_days bucket function).
n	Number	Number of occurrences in the source CDM database with the strata and results.
accumulated_probability	Float	Accumulating probability for the strata.

Table osim_cond_era_count_prob

Fn(cond_era_count | condition_concept_id, cond_count_bucket, time_remaining)

The osim_cond_era_count_prob table contains the probabilities for the total number of condition eras a person has for the given condition. The probabilities are stratified by condition concept, total number of distinct conditions, and full semi-years remaining on the data of the initial condition occurrence.

Procedure: ins_cond_era_count_prob()

Field Name	Format	Description / Comment
condition_concept_id	Number	(Stratum) Distinct condition concept
cond_count_bucket	Number	(Stratum) Bucketed count of distinct condition concepts.
time_remaining	Number	(Stratum) Person observation duration in full semi-years. A value of zero represents less than 6 months observed.
cond_era_count	Number	(Result) Total number of condition eras for the condition concept.
n	Number	Number of occurrences in the source CDM database with the strata and results.
accumulated_probability	Float	Accumulating probability for the strata.

Table osim_cond_reoccur_probability

Fn(delta_days | condition_concept_id, age_range, time_remaining)

The osim_cond_reoccur_probability table contains the probabilities for the days between subsequent reoccurrences of the same condition for the given condition. The probabilities are stratified by condition concept, total age range, and full semi-years remaining from the previous condition occurrence.

Procedure: ins_cond_days_before_prob()

Field Name	Format	Description / Comment
condition_concept_id	Number	(Stratum) Distinct condition concept.
age_range	Number	(Stratum) Bucketed age at beginning of condition era.
time_remaining	Number	(Stratum) Person observation duration remaining in full sem- years at beginning of condition era. A value of zero represents less than 6 months observed.
delta_days	Number	(Result) Days from prior to next condition (see the round_days bucket function).
n	Number	Number of occurrences in the source CDM database with the strata and results.
accumulated_probability	Float	Accumulating probability for the strata.

Table osim_drug_count_prob

Fn(drug_count | gender_concept_id, age_bucket, condition_count_bucket)

The osim_drug_count_prob table contains the probabilities for the total number of drug concepts a person has. The probabilities are stratified by gender, bucketed age range at beginning of observation period, and total number of distinct conditions.

Procedure: ins_drug_count_prob()

Field Name	Format	Description / Comment
gender_concept_id	Number	(Stratum) Distinct gender concept ID from the analyzed source CDM database.
age_bucket	Number	(Stratum) Bucketed age at beginning of observation period.
condition_count_bucket	Number	(Stratum) Bucketed count of distinct condition concepts.
drug_count	Number	(Result) Number of distinct drug concepts.
n	Number	Number of occurrences in the source CDM database with the strata and results.
accumulated_probability	Float	Accumulating probability for the strata.

Table osim_cond_drug_count_prob

Fn(drug_count | condition_concept_id, interval_bucket, age_range, drug_count_bucket, cond_count_bucket)

The osim_cond_ _drug__count_prob table contains the probabilities for the total number of drug concept draws a person should have for a given condition. This is the probability of the number of drugs in a gap between days with conditions. The probabilities are stratified by condition, bucketed days between conditions, bucketed age range on condition date, total number of distinct drugs, and total number of distinct conditions.

Procedure: ins_cond_drug_count_prob()

Field Name	Format	Description / Comment
condition_concept_id	Number	(Stratum) Distinct condition concept.
interval_bucket	Number	(Stratum) Bucketed number of days between subsequent days with conditions (see interval_bucket function).
age_bucket	Number	(Stratum) Bucketed age at beginning of condition era.
condition_count_bucket	Number	(Stratum) Bucketed count of distinct condition concepts.
drug_count_bucket	Number	(Stratum) Bucketed count of distinct drug concepts.
drug_count	Number	(Result) Number of drug concept draws to perform for this condition.
n	Number	Number of occurrences in the source CDM database with the strata and results.
accumulated_probability	Float	Accumulating probability for the strata.

Table osim_cond_first_drug_prob

Fn(drug_concept_id, delta_days | condition_concept_id, interval_bucket, gender_concept_id, age_range, condition_count_bucket, drug_count_bucket, day_cond_count)

The osim_cond_first_drug_prob table contains the counts and probabilities of the transition from a condition era to the drug concept of a first occurrence drug era. The duration in days until the drug era occurrence is also returned. The transition is further stratified by the duration of the gap to the next day with condition(s), gender, age range, distinct number of conditions range, and the number of conditions on the day preceding the gap.

During analysis, the probability for every drug in the condition gap is divided evenly among the conditions on the date preceding the gap.

Procedure: ins_cond_first_drug_prob()

Field Name	Format	Description / Comment
condition_concept_id	Number	(Stratum) Distinct condition concept.
interval_bucket	Number	(Stratum) Bucketed number of days between subsequent days with conditions (see duration_days_bucket function).
gender_concept_id	Number	(Stratum) Distinct gender concept ID from the analyzed source CDM database.
age_bucket	Number	(Stratum) Bucketed age at beginning of condition era.
condition_count_bucket	Number	(Stratum) Bucketed count of distinct condition concepts.
drug_count_bucket	Number	(Stratum) Bucketed count of distinct drug concepts.
day_cond_count	Number	(Stratum) Number of distinct conditions on the day of the condition preceding the gap.
drug_concept_id	Number	(Result) Drug concept occurring in the gap. A value of -1 is used to designate the probability of no drug.
delta_days	Number	(Result) Days from condition date until the drug era (see the round_days bucket function).
n	Number	Number of occurrences in the source CDM database with the strata and results.
accumulated_probability	Float	Accumulating probability for the strata.

Table osim_drug_era_count_prob

Fn(drug_era_count, total_exposure | drug_concept_id, drug_count_bucket, condition_count_bucket, age_range, time_remaining)

The osim_drug_era_count_prob table contains the probabilities for the total number of drug eras and total exposure duration of all the eras a person has given the drug concept, bucketed distinct number of drugs and conditions, bucketed age at first exposure, and full semi-years remaining at first exposure.

Procedure: ins_drug_era_count_prob()

Field Name	Format	Description / Comment
drug_concept_id	Number	(Stratum) Distinct drug concept.
drug_count_bucket	Number	(Stratum) Bucketed count of distinct drug concepts.
condition_count_bucket	Number	(Stratum) Bucketed count of distinct condition concepts.
age_range	Number	(Stratum) Bucketed age at beginning of condition era.
time_remaining	Number	(Stratum) Person observation duration remaining in full semi-years at beginning of first drug era. A value of zero represents less than 6 months observed.
drug_era_count	Number	(Result) Count of total drug eras for the drug concept.
total_exposure	Number	(Result) Bucketed total number of exposure days for drug concept (see round_days function).
n	Number	Number of occurrences in the source CDM database with the strata and results.
accumulated_probability	Float	Accumulating probability for the strata.

Table osim_drug_duration_probability

Fn(total_duration | drug_concept_id, time_remaining, drug_era_count, total_exposure, total_duration)

The osim_drug_duration_probability table contains the probabilities of total durations (end date of last exposure – start of first exposure) of a drug for the total number of drug eras, rounded total drug exposure, and time remaining after initial exposure.

Procedure: ins_drug_duration_probability()

Field Name	Format	Description / Comment
drug_concept_id	Number	(Stratum) Distinct drug concept.
time_remaining	Number	(Stratum) Person observation duration remaining in full semi-years at beginning of first drug era. A value of zero represents less than 6 months observed.
drug_era_count	Number	(Stratum) Count of total drug eras for the drug concept.
total_exposure	Number	(Stratum) Bucketed total number of exposure days for drug concept (see round_days function).
total_duration	Number	(Result) Bucketed total duration for exposure from start of initial exposure to end of final exposure for drug concept (see round_days function).
n	Number	Number of occurrences in the source CDM database with the strata and results.
accumulated_probability	Float	Accumulating probability for the strata.

Simulated CDM Tables

The simulation phase OSIM 2 populates CDM patient data tables containing the simulated Persons and related data. These are "analysis-ready" files, produced in the OMOP Common Data Model format, that are created using the attributes found in the Simulation Attribute Tables. There are four Simulated Person files produced, which are described in detail below. The current version analyzes and simulates persons, observation periods, 0-day persistence condition eras, and 30-day persistence drug eras.

During simulation, temporary drug and condition era tables are populated for each person, and the eras are then copied to the final output tables at the completion of each person's simulated eras.

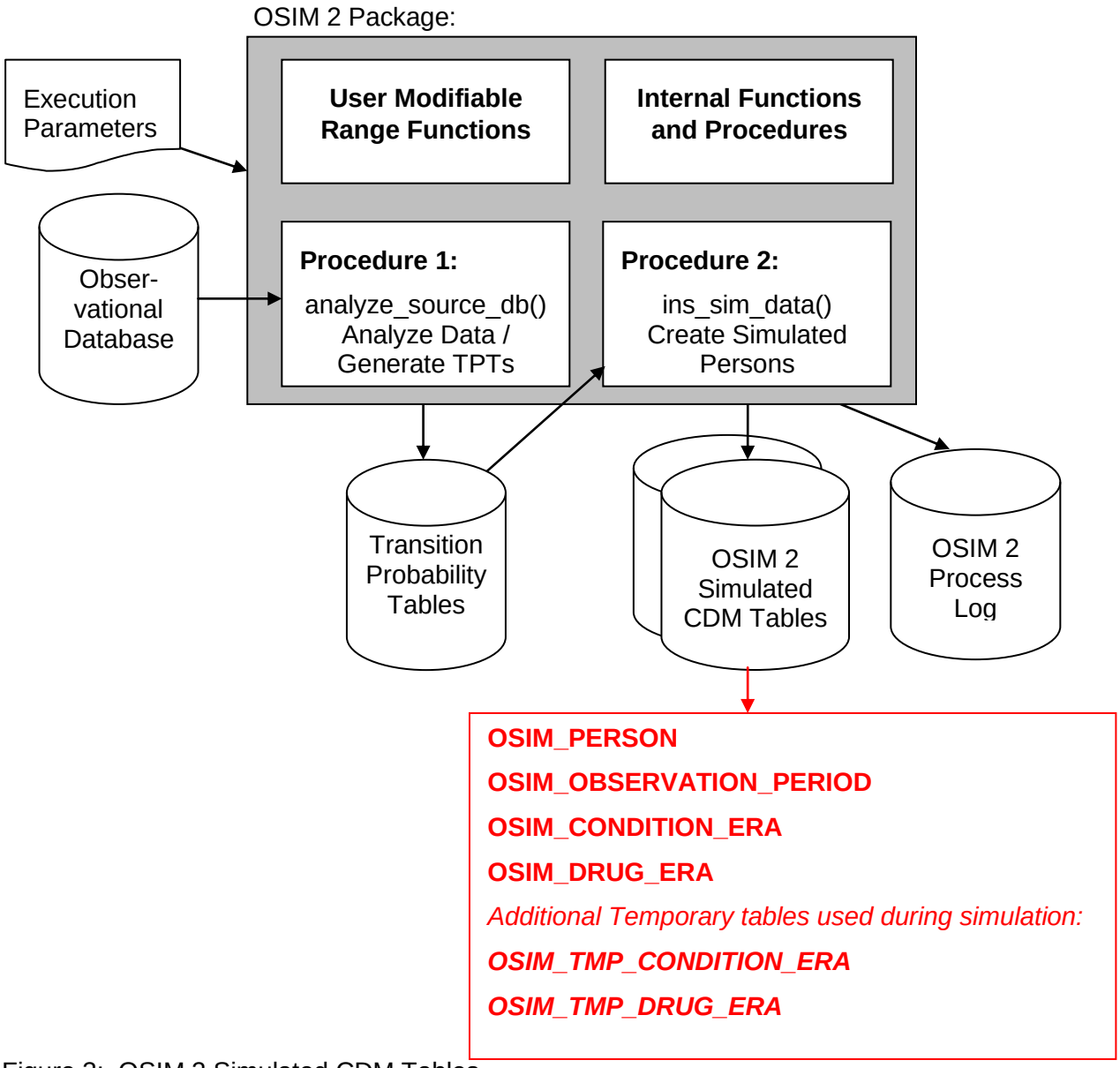


Figure 3: OSIM 2 Simulated CDM Tables

OSIM_PERSON

The Simulated Person Table **OSIM_PERSON** contains the demographic characteristics of the simulated persons generated during this execution of the OSIM. The number of persons contained within this file is controlled by the input parameter: *person_count.* The contents of this file are shown in the following table.

Field Name	Format	Description / Comment
PERSON_ID	Number	Sequential generated identifier that uniquely identifies each person.
YEAR_OF_BIRTH	Number	Year of birth is simulated from the observed probabilities based on the person gender.
GENDER_CONCEPT_ID	Text	Gender Concept ID assigned based on the observed probability in the source CDM database.
RACE_CONCEPT_ID	Text	*Null*
LOCATION_CONCEPT_ID	Text	*Null*
SOURCE_PERSON_KEY	Text	*Null*
SOURCE_GENDER_CODE	Text	*Null*
SOURCE_LOCATION_CODE	Text	*Null*
SOURCE_RACE_CODE	Text	*Null*

OSIM_OBSERVATION_PERIOD

The Simulated Observation Period Table **OSIM_OBSERVATION_PERIOD** contains the range of calendar time when a person can have recorded drug exposures and condition occurrences, which may be defined by a person's enrollment in the insurance plan. In the simulated data, this period is used to censor the person's data.

Field Name	Format	Description / Comment
OBSERVATION_PERIOD_ID	Number	Sequential generated identifier that uniquely identifies each Observation Period
OBSERVATION_START_DATE	Date	This is the start date of the observation period. It is assigned by selecting a random value from a uniform distribution with a lower bounds of the observed **MinDatabaseDate** and an upper bound of the observed **MaxDatabaseDate** – the person's simulated observation duration.
OBSERVATION_END_DATE	Date	This is the end date of the observation period. It is the simulated OBSERVATION_START_DATE + the simulated person observation duration.
PERSON_ID	Number	Sequential generated identifier from the OSIM_PERSON file that uniquely identifies each person.
PERSON_STATUS_CONCEPT_ ID	Number	*Null*
RX_DATA_AVAILABILITY	Text	This indicator will be set to "Y" for all OSIM persons
DX_DATA_AVAILABILITY	Text	This indicator will be set to "Y" for all OSIM persons
HOSPITAL_DATA_AVAILABILITY	Text	This indicator will be set to "N" for all OSIM persons
CONFIDENCE	Text	*Null*

OSIM_CONDITION_ERA

The Simulated Condition Era Table **OSIM_CONDITION_ERA** contains the simulated 0-day persistence window condition eras recorded for each person.

Field Name	Format	Description / Comment
CONDITION_ERA_ID	Number	Sequential generated identifier to uniquely identify this person condition.
CONDITION_ERA_START_DATE	Date	This is the date that this condition era was recorded in the patient's record.
PERSON_ID	Number	Generated identifier that uniquely identifies each person.
CONFIDENCE	Text	*Null*
CONDITION_END_DATE	Date	Will be same as CONDITION_ERA_START_DATE.
CONDITION_CONCEPT_ID	Number	CONDITION_CONCEPT_ID from the observed condition probability analysis.
CONDITION_OCCURRENCE_TYPE	Text	'64' to designate 0-day persistence window eras, '65' for 30-day windows.
CONDITION_OCCURRENCE_COUNT	Text	1

OSIM_DRUG_ERA

The Simulated Drug Exposure Table **OSIM_DRUG_ERA** contains the drug 30-day persistence window drug eras simulated for each person.

Field Name	Format	Description / Comment
DRUG_ERA_ID	Number	System generated identifier to uniquely identify this person drug exposure.
DRUG_ERA_START_DATE	Date	Start date of the drug era. It is simulated by randomly distributing all drug eras for a single drug over the person's total drug duration period.
DRUG_ERA_END_DATE	Date	End date of the drug era. It is simulated by randomly distributing all drug eras for a single drug over the person's total drug duration period.
PERSON_ID	Number	PERSON_ID from the OSIM_PERSONS file.
DRUG_EXPOSURE_TYPE	Text	'6' to designate 0-day persistence window eras, '7' for 30-day windows.
DRUG_CONCEPT_ID	Number	DRUG_CONCEPT_ID from the observed drug probability analysis.
DRUG_EXPOSURE_COUNT	Number	*Currently not simulated. Set to 1.*

Outcomes Table (Optional)

OSIM 2 has an optional outcome phase that can adjust known drug / condition outcome prevalence. The optional method, **ins_outcomes()**, requires the manual population of the OSIM_DRUG_OUTCOME table. The method will analyze the simulated prevalence, and add or remove condition eras as specified by the additional **OSIM_DRUG_OUTCOME** columns.

TABLE OSIM_DRUG_OUTCOME

Field Name	Format	Description / Comment
RISK_OR_BENEFIT	Text	'risk' or 'benefit' outcome type.
DRUG_CONCEPT_ID	Number	Outcome Drug concept ID.
CONDITION_CONCEPT_ID	Number	Outcome condition concept ID.
RELATIVE_RISK	Float	Percentage of simulated persons with the drug expected to have outcome.
OUTCOME_RISK_TYPE	Text	• **first exposure** – outcomes are only added to first drug exposures • **any exposure** – outcomes may be added to any drug exposure • **insidious** – outcomes are randomly added on a date during any exposure • **accumulative** – outcomes are added during any drug exposure, with accumulating probability over time
OUTCOME_ONSET_DAYS_MIN	Number	Minimum days from drug exposure start date for outcome to occur; this column can be set for any outcome_risk_type.
OUTCOME_ONSET_DAYS_MIN	Number	Maximum days from drug exposure start date for outcome to occur; this column only applies to first and any exposure types; value must be >= outcome_onset_days_min.

Revision History

Revision Date	Person Resp.	Reason for Revision
29 Oct 2010	R. Murray	Draft Version
08 Dec 2010	R. Murray	Additional parameters for parallelization, and optional outcomes description
30 Dec 2010	R. Murray	Final Version
03 Jan 2011	R. Murray	Updated Corporate Logo
20 Jan 2011	R. Murray	Minor Corrections