integrated intelligence

Design and Validation of a Data Simulation Model for Longitudinal Healthcare Data Method Validation

Presenter: Rich Murray

AMIA Presentation

October 26, 2011

UBC®
United BioSource Corporation
Evidence Matters®

FNIH
Foundation for the National Institutes of Health

Observational Medical Outcomes Partnership

"The OMOP partnership has conducted a two-year initiative to research methods that are feasible and useful to analyze existing healthcare databases to identify and evaluate safety and benefit issues of drugs already on the market"

source: omop.fnih.org

- Conducted methodological research to empirically evaluate the performance of alternative methods on their ability to identify true drug safety issues
- Developed tools and capabilities for transforming, characterizing, and analyzing disparate data sources
- Established a shared resource so that the broader research community can collaboratively advance the science

FNIH Foundation for the National Institutes of Health
UBC United BioSource Corporation Evidence Matters®

Context for Simulated Data

Observational Data (Large claims and EHR databases)	Simulated Data
Data is “noisy” (confounding)	Model both adverse drug reactions and confounding
Data capture process provides further distortion	Simulate data capture process
Limited gold standards for objective measurement	Known characteristics provide “truth” for measurement
Access limited & expensive with end user restrictions	Data freely & widely available & there are no privacy issues
Disparate data formats / coding schemes	Use of Common Data Model mitigates format issues

Simulated data, with known properties and characteristics, can facilitate systematic evaluation & comparison among methods providing objective gold standard

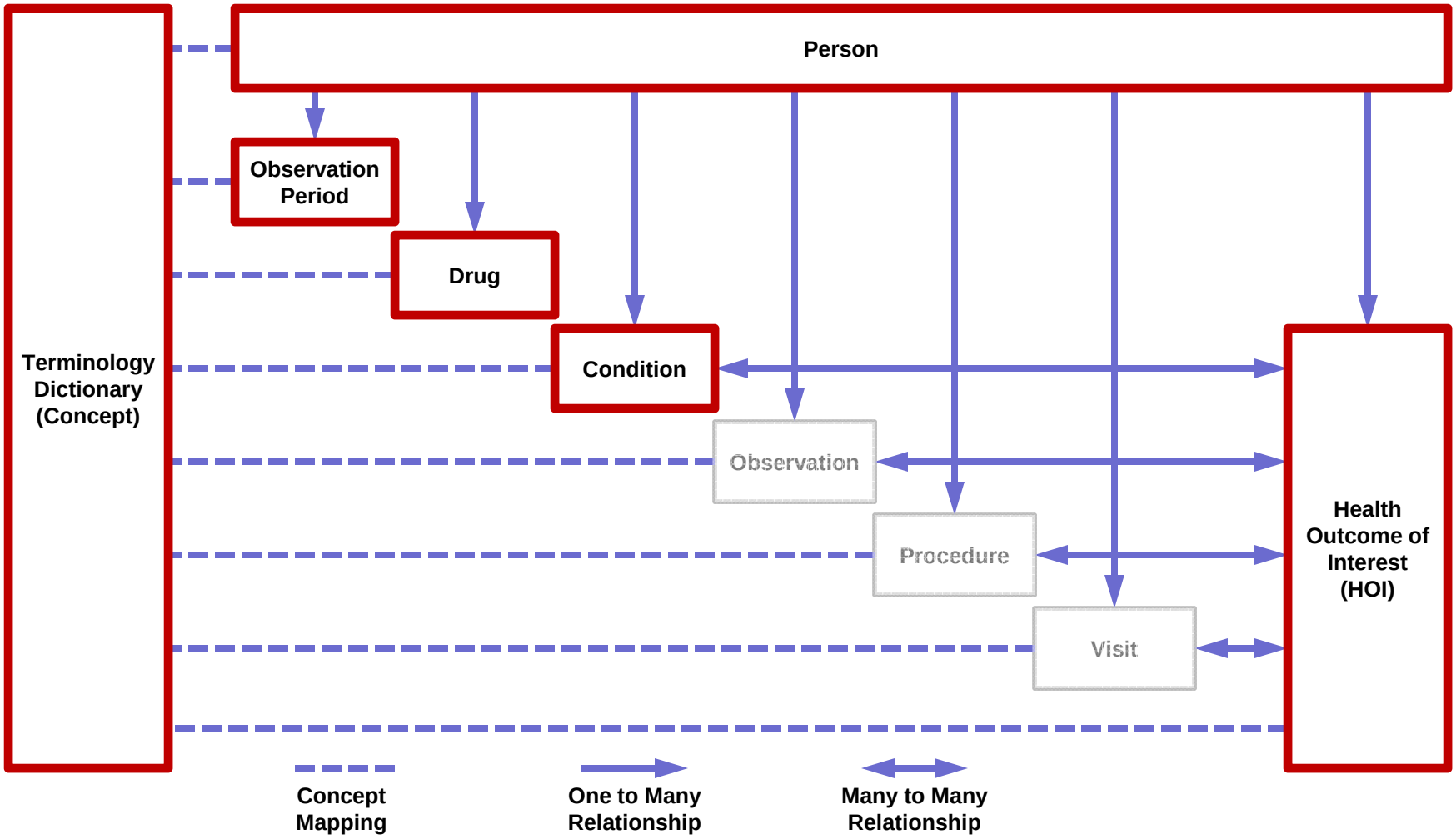
FNIH Foundation for the National Institutes of Health
UBC United BioSource Corporation
Evidence Matters®

Goals of the Simulation Model

- **Construct large-scale, high-fidelity, simulated database to complement "real" data experiments for methodological research**
  - Millions of persons
  - Thousands of distinct drugs and conditions
- **Mimic characteristics of real observational data**
  - Similar demographics
  - Similar condition and drug prevalence
  - Similar condition and drug timing
  - Similar recording anomaly characteristics
- **Signal injection component to provide known "Ground Truth" for developing outcome detection methods**
- **An open source software application that uses the OMOP Common Data Model (CDM)**

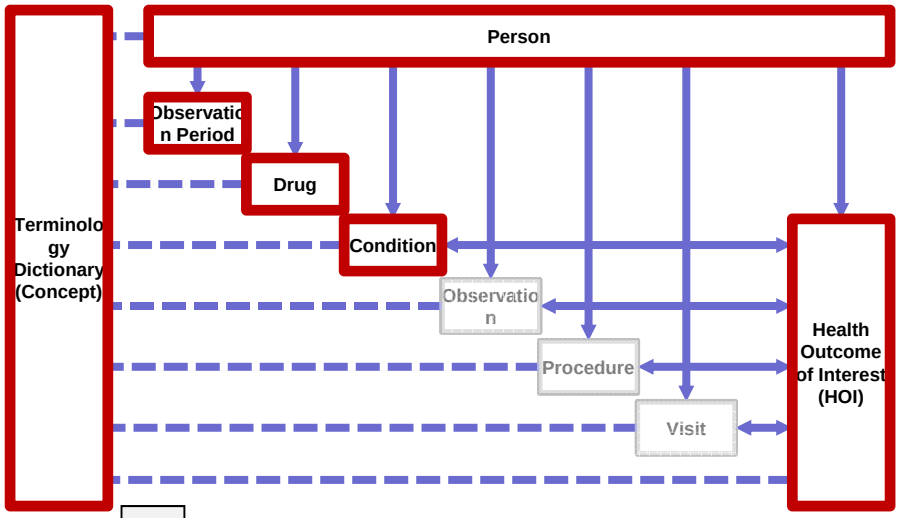
FNIH Foundation for the National Institutes of Health
UBC United BioSource Corporation
Evidence Matters®

OMOP Common Data Model Domains within OSIM 2 Simulation



Characteristics and transition probabilities are observed in any CDM database

OMOP Common Data Model Domains within OSIM 2 Simulation



Analyze any CDM database

Transition probability matrices

Simulate Data

Drug Characteristics

- Drug Counts
- Number and Duration of Drug Exposures
- Number of First Incident Drugs following Condition
- Condition to First Incident Drug Transitions and Time Between

FNIH Foundation for the National Institutes of Health

UBC United BioSource Corporation Evidence Matters®

Method Evaluation

Feasibility of the method to apply across disparate real-world healthcare databases

- Executed analysis phase of the model on five real world databases in the OMOP lab to establish execution time and space requirements

Consistency of the simulated data to reflect the key characteristics of real data

- Developed a standard “dashboard” comparing key summary statistics between 5the source database and the simulated data
  - Population characteristics
  - Condition and Drug Counts per person
  - Condition and Drug Prevalence
  - Condition and Drug Co-occurrence

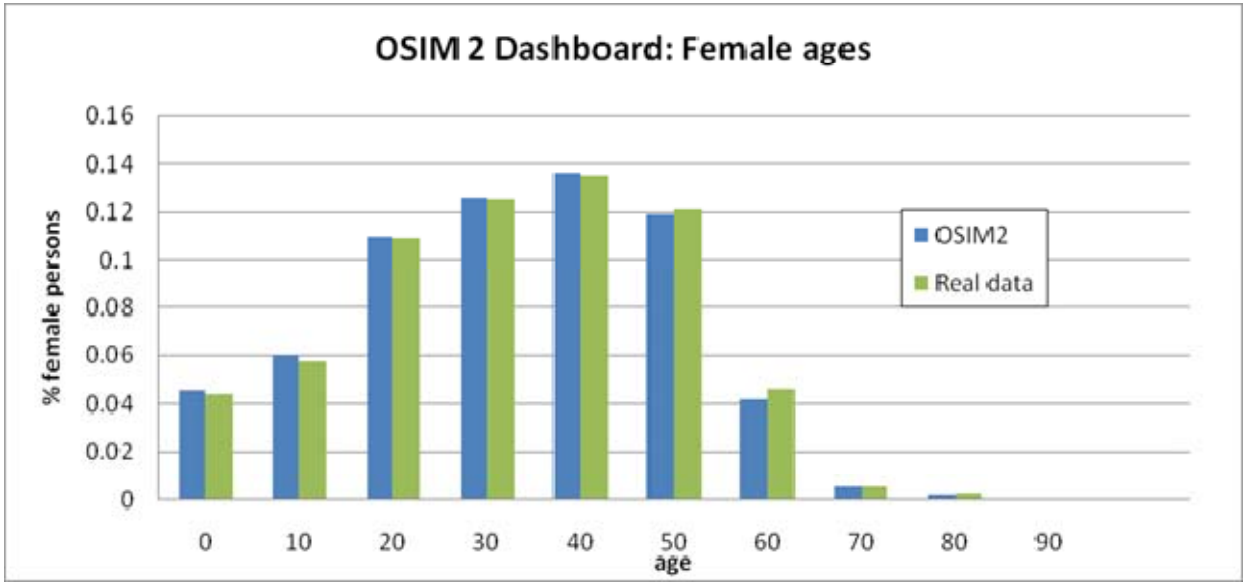
FNIH Foundation for the National Institutes of Health
UBC United BioSource Corporation
Evidence Matters®

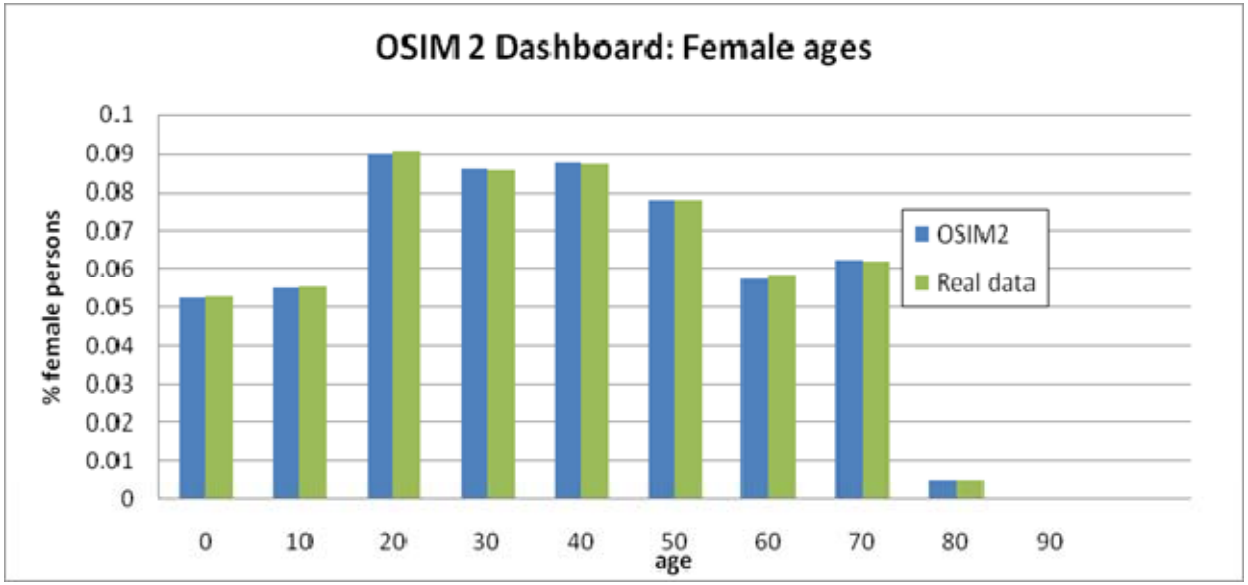
Physical Characteristics of Analysis and Transitions

	MSLR	MDCR	MDCD	CCAE	GE
Persons (millions)	1.5	4.4	11.1	58	11
Approximate Size of Source Tables (gigabytes)	2.7	7.9	20	104	19.8
Analysis Time (hours)	5.25	26.5	49.5	148	15.5
Size of Attribute Tables (gigabytes)	3.9	4.3	5.5	14.8	5.5

MSLR -- Thomson Reuters MarketScan® Lab Database with inpatient and outpatient medical claims as well as pharmacy dispensing records
MDCR -- MarketScan Medicare Supplemental and Coordination of Benefits Database contains administrative claims for 4.4 million retirees with Medicare supplemental insurance paid for by employers
MDCD -- MarketScan Medicaid Multi-State Database (MDCD) contains administrative claims data for Medicaid enrollees from multiple states
CCAE -- MarketScan Commercial Claims and Encounters (CCAE) captures private de-identified administrative claims from inpatient and outpatient visits and pharmacy claims of multiple insurance plans
GE -- contains patient-level data of 11 million persons captured from a consortium of providers using the GE Centricity system in their outpatient and specialty practices

Simulation Results – Population Characteristics





- Claims Data Simulation (MSLR)

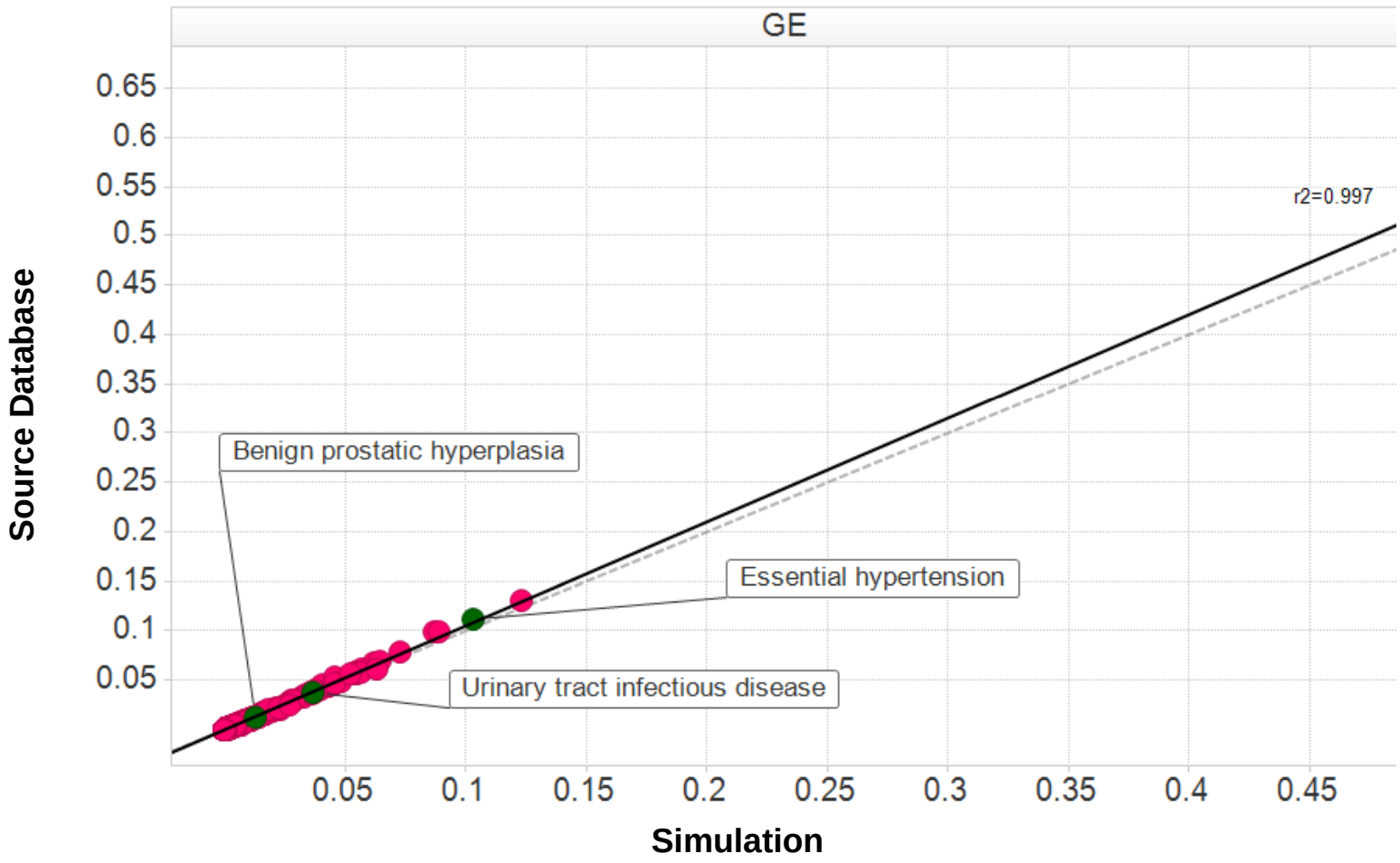
Female Ages

- EMR Data Simulation (GE)

FNIH Foundation for the National Institutes of Health

UBC United BioSource Corporation *Evidence Matters®*

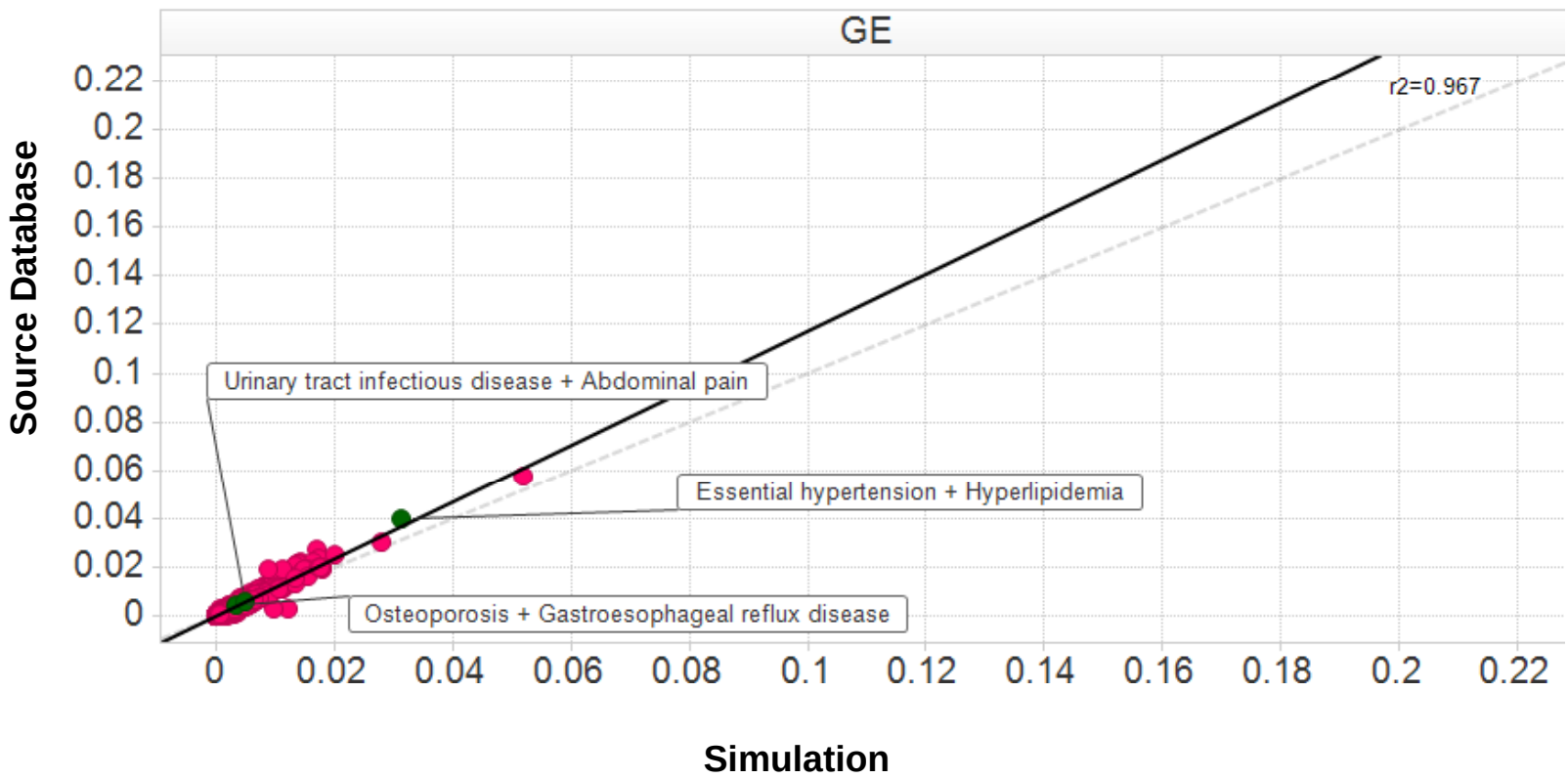
Condition Prevalence Dashboard (GE)



Comparing condition prevalence for all conditions (GE)

FNIH Foundation for the National Institutes of Health
UBC United BioSource Corporation
Evidence Matters®

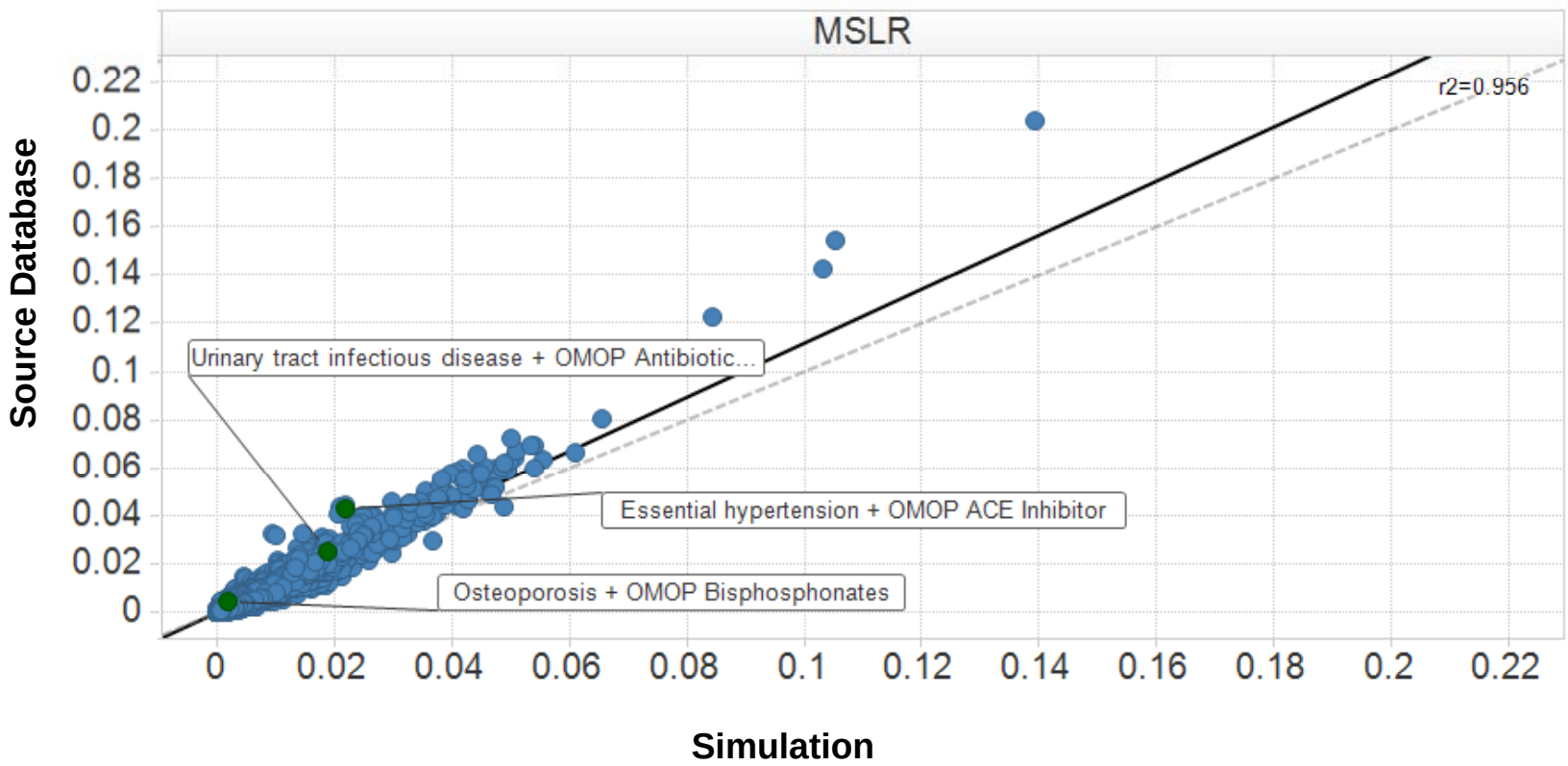
Condition Co-occurrence Dashboard (GE)



Comparing condition co-occurrence for 200 most prevalent conditions (GE)

FNIH Foundation for the National Institutes of Health
UBC United BioSource Corporation
Evidence Matters®

Condition / Drug Co-occurrence Dashboard (MSLR)



Comparing condition / drug co-occurrence for 200 most prevalent conditions and drugs (MSLR)

FNIH Foundation for the National Institutes of Health
UBC United BioSource Corporation
Evidence Matters®

Current Limitations

- First-order Markov model doesn't reflect full complexities in data
- Currently models one observation period per person
- Does not model clustered nature of encounter-based data
- Does not model multi-ingredient drugs well (the CDM drug eras are at an ingredient granularity)
- Does not simulate procedures or observations
- Does not model emergence or removal of drugs over time

FNIH Foundation for the National Institutes of Health
UBC United BioSource Corporation
Evidence Matters®

Conclusions

Even with the current limitations:

- OSIM 2 effectively create a performance benchmark that can facilitate methodological research
- By directly modeling characteristics of the healthcare data, recording anomalies and errors are automatically simulated with similar frequencies
- The resulting simulated data is ideally suited for testing and evaluating methods designed to run against real healthcare datasets, but has other potential uses anywhere inexpensive EHR data is needed with no privacy issues (software development and testing, training, etc.)

FNIH Foundation for the National Institutes of Health
UBC United BioSource Corporation
Evidence Matters®

Find out more

OSIM2 represents an informatics solution to enable the systematic exploration of the performance of alternative data analytic methods in their ability to identify true effects and discern from false positive findings

http://omop.fnih.org/OSIM2

FNIH Foundation for the National Institutes of Health
UBC United BioSource Corporation
Evidence Matters®