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

Richard E Murray¹, Patrick B Ryan, PhD²,³, Stephanie J Reisinger¹

¹United BioSource Corporation, Harrisburg, PA, United States
²Johnson & Johnson Pharmaceutical Research and Development, Titusville, NJ, United States
³Observational Medical Outcomes Partnership; Foundation for the National Institutes of Health, Bethesda, MD, United States

Evaluating performance characteristics of analytic methods developed to identify treatment effects in longitudinal healthcare data has been hindered by lack of an objective benchmark to measure performance. Relationships between drugs and subsequent treatment effects are not precisely quantified in real-world data, and simulated data offer potential to augment method development by providing data with known, measurable characteristics. However, the use of simulated data has been limited due to its inability to adequately reflect the complexities inherent in real-world databases that are necessary for effective method development. The goal of this study was to develop and evaluate a model for simulating longitudinal healthcare data that adequately captures these complexities. An empiric design was chosen that utilizes the characteristics of a real healthcare database as simulation input. This model demonstrates the potential for simulated data with known characteristics to adequately reflect complex relationships among diseases and treatments as recorded in healthcare databases.

BACKGROUND

Analysis of longitudinal healthcare data, such as electronic health records and administrative claims, provides opportunities to better understand the effects of medical interventions. Two applications for this type of research, active drug safety surveillance and comparative effectiveness research, have gained recent focus due to Congressional mandates including the Food and Drug Administrative Amendments Act of 2007,¹ and the, American Recovery and Reinvestment Act of 2009 [2] which lead to the creation of the Patient Centered Outcomes Research Institute. Both mandates require better identification of drug-related treatment effects and require increased evidence generation of alternative treatments to facilitate better and more cost-effective medical decision making.

To address the need for the generation of more and better evidence related to the effects of drug treatments, further methodological research is needed to develop analytical methods that can be systematically applied to longitudinal data to provide accurate measures of those effects. Such methodological research typically requires some benchmark against which to measure performance. In this context, a desired performance benchmark is a well characterized database with known, measurable relationships between drug exposures and subsequent treatment effects. Unfortunately, real-world healthcare data sources vary significantly in how clinical observations are recorded depending on the data capture process and the population represented. This variability makes it difficult to determine if real clinical effects are truly observable in these sources, and whether the observed effect estimates should be expected to be consistent with the known effect. In addition, a significant limitation to the use of real healthcare data for methodological research is that access to the data is often limited due to cost, patient privacy and confidentiality issues.

By addressing some of the issues inherent with the use of real-world healthcare data, simulated data offers the potential to augment methodological research for measurement of treatment effects. However, a significant weakness of simulated data has been an inability to capture the complex relationships among the disease and treatment information recorded in healthcare databases, a consequence of intricacies related to disease progression, physician / patient interactions, as well as the actual recording of this information into an electronic health record. These complexities introduce confounding factors into the data that may bias the identification and measurement of drug treatment effects; therefore it is important that any methods developed are able to identify and control for these factors.
Simulation models previously described in the literature have focused on specific diseases and biological disease progression such as influenza,[3] metachronous colorectal cancer,[4] and recurrent infections.[5] For the purposes of systematic identification of drug treatment effects that span multiple disease areas, these models are insufficient beyond their disease area of focus. In addition, disease focused simulations do not address how disease information is actually recorded in healthcare databases, which is an important confounding factor of healthcare data that must be accounted for when identifying potential treatment effects. Other models have taken the approach of “injecting” drug treatment effects with measurable characteristics into real-world data.[6, 7] While this approach provides signals that can be objectively measured, the background database is poorly characterized making it more difficult to identify and account for factors that may confound the identification of real drug treatment effects.

To facilitate method development testing, the Observational Medical Outcomes Partnership (OMOP)[8] carried out the development of a novel simulation program, utilizing an empiric approach, in an attempt to model complexities found in real-world healthcare data. Unlike previous models, this approach does not model biological processes, but instead models the characteristics of the data itself. The initial version of the Observational Medical Dataset Simulator (OSIM) utilized a series of user-defined probability tables to model first-order univariate effects such as population demographics and prevalence distributions of drugs and conditions. The initial program also addressed some of the healthcare database recording anomalies by assigning specificity and sensitivity probabilities to every drug and condition. However, OSIM insufficiently modeled derived data characteristics such as relationships between and among conditions and drugs, a limitation which became apparent during method validation. OSIM2, described in this paper, extends the OSIM research and addresses these limitations. The OSIM2 model is based on empiric data characteristics that are extracted from real-world healthcare databases as part of the simulation. Using these empiric characteristics, OSIM2 creates simulated data containing persons with longitudinal records of diseases and treatments, which are constructed using the extracted real-world characteristics.

METHOD DEVELOPMENT

Overview of the Model
OSIM2 is an empirical simulation model of longitudinal patient data, including diseases and treatments. The simulation creates individual medical record data based on a detailed analysis of a real observational database. To model drug adverse events, the user can optionally specify a list of drug / outcome associations with specific characteristics that are injected into the data. OSIM2 can create a simulated dataset based on the characteristics of any healthcare database that conforms to the OMOP common data model (OMOP CDM)[9] format. OSIM2 comprises 2 modules, Analysis and Simulation, which can be executed together or independently. Figure 1 below shows an overview of the OSIM2 process. The full OSIM2 program (as an Oracle SQL stored procedure), including detailed documentation and training material is freely available through the OMOP website at http://omop.fnih.org/OSIM2. This paper provides a high-level summary of the model and describes its application to two observational databases.

OSIM2 Analysis Module
The OSIM2 Analysis module performs a detailed analysis and extraction of characteristics, as probability distributions, from an existing healthcare dataset. The analysis is purely empirical using no prior knowledge of biological or medical associations; all resulting probabilities are based solely on observed occurrence rates in the source dataset. The results of the Analysis module are stored in Simulation Attribute tables which contain descriptive metrics, multinomial distributions, and single-state transition probability distributions. The state table distributions are used to model derived data characteristics, including the transition from no disease to incident diagnoses, the transition from incident diagnoses to subsequent reoccurrences of the same diagnosis, and from recorded diagnoses to subsequent drug (prescription) occurrences. Each of these transition matrices are populated based on analysis of the source database and are described in more detail in Figures 2 and 3. An additional optional table can be created manually by the user that contains the attributes of drug treatment effects (outcomes). These
tables, which can be saved and distributed independently from the original source database, are used as input to the Simulation module of OSIM2 to simulate observational data with similar characteristics.

**Figure 1**: Overview of OSIM2 Process. The analysis uses real observational data as input to produce 13 empirical distributions, which are then used to simulate patients with those characteristics.

**OSIM2 Simulation Module**

The OSIM2 Simulation module uses an individual Monte Carlo approach where persons are generated using a random probability at each decision point to select a value from the empiric multinomial distributions generated during the Analysis module. The person simulation model follows a four-stage process of 1) creating a simulated population (with association person-level demographics and periods of observation), 2) generating a set of background diseases for each individual, 3) assigning treatment (drug exposure) to the patients based on their underlying disease, and 4) injecting causal effects between treatments and subsequent outcomes. These stages are described in more detail below.

**Stage 1: Creating a Simulated Population**. Simulated persons are each assigned gender, age, number of distinct conditions, and observation duration. Gender is assigned by a random draw according to the observed *Gender Probability* multinomial distribution. Age is assigned according to the *Age Probability* multinomial distribution for the assigned gender. The number of distinct conditions is assigned according to the *Distinct Condition Count Probability* multinomial distribution for the assigned gender and age. The number of distinct conditions is used by the simulation as the general measure of wellness for each person, and condition count ranges, (by default 0-2, 3-7, 8-25, 26+) are used as a strata in many of the probability tables. Each person’s observation duration is assigned from the *Observation Time Probability* multinomial distribution for persons of the same gender, age, and condition count range.

**Stage 2: Generating a Set of Background Diseases for Each Individual**. Conditions are recorded observations about a disease; they represent indications, co-morbidities, outcomes and/or adverse events. One condition can have multiple recorded occurrences during a person’s observation period. In OSIM2, incident conditions for each person are simulated based upon the *First Incident Condition Transitions* matrix containing the multinomial distributions for changing from the current condition state to the next condition state, which is stratified by gender, age range (0-5, 6-13, 14-19, 20-54, 55-69, 70+), and condition count range. The time interval between the incident conditions is also drawn from this distribution. An incident condition by definition can never transition to the same condition or to any other condition that has already been assigned to that person. For each person, additional incident conditions continue to be simulated until the number of conditions associated with that person is reached.
For each incident condition, the number of subsequent condition occurrences to record for that condition is simulated using the *Condition Occurrence Count Probability* distribution, which is stratified by the condition, gender, age range, condition count range, and time remaining in the observation period (in full semi-year periods). The time interval between each recorded occurrence is randomly drawn based from the *Condition Re-occurrence Transitions*, a multinomial distribution of condition to same condition transitions that is further stratified by condition count range and time remaining in the observation period. Using time remaining to stratify interval durations ensures that the simulated observation period durations are maintained by never returning durations that extend beyond the time remaining. Figure 2 provides a graphical representation of the condition simulation, using first order transition matrices to simulate incident conditions and subsequent occurrences of each condition.

**Figure 2:** Condition Simulation Model

![Condition Simulation Model](image)

The Condition Simulation uses the state transition tables generated in the Analysis Module to model the occurrence of incident conditions, followed by the subsequent reoccurrences
- C1-Cn: reflect all incident conditions in the database
- Ci(m): reflects the m number of occurrences of condition C in the database
- \( p_{ij} = \Pr(C_{k+1}=j|C_k=i, C_1..C_k \neq j) = \Pr(C_{k+1}=j|C_1=c_1, C_2=c_2, ..., C_k=c_k) \)
- Time-homogenous: probabilities are independent of k
- Null recurrent: condition j cannot be revisited
- \( p_{ij} \) can be estimated from real data based on frequency of condition co-occurrence
- Model replicated within age * gender * condition count strata

**Stage 3: Assigning Treatment (Drug Exposure) to the Patients Based on their Underlying Diseases.** Drug treatments occur throughout a patient observation period to treat underlying conditions. It can be difficult from empirical observation to fully derive the condition associated with each treatment. We derived the most probable associations by recording a relationship between conditions and initial drug therapies that follow in the gap between that condition and the next recorded condition. Condition / Drug relationships are simulated with the *Condition First Incident Drug Transitions* matrix containing the multinomial distributions for changing from the current condition state to the next drug state. In the observational data, it is common for several conditions and drugs to occur in clusters on the same day since they are recorded during an office visit or hospitalization. Multiple conditions occurring on the same day preceding the gap are each given an equal probability for the association to the following initial drug therapies.
During the simulation process, the treatment simulation draws for the number of distinct drug therapies to simulate for each person from the Distinct Drug Count Probability table, which is stratified by gender, age range, and condition count range. After determining the number of distinct drugs to simulate, the simulation examines each simulated condition record (Figure 3), not just incident conditions. The simulated conditions are processed in a random order to prevent simulated drug therapies from clustering at the beginning of patient’s observation period. A likely number of drugs to associate to the condition in the subsequent gap is drawn from the Condition Drug Count Probability table, and that number of draws is made for an initial distinct drug therapy. Initial drug therapies continue to be simulated until the distinct number of therapies is reached.

For each simulated distinct drug therapy, the simulation randomly draws for the total number of drug exposures, the combined sum of all drug exposures, and the calendar duration from the start of the first exposure to the end of the last from the Drug Occurrence Count and Drug Duration Probability table. The simulation then randomly distributes the desired number of drug exposures throughout the simulated total drug duration. Figure 3 provides a graphical representation of the simulation of assigning drug therapies and exposures given conditions.

Figure 3: Drug Simulation Model

![Drug Simulation Model Diagram]

The drug simulation uses the state transition table generated in the Analysis Module to model incident drug occurrences in gaps between conditions, followed by the subsequent drug reoccurrences.

- \( C_1^{\rightarrow}C_n \): reflect all conditions in the database
- \( D_1^{\rightarrow}D_m \): reflect all initial drug instances in the database
- \( D_{i(q)} \): reflect the \( q \) number of occurrences of drug \( D_i \) in the database
- \( p_{ij} = Pr(D_{k+1}=j|C_k=i, D_1..D_k\neq j) \)
- \( p_{ij} \) can be estimated from real data based on frequency of condition/drug co-occurrence
- Model replicated within age * gender *condition count * drug count strata
- Model preserves conditional independence between drugs and subsequent conditions

Stage 4: Injecting Causal Effects between Treatments and Subsequent Outcomes. After all persons have been generated, the final stage of the Simulation algorithmically introduces known relationships between drugs and outcomes. Treatment effects are defined by the drug, the condition, the effect size, and the time-to-event relationship. Multiple treatment effects can be injected into the simulated dataset simultaneously. The relative risk is a multiplier to be applied to the existing background rate during the time-at-risk for the exposed population to determine the number of conditions to add or remove.
METHOD EVALUATION

Two key aspects of OSIM2 were evaluated: 1) feasibility of the method to apply across disparate real-world healthcare databases and 2) consistency of the simulated data to reflect the key characteristics of real data.

Feasibility

To test the feasibility of the approach, OSIM2 was executed against two different types of healthcare databases, Thomson Reuters MarketScan® Lab Database (MSLR) claims database and GE Centricity (GE) electronic medical record database. Two simulated databases were created from these databases, each containing 1 million records. In the final step of the simulation, each database was injected with pre-defined drug / outcome relationships.

MSLR is 1.5 million person employer-based aggregated claims database, with inpatient and outpatient medical claims as well as pharmacy dispensing records. Execution of the OSIM2 Analysis module against MSLR database required 5.25 hours and the resulting Simulation Attribute tables of database characteristics were approximately 3.9 gigabytes. Using these tables as input, the OSIM2 simulation module simulated 1 million persons in 20.3 hours.

The GE database 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. Execution of the OSIM2 Analysis module against the GE database required 15.5 hours and resulted in approximately 5.5 gigabytes of Simulation Attributes tables. Using these tables as input, the OSIM2 simulation module simulated 1 million persons in 59.9 hours.

Additionally, OSIM2 was executed on three other Thomson Reuters healthcare databases. MarketScan Medicaid Multi-State Database (MDCD) contains administrative claims data for Medicaid enrollees from multiple states for 11.1 million persons. MarketScan Medicare Supplemental and Coordination of Benefits Database (MDCR) contains administrative claims for 4.4 million retirees with Medicare supplemental insurance paid for by employers, including services provided under the Medicare-covered payment, employer-paid portion, and any out-of-pocket expenses. And finally, the MarketScan Commercial Claims and Encounters (CCAE) captures private de-identified administrative claims from inpatient and outpatient visits and pharmacy claims of multiple insurance plans for 58 million persons. Table 1 describes the technical and physical characteristics of executing OSIM2 Analysis Module against each database.

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

Table 1: Technical and physical characteristics of executing OSIM2 Analysis Module against each database

Consistency of Results

To evaluate the consistency of the simulated data to model real world complexities, a dashboard was developed, containing a series of queries comparing characteristics of the simulated data to those same characteristics found in real-world populations. The characteristics that were evaluated (Table 2) include database summary statistics (persons, conditions and drugs per person), demographic distributions, number of distinct conditions and condition occurrences, condition / condition co-occurrence, condition prevalence, number of distinct drugs and drug occurrences, drug prevalence, condition / drug co-occurrence, and drug / drug co-occurrence. The population
characteristics illustrated in Table 2 are modeled with high fidelity compared to the source data. The age and gender of patients in each of the simulated databases accurately reflects the distributions of those patients found within the source data, although there are differences in the characteristics of simulated databases that represent underlying differences in claims and EHR data. In both simulated databases there is a higher percentage of females; and although the mean age is similar in both types of data, there are more of both the oldest and youngest patients found within the EHR source and simulated data.

<table>
<thead>
<tr>
<th>Gender</th>
<th>OSIM2 GE</th>
<th>OSIM2 MSLR</th>
<th>MSLR GE</th>
<th>MSLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>57.6%</td>
<td>64.7%</td>
<td>57.5%</td>
<td>64.8%</td>
</tr>
<tr>
<td>Male</td>
<td>42.4%</td>
<td>35.3%</td>
<td>42.5%</td>
<td>35.2%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>OSIM2 GE</th>
<th>OSIM2 MSLR</th>
<th>MSLR GE</th>
<th>MSLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAN (STDEV)</td>
<td>39.6 (22.0)</td>
<td>39.7 (22.0)</td>
<td>38.1 (17.5)</td>
<td>37.1 (17.7)</td>
</tr>
<tr>
<td>&lt;10</td>
<td>11.0%</td>
<td>8.7%</td>
<td>10.9%</td>
<td>9.0%</td>
</tr>
<tr>
<td>10-19</td>
<td>10.1%</td>
<td>9.4%</td>
<td>10.1%</td>
<td>9.7%</td>
</tr>
<tr>
<td>20-29</td>
<td>14.0%</td>
<td>13.7%</td>
<td>14.0%</td>
<td>13.8%</td>
</tr>
<tr>
<td>30-39</td>
<td>14.2%</td>
<td>17.5%</td>
<td>14.3%</td>
<td>17.7%</td>
</tr>
<tr>
<td>40-49</td>
<td>15.1%</td>
<td>21.5%</td>
<td>15.2%</td>
<td>21.6%</td>
</tr>
<tr>
<td>50-59</td>
<td>13.8%</td>
<td>20.1%</td>
<td>13.8%</td>
<td>19.9%</td>
</tr>
<tr>
<td>60-69</td>
<td>10.5%</td>
<td>7.8%</td>
<td>10.4%</td>
<td>7.2%</td>
</tr>
<tr>
<td>70-79</td>
<td>10.5%</td>
<td>0.9%</td>
<td>10.6%</td>
<td>0.9%</td>
</tr>
<tr>
<td>80-89</td>
<td>0.8%</td>
<td>0.3%</td>
<td>0.8%</td>
<td>0.3%</td>
</tr>
<tr>
<td>90-99</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>100+</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observation Months</th>
<th>MEAN (STDEV)</th>
<th>OSIM2 GE</th>
<th>OSIM2 MSLR</th>
<th>MSLR GE</th>
<th>MSLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAN (STDEV)</td>
<td>24.9 (30.6)</td>
<td>24.0 (31.2)</td>
<td>24.0 (10.1)</td>
<td>21.3 (12.2)</td>
<td></td>
</tr>
<tr>
<td># of distinct drugs/person</td>
<td>8.9</td>
<td>9.0</td>
<td>11.2</td>
<td>11.2</td>
<td></td>
</tr>
<tr>
<td># of total drug records/person</td>
<td>15.0</td>
<td>13.7</td>
<td>17.3</td>
<td>15.4</td>
<td></td>
</tr>
<tr>
<td># of distinct conditions/person</td>
<td>6.9</td>
<td>6.9</td>
<td>14.6</td>
<td>16.3</td>
<td></td>
</tr>
<tr>
<td># of total condition records/person</td>
<td>7.4</td>
<td>7.3</td>
<td>21.4</td>
<td>22.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Population characteristics of the source dataset are maintained in the simulated data

Longitudinal evaluation included comparative analysis of time between similar conditions, time between sequential conditions, number of days with conditions, time between first and last conditions of the same type, time between similar drug starts, time between consecutive drug starts, and number of days with drug starts. Correlation coefficients (R^2) were estimated for agreement in condition and drug prevalence, drug exposure length, and condition / condition, condition / drug, and drug / drug co-occurrence. Both occurrence count and longitudinal comparisons yielded high fidelity results with R^2 coefficient values above 0.95. Figures 4-5 provide partial dashboard results for the two simulations described above. Very similar correlative results were obtained for all five
The condition prevalence graphs shown in Figure 4 plot the proportion of patients with each condition in the source data (y-axis) vs. the proportion of patients with each condition in the simulated data (x-axis). The prevalence of individual conditions is highly correlated between the simulated and source data in both databases ($R^2$: GE= 0.997, MSLR=0.954). Three conditions have been highlighted to demonstrate the condition assignment consistency with gender as well as differences between the source databases that are reflected in the simulation: essential hypertension is very prevalent overall, and among both genders; urinary tract infection is more commonly associated with females; and benign prostate hyperplasia should be gender specific, but the simulation maintains the very small prevalence in females that occurs in the source database. In addition urinary tract infectious disease occurs at an approximate rate of 5% in the both the real-world and the simulated GE data, while it occurs at a rate of approximately 13% in both the real world and the simulated MSLR data.

Figure 5 shows the proportion of patients with co-occurrence between each condition in the real and simulated data in the top panel, and comparison of real vs. simulated data for the prevalence of condition-drug co-occurrence in the bottom panel. There is good agreement between co-morbidities ($R^2$: GE= 0.967, MSLR= 0.952), with common diseases, like hypertension and hyperlipidemia, co-occurring at similar rates in both databases Drug treatments commonly co-occur with the indicated disease, such as ACE inhibitors following hypertension or bisphosphonates after osteoporosis diagnosis Results indicate that condition co-occurrence is highly correlated between the simulated and source datasets, although co-occurring conditions with the highest prevalence occur slightly more often in source versus simulated data in the MSLR database simulation. This pattern is not seen in the GE condition co-occurrence simulation. Condition / drug co-occurrence pairs exhibit similar results, but the pairs with the highest prevalence occur slightly more often in both the source MSLR and GE databases.
DISCUSSION
This paper describes an empirical model developed to simulate longitudinal healthcare data by first extracting key characteristics of real world healthcare databases, and then executing a simulation model that uses those characteristics to create simulated patients with longitudinal records of condition and drug occurrences. We demonstrated the feasibility and assessed the performance of the approach by simulating two real-world healthcare databases. OSIM2 was found to model several of the key complexities found within healthcare data with high fidelity, making it a valuable resource for augmenting observational data methods research and development. The model preserves conditional independence so that no treatment-outcome associations exist in the simulated data unless injected, which can be very useful for testing the performance of outcome detection methods by applying them to the simulation data before and after injection. OSIM2 provides method developers with simulated data that contain well characterized drug treatment effects against a background of data containing confounding factors that may be present in real-world healthcare data.

There are several limitations to the simulation model described in this paper. While OSIM2 models person-level longitudinal data by generating background conditions and drug treatments in a temporal fashion, it does not precisely approximate the clustered nature of encounter-based data. The model assumes the relationships between conditions and drugs are stable over time, which means changes in coding behavior, physician prescribing patterns, and new product introductions are not accurately represented. The first-order condition model does not represent the true complexity of co-morbidities and disease progression. Similarly, if the drug records analyzed by OSIM2 are individual ingredients (vs. branded drugs), the simulation program is unable to accurately model combination products. In addition, the OSIM2 model does not consider the effects that dose or formulations may have on treatment effects and the length of drug exposures has no dependency on primary or co-morbid conditions triggering the therapy. The current version of the model does not include all possible fields available in the OMOP CDM (e.g., race, geography) or all available data elements that can be captured in observational sources (such as visits, procedures, and laboratory results). Data characteristics that are not explicitly described in the empiric distribution tables may not be accurately represented within the simulated datasets, such as co-prescription of drugs. Given these limitations and simplifying assumptions, methods evaluation against OSIM2 should be considered a best-case
scenario, as methods will be unlikely to yield more accurate predictive performance against real-world data given the added complexities and potential sources of bias. OSIM2 was designed to facilitate observational analysis methods development for areas such as active medical product safety surveillance and comparative effectiveness. Simulated healthcare data may have other potential uses, such as in quality improvement and clinical decision support, but is limited both to the scope of data elements contained within the OMOP CDM as well as to the source data for which the simulation is run against. Those skeptical of the use of observational EHR data for knowledge discovery should be more skeptical of the simulated EHR data given the additional level of assumptions that are imposed.

Even with the limitations imposed by the first-order condition model, the known characteristics of the simulated data produced by OSIM2 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. Drug / outcome pairs that are injected into the model can represent both positive and negative treatment effects, making the model appropriate for both drug safety and comparative effectiveness research. A simulation program such as OSIM2 can provide a mechanism to facilitate data exchange across the research community to understand differences in the underlying data characteristics of alternative data sources, without jeopardizing patient privacy or commercial interests. OSIM2 represents an informatics solution to enable the systematic exploration of the performance of alternative methods in their ability to identify true effects and discern from false positive findings, as well as the standardized evaluation the impact of data characteristics (such as database size, source population) and different drug-outcome scenarios on method behavior.

All stakeholders (medical products manufacturers, regulators, payers, providers, and patients) require reliable evidence about the effects of healthcare interventions to inform medical decision-making. Empirical research is necessary to assess the accuracy of observational evidence. Simulated data offers a promising opportunity to engage the scientific community in studying the outstanding research questions in a systematic, transparent, and reproducible fashion.

REFERENCES

4. Fatih Safa Erenay, PhD, Oguzhan Alagonz, PhD, Ritesh Banerjee, PhD, Robert R. Cima, MD, MA. Estimating the Unknown Parameters of the Natural History of Metachronous Colorectal Cancer Using Discrete-Event Simulation. Downloaded from mdm.sagepub.com at Ortho-McNeil Janssen Scientific Affairs on March 3, 2011.