Feasibility and utility of applications of the common data model to multiple, disparate observational health databases

RECEIVED 30 June 2014
REVISED 2 October 2014
ACCEPTED 11 November 2014
PUBLISHED ONLINE FIRST 10 February 2015

Erica A Voss[1], Rupa Makadia[1], Amy Matcho[1], Qianli Ma[1], Chris Knoll[1], Martijn Schuemie[1], Frank J DeFalco[1], Ajit Londhe[2], Vivienne Zhu[1], Patrick B Ryan[1]

ABSTRACT

Objectives To evaluate the utility of applying the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) across multiple observational databases within an organization and to apply standardized analytics tools for conducting observational research.

Materials and methods Six deidentified patient-level datasets were transformed to the OMOP CDM. We evaluated the extent of information loss that occurred through the standardization process. We developed a standardized analytic tool to replicate the cohort construction process from a published epidemiology protocol and applied the analysis to all 6 databases to assess time-to-execution and comparability of results.

Results Transformation to the CDM resulted in minimal information loss across all 6 databases. Patients and observations excluded were due to identified data quality issues in the source system, 96% to 99% of condition records and 90% to 99% of drug records were successfully mapped into the CDM using the standard vocabulary. The full cohort replication and descriptive baseline summary was executed for 2 cohorts in 6 databases in less than 1 hour.

Discussion The standardization process improved data quality, increased efficiency, and facilitated cross-database comparisons to support a more systematic approach to observational research. Comparisons across data sources showed consistency in the impact of inclusion criteria, using the protocol and identified differences in patient characteristics and coding practices across databases.

Conclusion Standardizing data structure (through a CDM), content (through a standard vocabulary with source code mappings), and analytics can enable an institution to apply a network-based approach to observational research across multiple, disparate observational health databases.

Key words: database, factual vocabulary, controlled health services research, medical informatics observational study

BACKGROUND AND SIGNIFICANCE

Observational health data sourced from electronic health records (EHRs), insurance/administrative claims, hospital billing, clinical registries, and longitudinal surveys are of increasing importance for research in population health. The reuse of data already collected from these various sources provides researchers with large, heterogeneous patient populations that are geographically dispersed, at generally lower costs than if data were collected by prospective data collection or randomized clinical trials.[1–3] Medical product safety surveillance is one area that has garnered substantial attention in recent years. Several initiatives are currently underway to develop the science and technology to leverage observational data to study the effects of medical products, including Mini-Sentinel, EU-ADR,[12] and Observational Medical Outcomes Partnership (OMOP).[4–12] One consistent theme across these initiatives is the recognition that standardization of a data model and vocabulary is imperative to performing efficient research, clinical discovery, and adverse event surveillance.

One of the obstacles to using observational data is gaining sufficient understanding of each data source; each is unique and even databases sourced from the same type of data can have large differences in schema, format, and coding usages. Comparing 2 popular US administrative claims datasets, Optum Clinformatics DataMart (Optum; Optum, Inc, Eden Prairie, Minnesota) and Truven Health MarketScan (Truven Health Analytics Inc, Ann Arbor, Michigan) Commercial Claims and Encounters (CCAE), demonstrates some of the challenges.

Correspondence to Erica A Voss, 920 Route 202, Raritan, NJ 08869; evoss3@its.jnj.com

For affiliation see end of article.

Since US claims databases are derived off two standard forms, health Insurance Claim Form-1500 and Universal Billing form 92, one might assume the 2 databases would have similar content and structure; however, this is not the case. When looking for conditions in Optum, 1 data table contains 5 columns with International Classification of Diseases, Ninth Revision (ICD9) codes, and CCAE houses 4 data tables with 6 to 16 columns containing the same codes. When working across data sources from different countries, there are added challenges with the usage of different source vocabularies. Within the Clinical Practice Research Datalink (CPRD), diagnoses are coded using Read Codes instead of ICD9s, and such building definitions of diseases to go across databases with different source vocabularies would require multiple, independent code lists.

In addition to the disparate coding standards, data source–specific proprietary vocabularies create additional challenges. Premier Perspective (Premier; Premier, Inc, Charlotte, North Carolina), a US hospital billing dataset, has its own proprietary billing codes, which can be extremely important in understanding what drugs were dispensed and procedures performed during a visit, but no other databases use these codes. Idiosyncrasies between datasets, both in their format and coding practices, make them difficult and time-intensive to use for research in a systematic manner.

Of the several observational research initiatives that have identified standardization as necessary to work with the data, OMOP, specifically, has developed the OMOP Common Data Model (CDM) v4[13] and OMOP Vocabulary v4[14] to address the standardization issue. The motivation behind the OMOP CDM is to enable transformation of data from diverse observational databases into a common format with a standardized vocabulary, which can then be used to perform systematic analysis.[3,5,15–22] The OMOP CDM is a person-centric model that accommodates different data domains typically found within observational data (demographics, visits, condition occurrences, drug exposures, procedures, and laboratory data). Each individual data domain is modeled as a specific table which supports capture of data elements specific to that domain (ie, DAYS_SUPPLY is a column in the DRUG_EXPOSURE table within this model) and is designed to enable queries in an efficient manner. The OMOP Vocabulary, used to standardize the codes or terminologies used within the raw data, is tightly intertwined within the OMOP CDM. For each domain, 1 or more vocabularies are defined as the standard reference vocabulary set to which all source-coding systems are mapped. For example, for drugs, the standard reference vocabulary is RxNorm, and the OMOP Vocabulary contains mappings from other dictionaries to RxNorm. Drug exposures that may be captured in US databases through National Drug Codes can also be coded as procedures using ICD9-Procedure (ICD9-PROC) and as Healthcare Common Procedure Coding System codes. CPRD uses its own standard, Multilex, for drugs. The OMOP Vocabulary allows all these source codes to be translated into RxNorm[23] during transformation to the OMOP CDM. If a researcher wants to find a specific active ingredient across CDMs, a standard RxNorm concept can be used to retrieve all drug exposure records in 1 standardized table regardless of how the raw data were structured or coded.

Many organizations have access to multiple patient-level datasets and attempt to conduct analyses across these sources to answer research questions of interest to the institution. For example, pharmaceutical research organizations may license deidentified administrative claims and electronic health records datasets from multiple sources. To date, no literature has demonstrated the potential use of the OMOP CDM across multiple, disparate databases within 1 institution.

OBJECTIVES

The objective of this study is to explore the benefits and costs associated with standardizing a network of disparate observational health databases into the OMOP CDM and Vocabulary. We aim to evaluate the standardization process in terms of its impact on the quality, efficiency, and consistency of observational database research. We aim to demonstrate how standardization can work in practice through the replication of the cohort construction process, using an existing epidemiology protocol published by the US Food and Drug Administration that compares the use of warfarin versus rivaroxaban in patients with atrial fibrillation.

MATERIALS AND METHODS

We used 6 databases for this research: Premier, Optum, CPRD, CCAE, Truven Health MarketScan Medicaid (MDCD), and Truven Health MarketScan Medicare Supplemental (MDCR). Table 1 provides high-level information about each database. Optum, CCAE, MDCD, and MDCR are claims databases. Premier is a hospital billing database and CPRD is a UK general practitioners (GPs) database. Depending on the specific licensing agreement, it is possible to have data that spans more or less time than reported here. The use of Optum, Premier, CCAE, MDCD, and MDCR was reviewed by the New England Institutional Review Board and was determined to be exempt from broad Institutional Review Board approval as this project did not involve human subject research. Approval for CPRD was provided by the Independent Scientific Advisory Committee.

OMOP CDM transformation

The process of extracting, transforming, and loading (ETL) data into the OMOP CDM differs for each database. We describe the general process and then highlight database specifics.

When building the ETL, the data were first categorized with an open source tool called WhiteRabbit,[24] listing all tables, fields, and distinct values in those fields. WhiteRabbit analyzes the structure and content of a database and exposes data anomalies that the ETL will need to handle. Prior to developing this tool, CDMs were transformed based on experience with the data, and we found that the exceptions within the data were often more numerous than foreseen and required considerable time to handle. It should be highlighted that researcher experience with a data source, in addition to the insights

Table 1: High-level Information about each dataset

Statistic	Premier Perspective	Optum	CPRD	Truven CCAE	Truven MDCR	Truven MDCD
High-level Description	A hospital transactional database that includes emergency, inpatient, and outpatient visits for patients who visit a Premier hospital. Includes commercially insured, government plans, and charity care.	An administrative health claims database for members of United Healthcare, who enrolled in commercial plans (including ASO, 36.31 M), Medicaid (prior to July 2010, 1.25 M), and Legacy Medicare Choice (prior to January 2006, 0.36 M) with both medical and prescription drug coverage.	Anonymized longitudinal electronic health records from primary care practices in UK. Patient management system with many aspects of patient care covered, including diagnoses, prescriptions, signs and symptoms, procedures, labs, lifestyle factors, clinical, and administrative/social data	An administrative health claims database for active employees, early retirees, COBRA continues, and their dependents insured by employer-sponsored plans (individuals in plans or product lines with fee-for-service plans and fully capitated or partially capitated plans).	An administrative health claims database for Medicare-eligible active and retired employees and their Medicare-eligible dependents from employer-sponsored supplemental plans (predominantly fee-for-service plans). Only plans where both the Medicare-paid amounts and the employer-paid amounts were available and evident on the claims were selected for this database.	An administrative health claims database for the pooled healthcare experience of Medicaid enrollees from multiple states.
Source Codes Used	–	–	–	–	–	–
• Conditions	ICD9	ICD9	Read	ICD9	ICD9	ICD9
• Drugs	Premier Standard Charge Code	NDCs, HCPCs, ICD9-PROC	Multilex, native immunization codes	NDCs, HCPCs, ICD9-PROC	NDCs, HCPCs, ICD9-PROC	NDCs, HCPCs, ICD9-PROC
• Lab Data	Premier Standard Charge Code	LOINC[a]	Native test codes	LOINC[a]	LOINC[a]	–
Region	United States	United States	United Kingdom of Great Britain	United States	United States	United States
Date Ranges	December 1998 - 2013	October 2005 - December 2012	January 1987 - July 2013	January 2000 - October 2013	January 2000 - October 2013	January 2006 - October 2012
No. of Overall Patient Count	100 092 900	36 229 849	11 485 373	108 589 866	8 216 678	16 172 699
Age at Start in Database, mean (SD), y	38.80 (24.33)	31.43 (18.95)	32.98 (23.07)	31.20 (18.13)	72.36 (8.10)	22.45 (22.56)

Abbreviations: Optum, Optum Clinformatics DataMart; CPRD, Clinical Practice Research Datalink; Truven CCAE, Truven Health MarketScan Commercial Claims and Encounters; Truven MDCD, Truven Health MarketScan Medicaid; Truven MDCR, Truven Health MarketScan Medicare Supplemental; SD, Standard Deviation; ICD9, International Classification of Diseases, Ninth Revision; ICD9-PROC, International Classification of Diseases, Ninth Revision–Procedure Codes; LOINC, Logical Observation Identifiers Names and Codes; NDC, National Drug Code; HCPC, Healthcare Common Procedure Coding.

[a]Results for laboratory tests processed by large national lab vendors that provide data back to the data vendor.

provided by WhiteRabbit, substantially reduced the number of iterations required to successfully account for potential data conversion issues. Next, we documented each ETL with a tool called RabbitInAHat. This interactive application takes the results from WhiteRabbit and allows the user to connect data tables and columns from the raw dataset to where they will map into the OMOP CDM dataset. The output RabbitInAHat is a requirements document for building an ETL. Using this document, a CDM Builder program was developed to transform raw data into the CDM. We implemented CDM Builders as a C# application that processed ETL logic on a distributed, parallelized computing infrastructure. CDM Builder development included several rounds of testing where another developer would perform independent programming to recreate logic with SAS or SQL and iterating until results matched with CDM Builder results. Once a CDM was deemed valid, it was then released to researchers within the organization.

Since every observational dataset is unique, each CDM Builder has unique properties, some of which are discussed below. Full information on individual CDM transformation can be found on the OMOP website (http://omop.org/CDM).[25–29]

Premier

In Premier, all charges are recorded as standard charge codes, which are free text. By applying fuzzy string text matching to these records, we were able to map drugs and procedures to standard vocabularies. Additionally, we converted the provided within-visit chronology of events to approximate dates to allow standard analytics to be used.

Optum

We developed a standard convention for defining visits from administrative claims data based on revenue codes, which allowed consistent application across Optum and the Truven datasets. The heuristic enabled disambiguation between outpatient visits, emergency department visits, and inpatient admissions while also consolidating multiple claims that are part of the same episode of care.

CPRD

All lifestyle and clinical data such as smoking status and body mass index were transformed to the CDM. CPRD raw lifestyle/clinical data are housed in 2 tables. Within these tables, each data category (eg, smoking) has a varying number of data elements (eg, status, cigarettes per day, cigars per day), and these data elements are associated with varying lookups. We created an algorithm to process all data elements in the same manner despite the unusual format described above. In addition, because drug exposure duration was only provided for 7% of prescriptions, an algorithm was developed and extensively validated to impute days supplied for a drug record.

Truven CCAE

CCAE has health risk assessment data available, which contains self-reported biometrics, health status, risk behaviors, and behavioral change data. We loaded the data into the observation table with each survey item as 1 unique observation source value, and every reported item for each person on a certain date created 1 row in the observation table.

Each CDM-ETL process includes logic to exclude source data that we do not believe is of sufficient quality for research purposes (these decisions are made with all use cases of the CDM in mind, not just for a specific research question). For example, we applied a consistent set of logic that excluded patients if the source data indicated multiple genders or multiple year of birth records that were more than 2 years apart, because we found these instances suggested source data errors or inaccurate patient identifiers that were being applied to multiple individuals. The development of the ETL enabled a transparent process to codify and document issues with the raw data and to apply consistent decisions about which data should be made available for researchers. Different research groups may choose to make different decisions within their CDM implementations, but the process of designing and implementing an ETL specification allows those decisions to be exposed to the broader research community.

Leveraging the OMOP Vocabulary

The OMOP CDM provides a common format for diverse raw dataset, and integration of the OMOP Vocabulary into the CDM is a requisite process (detailed information on the OMOP Vocabulary and its curation and maintenance process can be found at http://omop.org/Vocabularies). The OMOP Vocabulary is a downloadable dataset that aids in translating source codes (eg, ICD9 or National Drug Codes) during the ETL process into what OMOP considers standardized terminologies (eg, Systemized Nomenclature of Medicine [SNOMED] or RxNorm). This transformation allows different observational datasets to essentially "speak the same language" when a researcher performs an analysis. Not all source codes from observational data can be found within the OMOP Vocabulary. Some codes are proprietary to the database or other source code sets have not yet been integrated with the Vocabulary. Any code lookup that does not currently exist in the OMOP Vocabulary will be created and appended to the OMOP Vocabulary.

Analysis across datasets

To demonstrate the utility of standardizing disparate data sources into a CDM, we replicated a published observational study protocol and evaluated the quality of a standardized approach and time-to-execution. As an exemplar, we used the Mini-Sentinel analysis of the comparative effectiveness of rivaroxaban versus warfarin on various outcomes in patients with atrial fibrillation.[30] We developed a standardized analytic routine that replicated the cohort definitions within the protocol and applied the analytic program across all 6 databases to compare the impact of the inclusion criteria on the proportion of patients qualifying for the study.

Specifically, we identified all new users of each target drug (warfarin and rivaroxaban) who satisfied the following 7 criteria of the original study: (1) had at least 183 days of nonexposure before the first target drug exposure; (2) had at least 1 atrial fibrillation or atrial flutter diagnosis code within the 183-day window prior to first exposure; (3) did not have any prior diagnosis or procedure codes indicative of long-term dialysis; (4) did not have any prior diagnosis or procedure codes indicative of kidney transplant; (5) did not have any prior diagnosis or procedure code indicative of mitral stenosis or mechanical heart valve; (6) did not have any prior procedure code indicative of joint replacement or arthroplasty surgery; and (7) did not have prior use of any anticoagulant (warfarin, rivaroxaban, dabigatran, or apixaban). For each target drug, we created 2 cohorts: new users of the drug (defined by satisfying criteria No. 1), and the subset of those new users of the drug who satisfied the remaining 6 criteria. For each cohort, we produced a standardized descriptive summary of the population, including demographics (gender and age distribution), comorbidities (prevalence of conditions in time window prior to cohort entry), concomitant medications (prevalence of drug exposure in time window prior to cohort entry), and service utilization (prevalence of procedures in time window prior to cohort entry). We measured the execution time for the standardized analytic routine when applied to each target drug across all 6 databases. Analyses were conducted on a Microsoft Server 2008 (Microsoft Corporation, Redmond, Washington) with an AMD Opteron 6172 (Advanced Micro Devices, Inc, Sunnyvale, California), 2.10 GHz, 2 processors, 24-core CPU, and 256 GB of RAM. Each CDM was stored in a separate database within an instance of Microsoft SQL Server 2012 (Microsoft Corporation, Redmond, Washington).

Appendix 1 contains the standard concepts and corresponding source codes that were used to define each of the core concepts required within the prespecified protocol.

RESULTS

When transforming a raw dataset into a common format, information loss is a concern.[20] Table 2 explores data loss from 4 perspectives: (1) exclusion of patients; (2) data records excluded because they were outside defined observation periods; (3) data types in the raw schema which could not be loaded into the OMOP CDM because there were no equivalent tables or data fields; and (4) source codes which could not be mapped to the common OMOP Vocabulary coding systems. Less than 2% of patients were excluded in Premier, Optum, and MDCD; however, for CPRD, CCAE, and MDCR almost a quarter of the patients were excluded. The primary reason for patient exclusion in all the databases was because the source data had anomalies that made us believe the data was not of sufficient quality for research purposes. As previously mentioned, we applied a consistent set of logic that excluded patients if the source data indicated multiple genders or multiple year of birth records that were more than 2 years apart. We also excluded patients with a year of birth less than 1900 or greater than the current year because these were considered to be an irreconcilable data anomaly. In Truven CCAE and MDCR, the primary reason for patient exclusion was the requirement for each patient to have had at least 1 period of observation with both medical and pharmacy coverage, since the majority of our research is drug safety surveillance and comparative effectiveness research where it is necessary to have information about both drug exposure and outcome incidence. We applied this logic against the entire dataset so that it was consistently applied within all specific research studies. In CPRD, patients were only included if they met CPRD-defined acceptability criteria and had valid observation time in the database. An observation period was defined as the period for which we believed we had data capture for a person and, most importantly, when absence of recorded events could be interpreted (up to a point) as absence of events. We saw only a small loss of information by discarding events outside of observation (only considering data for patients who were included in the CDM), ranging from 0.0% (Premier) to 1.9% (MDCR) with the exception of CPRD, which had an information loss of 21.7% (this loss comprised prior history records that GPs submitted later in time). In all CDMs, all data domains were included—there was no data domain in the raw data that could not be transformed into the CDM format.

Not all source codes could be mapped to an OMOP Vocabulary concept; unmapped codes were assigned a concept ID of 0. All source data were still maintained within the CDM, regardless of whether the source code could be mapped into one of the standardized vocabularies. In Premier, CPRD, CCAE, MDCD, and MDCR, we were able to map 92.3% (Premier) to 98.2% (CPRD) of the unique condition source codes to a code in the OMOP common coding system (SNOMED for conditions), corresponding to 96.8% (Premier) to 99.8% (CPRD) of the data records. For Optum, 29% of the condition source codes could be mapped; however, this represented 98.7% of the data records (ie, there were many codes that we could not map for Optum, but most of them were not valid codes or were not commonly used). For the drug codes Premier, Optum, CCAE, MDCD, and MDCR, all had between 81.0% (MDCR) to 86.6% (Premier) of the unique source codes mapped to the common coding system (RxNorm), and those drug source codes represented 90.5% (Premier) to 98.6% (MDCR) of the data records (for Premier the majority of the drop off was due to unmapped standard billing). For CPRD, only 38.9% of the drug source codes could be mapped, representing 89.9% of the total data records; the majority of most prevalent unmapped drug exposures in the data were medical devices/supplies and over-the-counter products.

Once the datasets had been transformed into the CDM, it became straightforward to develop standardized analytics that could be applied consistently across all databases. Figure 1 depicts an example of a standardized tool built as a web application. The tool generates side-by-side visualizations of the CDM data, showing the total number of distinct patients, duration of observation, gender distributions, types of patient visits (ie, emergency department, inpatient, outpatient, and longer term care), age at first observation, and years of first

Table 2: Understanding data loss in CDM transformation

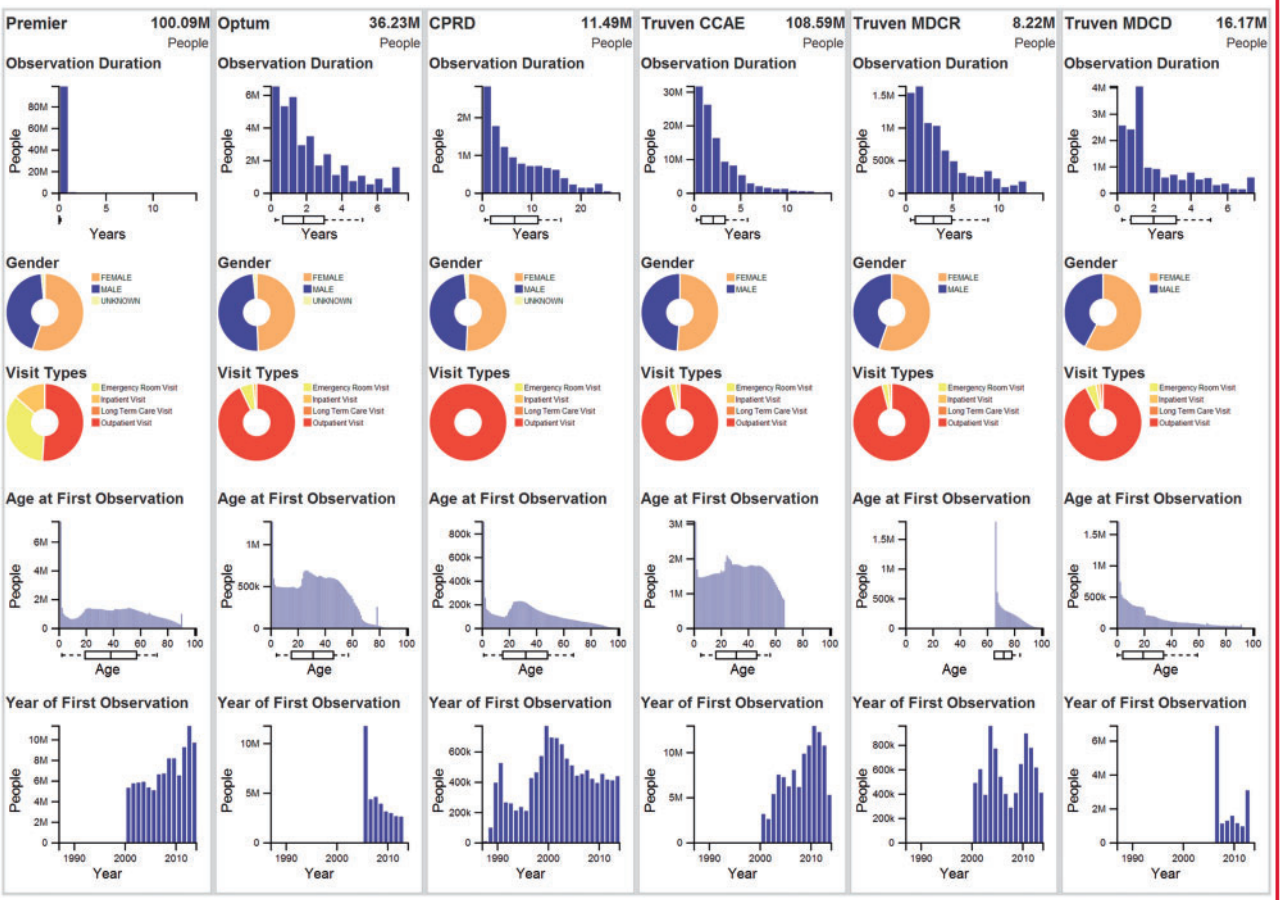
Code Counts	Premier Perspective	Optum	CPRD	Truven CCAE	Truven MDCR	Truven MDCD
Patients excluded, No. (%)	1 354 310 (1.4)	1077 (<0.1)	3 751 558 (24.6)	37 140 364 (25.5)	2 834 999 (25.7)	44,277 (0.27)
Excluded rows outside observation periods, No. (%)	0 (0.0)	1 356 281 (<0.1)	839 237 761 (21.7)	129 235 806 (1.4)	41 905 900 (1.9)	4 669,939 (0.25%)
Information not supported by CDM	None	None	None	None	None	None
Code mapping	–	–	–	–	–	–
Condition codes	ICD9s	ICD9s	Read	ICD9s	ICD9s	ICD9s
No. of unique source codes	15 938	52 993	30 445	14 856	14 282	14,598
Mapped unique source codes, No. (%)	14 717 (92.3)	15 377 (29.0)	29 890 (98.2)	14 325 (96.4)	13 824 (96.8)	14 146 (96.9)
No. of total records	1 526 743 203	1 408 044 548	131 206 276	3 462 089 538	837 145 789	891,097 856
Total mapped records, No. (%)	1 478 322 372 (96.8)	1 390 271 348 (98.7)	130 998 307 (99.8)	3 427 233 910 (99.0)	824 166 146 (98.4)	883 173,325 (99.1)
Drug codes	Standard Charge Code	NDCs[a]	Multilex, Immunizations	NDCs[a]	NDCs[a]	NDCs[a]
No. of unique source codes	1 022 475	73 139	53 836	138 906	97 484	69,986
Mapped unique source codes, No. (%)	884 309 (86.6)	60 854 (83.2)	20 955 (38.9)	96 447(69.4)	78 965 (81.0)	57 435 (82.1)
No. of total records	3 217 360 412	765 800 100	1 143 757 300	2 632 232 959	824 675 757	394 531 395
Total mapped records, No. (%)	2 913 494 490 (90.6)	751 416 033 (98.1)	1 027 644 814 (89.9)	2 577 864 143 (97.9)	813 142 800 (98.6)	384 227 647 (97.4)

Abbreviations: CDM, Common Data Model; Optum, Optum Clinformatics DataMart; CPRD, Clinical Practice Research Datalink; Truven CCAE, Truven Health MarketScan Commercial Claims and Encounters; Truven MDCD, Truven Health MarketScan Medicaid; Truven MDCR, Truven Health MarketScan Medicare Supplemental; OMOP, Observational Medical Outcomes Partnership; ICD9, International Classification of Diseases, Ninth Revision; NDC, National Drug Code.

[a]This group may have multiple types of codes being used; however, we will focus on the largest contributor within the source data.

Voss EA, et al. J Am Med Inform Assoc 2015;22:553–564. doi:10.1093/jamia/ocu023, Research and Applications

Figure 1: Visualizations on observation data in the CDM.
Abbreviations: CDM, Common Data Model; Premier, Premier Perspective; Optum, Optum Clinformatics DataMart; CPRD, Clinical Practice Research Datalink; Truven CCAE, Truven Health MarketScan Commercial Claims and Encounters; Truven MDCD, Truven Health MarketScan Medicaid; Truven MDCR, Truven Health MarketScan Medicare Supplemental.



observation. This graphic illustrates that Premier has the shortest patient duration of less than 1 year (consistent with this database being hospital transactions) and CPRD has the longest duration of over 20 years (consistent with this database being GP-centric). For gender, some databases have about a 50/50 split between male and female (Optum, CPRD, and CCAE), while the others have more females (Premier, MDCR, and MDCD). This figure also shows that there are a small percentage of patients who are of unknown gender within the database. With the distribution of types of visits, we see that Premier has the most inpatient and emergency department visits among all the databases; outpatient data entirely comprises CPRD; and MDCD is the only database with long-term care data. Age at first observation highlights the age diversity—MDCR contains an elderly patient population, MDCD has a large proportion of patients, and the majority of patients in Optum and CCAE are fairly similar. Finally, the year of first observation shows the calendar years of data available for each dataset—CRPD has the most years of observation and MDCD has the fewest.

Analysis across datasets

Table 3 shows the cohort size and execution time across the 6 databases in our internal data network. Within the warfarin cohort, 5 databases had at least 10 000 new users and CCAE had more than 100 000 patients who started warfarin after November 2011 and had at least 183 days of prior observation. The proportion of new users that satisfied all inclusion criteria ranged from 12% (CCAE) to 39% (CPRD); the largest resulting cohort was found in the MDCR with 22 026 patients. The entire analysis (2 cohorts with 7 inclusion criteria and 4 descriptive summary reports run across a network of 6 databases) was executed in 16 minutes when run in parallel and would have been completed in 59 minutes had all analyses been executed in sequential fashion.

Premier is a hospital database in which the observation periods tend to be more episodic in nature. Without many qualifying patients, we decided that Premier was not appropriate for use in a long-term longitudinal study like this. While all summary statistics were generated on the resulting cohorts, we removed them from the manuscript to simplify the presentation.

Table 3: Cohort Size

Data Source	Warfarin				Rivaroxaban			
	No. of New Users	No. of Persons Matching All Criteria	Match Rate, %	Execution Time, MM:SS.ms	No. of New users	No. of Persons Matching All Criteria	Match Rate, %	Execution Time, MM:SS.ms
Premier	17	2	11.76	00:31.7	475	58	12.21	01:23.5
Optum	23 840	3890	16.32	05:18.9	9750	1797	18.43	02:29.0
CPRD	25 073	9860	39.33	04:46.8	1353	184	13.60	01:49.2
CCAE	100 768	12 153	12.06	15:59.6	53 321	8971	16.82	06:47.3
MDCR	67 370	22 026	32.69	10:44.1	34 212	9585	28.02	05:02.7
MDCD	10 165	1514	14.89	03:31.3	1605	157	9.78	01:43.6

Abbreviations: MM:SS.ms, minutes, seconds, milliseconds.

Table 4 highlights the impact of each inclusion criteria on the proportion of eligible patients among the new user cohorts. Across the databases, the requirement for having an atrial fibrillation or atrial flutter diagnosed within the prior 183 days was the most restrictive, with 16% to 44% of warfarin new users and 21% to 55% of rivaroxaban new users satisfying that criteria. This could be due to the drug being used for different indications or the diagnosis code not being recorded within the time window on the claims or EHR system.

We highlight some of the key statistics within the descriptive summaries in Table 5. Across the 5 databases, we saw substantial heterogeneity in the mean age and gender distribution. From the prevalence of prior conditions, we consistently observed across all databases that atrial fibrillation was more commonly recorded than atrial flutter, but CPRD also had a substantial number of patients that qualified based on codes of *atrial fibrillation and flutter* and *paroxysmal atrial fibrillation*. This difference reflects the difference in coding practice across health systems and the value in standardizing vocabulary and analytics to accommodate these variations in a consistent manner. For conditions that may be considered by researchers to be potential outcomes for a comparative effectiveness study of these 2 products (such as acute myocardial infarction, stroke, intracranial hemorrhage, gastrointestinal bleeding), there are substantial differences between the 2 cohorts in the baseline rate of these events prior to exposure that would require attention in order to conduct an appropriate study. Table 4 also highlights differences in drug usage where each source has been standardized to RxNorm, and we applied the OMOP vocabulary to aggregate individual products into drug classes using the World Health Organization Anatomical Therapeutic Chemical classification system.

DISCUSSION

The results of this paper highlight the feasibility and utility of applications of the OMOP CDM to multiple, disparate observational health databases. We highlight both the costs and benefits of such standardization.

One of the potential costs is loss of information. Table 2 shows that not all source codes may map into OMOP Vocabulary concepts. Most loss of information can be attributed to our exclusion rules, which were aimed at improving the quality of the data. By applying these rules during the ETL, all future analyses consistently benefitted from this curation. For the Truven datasets, we included only patients with both medical and prescription coverage to ensure we could research the effects of medical products, and this requirement accounted for about 25% of the patients dropped. Furthermore, sometimes during an ETL, we encountered other information that seemed questionable and therefore needed to be dropped. For example, in Optum and Premier we found cases where patient IDs seemed to be accidentally reused, making it impossible to untangle which data were associated to which person. In each of our databases, we conducted replication analyses using both the raw data and the CDM-transformed data as part of our quality assessment to determine that the transformation did

Table 4: Inclusion Rules

Inclusion rule	Optum	CPRD	CCAE	MDCR	MDCD
Warfarin Cohort, No. (%)					
Warfarin new users	23 840 (100)	25 073 (100)	100 768 (100)	67 370 (100)	10 165 (100)
Have atrial fibrillation or flutter	5093 (21)	11 075 (44)	16 202 (16)	28 499 (42)	1822 (18)
No codes suggestive of chronic dialysis	23 196 (97)	24 842 (99)	98 031 (97)	65 909 (98)	9801 (96)
No kidney transplant	23 761 (100)	25 044 (100)	100 387 (100)	67 211 (100)	10 122 (100)
No mitral stenosis or mechanical heart value	22 944 (96)	24 510 (98)	97 080 (96)	64 245 (95)	9914 (98)
No joint replacement/ arthroplasty surgery	18 344 (77)	22 946 (92)	77 709 (77)	53 675 (80)	9163 (90)
No other anticoagulant use in prior 183 days	23 376 (98)	25 009 (100)	98 831 (98)	65 141 (97)	10 074 (99)
Rivaroxaban Cohort, No. (%)					
Rivaroxaban new users	9750 (100)	1353 (100)	53 321 (100)	34 212 (100)	1605 (100)
Have atrial fibrillation or flutter	3133 (32)	280 (21)	13 696 (26)	18 916 (55)	339 (21)
No codes suggestive of chronic dialysis	9650 (99)	1344 (99)	52 688 (99)	34 016 (99)	1594 (99)
No kidney transplant	9740 (100)	1353 (100)	53 282 (100)	34 191 (100)	1602 (100)
No mitral stenosis or mechanical heart value	9608 (99)	1341 (99)	52 910 (99)	33 219 (97)	1585 (99)
No joint replacement/ arthroplasty surgery	5386 (55)	1140 (84)	32 503 (61)	24 516 (72)	1045 (65)
No other anticoagulant use in prior 183 days	8230 (84)	851 (63)	44 621 (84)	24 003 (70)	1206 (75)

not substantially alter prevalence of disease and treatment or analytical study results.[31,32]

With respect to loss due to code mapping, Optum had fewer codes mapped than others sources, but it reflected more than 90% of the data, which could have been attributed to invalid codes being infrequently used in practice. For example, there were records in Optum medical claims with a diagnosis code of 888.88 or 999.99. These terms are not valid ICD-9-CM codes and therefore are not mapped into the OMOP Vocabulary. It is also important to reinforce that while the CDM provides the opportunity to normalize all codes into a common reference standard that is applied consistently across all databases, the CDM also maintains the source codes from the raw data—the Vocabulary is not used to exclude data. As a result, while the CDM makes it efficient to perform cross-database analyses under a standard vocabulary, it fully supports specific research questions that require analysis with the local source codes (eg, Read Codes and Multilex drugs for CPRD).

One of the benefits of standardization is that data preprocessing steps can be included in the ETL, ensuring that these steps are uniformly applied to all subsequent studies. These steps include the several data quality curation steps mentioned above. Standardization also allows several individuals in an organization to specialize in the ETL of a particular data source while allowing many users to analyze the data without the need to understand all database-specific schema details.

The main benefit of standardization is demonstrated in our replication study. With 1 analytic routine, we were able to execute studies across 6 databases and generate a consistent set of results. Without the CDM, we would have required independent programming of each schema and results may not have been directly comparable due to differences in the source vocabulary. The replication study also demonstrated the considerable insights that could be gained by reviewing results across disparate datasets as we learned what findings were consistent (thereby potentially becoming robust against the different sources of bias that exist within each source). We also observed sources of heterogeneity that stimulated further research to better understand the underlying data to ensure an appropriate interpretation of the findings. The descriptive analysis across databases allowed us to conduct a feasibility assessment to determine if we had sufficient sample size, both within a database as well as across the network, to study the various health outcomes of interest. While these results indicate that we are not yet powered to explore all clinical endpoints at this time, the same standardized analytic routine can and will be applied after the quarterly refresh of each database, and the full protocol-based assessment can be executed as soon as the necessary population is available.

CONCLUSIONS

We have found that the time and resources required to establish a consistent platform using the OMOP CDM has had a substantial return on investment given the enhanced understanding of our observational databases we obtained; the improved quality of the data; and the increased efficiency we obtained in conducting the full portfolio of the observational analyses we supported. We have used the OMOP CDM to

Table 5: Cohort Summary

	Warfarin					Rivaroxaban				
	Optum	CPRD	CCAE	MDCR	MDCD	Optum	CPRD	CCAE	MDCR	MDCD
Demographics										
Total number of persons	3890	9860	12 153	22 026	1514	1797	184	8971	9585	157
Age at index, mean, y	64	74	57	78	62	61	75	56	77	61
Male, %	2637 (67.8%)	5492 (55.7%)	8604 (70.8%)	11608 (52.7%)	746 (49.3%)	1276 (71.0%)	94 (51.1%)	6495 (72.4%)	5272 (55%)	79 (50.3%)
Prevalence of conditions occurring in 90 days prior to cohort entry, %										
Atrial fibrillation	92.3	58.6	91.3	92.3	86.1	94.6	52.2	93.8	93.1%	91.1%
Atrial flutter	17.8	3.6	18.4	14.3	17.5	19.0	6.0	19.7	15.9	15.9%
Atrial fibrillation and flutter		24.9					19.0			
AF, Paroxysmal atrial fibrillation		10.3					14.7			
Acute myocardial infarction	3.3	0.5	3.2	3.3	2.7	1.7		1.1	1.7	1.3
Intermittent cerebral ischemia	5.3	2.5	3.6	5.8	3.6	3.6	4.9	2.5	4.7	5.1%
CVA, Cerebrovascular accident		2.7					9.8			
GI, Gastrointestinal hemorrhage	1.2	0.0	1.3	2.1	1.7	0.5		0.4	1.2	0.6
HF, Heart failure	2.1	2.3	2.5	2.3	4.0	1.3	1.6	1.1	1.4	3.2
Intracranial hemorrhage	0.3	0.0	0.3	0.2	0.5	0.1		0.0	0.1	
Essential hypertension	52.7	1.3	43.9	52.0	59.4	48.1	1.6	40.5	46.6	65.0
Hyperlipidemia	34.0	0.2	27.5	30.5	30.8	34.7	1.1	27.5	29.5	34.4
Type 2 diabetes mellitus	24.2	1.0	22.2	24.8	36.6	18.1		17.7	20.3	42.7
Prevalence of drugs occurring in 90 days prior to cohort entry, %										
ACE inhibitors, plain	33.2	39.5	33.0	33.4	40.4	27.2	40.2	28.3	30.2	41.4
Angiotensin II Antagonists, plain	14.4	16.2	14.2	19.4	10.0	18.3	22.3	16.3	23.1	12.7
Beta blocking agents, selective	49.7	60.5	49.5	51.6	38.5	47.2	60.3	49.8	50.0	42.7
HMG COA reductase inhibitors	43.6	51.1	38.2	50.2	38.4	40.9	60.3	35.3	50.9	43.9
Platelet aggregation inhibitors excl. heparin	11.3	57.9	9.5	14.7	21.5	9.6	56.5	7.5	15.1	22.3
Proton pump inhibitors	19.1	34.8	18.8	21.7	20.1	18.0	44.6	18.4	20.2	29.3
Salicylic acid and derivatives	1.4	52.2	1.7	1.6	11.6	0.7	47.8	1.4	1.2	7.6
Sulfonamides, plain	24.2	28.5	23.3	31.9	44.8	13.9	33.7	14.7	23.7	34.4
Thiazides, plain	17.5	16.7	16.4	19.6	13.6	17.6	15.8	17.4	20.8	20.4

conduct descriptive epidemiology research on the natural history of disease and treatment utilization patterns,[33] medical product safety surveillance,[34] comparative effectiveness,[35] and clinical trial feasibility assessment.[36] We believe the framework followed within our organization can be successful within other organizations with multiple observational data sources and demonstrates the potential for organizations working together as part of a network which can leverage standards in data structure, content, and analytics to support their research activities. In an evaluation of the association of fluoroquinalone exposure and the incidence of retinal detachment,[34] we applied multiple different study designs and analysis variants across 2 databases. The consistency of the findings across sources and methods provided a more comprehensive characterization of the magnitude of association than had been previously described in the literature. In a comparative effectiveness analysis of the relative incidence of abuse between 2 opioids,[35] we used a common analytic routine to generate source-specific estimates in 2 populations and used these results to evaluate database heterogeneity and produce a composite estimate with greater precision. In all of these cases, the ability to explore a potential association across multiple databases has proven

tremendously useful for strengthening our confidence in the clinical results.

ACKNOWLEDGEMENTS

This study is based in part on data from the Clinical Practice Research Datalink obtained under license from the UK Medicines and Healthcare products Regulatory Agency. However, the interpretation and conclusions contained in this study are those of the author/s alone. The protocol for this study (reference No. 14_076A) is provided with the submission and was approved by the Independent Scientific Advisory Committee. Also, we thank Anton Ivonov and Alexey Arestenko for their programming and data management efforts associated with our OMOP CDMs and OMOP Vocabulary. We thank David Longolucco for his technical editing support with this manuscript.

CONTRIBUTORS

All authors agreed to be accountable for the work and made substantial contributions through drafting and revising the work and approving the final work.

COMPETING INTERESTS

All authors are full-time employees of Janssen Research and Development, LLC, a unit of Johnson & Johnson, and the work on this study was part of their employment. They each hold pension rights from the company and own stock and stock options. Rivaroxaban is a marketed product of Janssen.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://jamia.oxfordjournals.org/.

REFERENCES

1. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol.* 2005;58(4):323–337.
2. Brookhart MA, Stürmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. *Med Care.* 2010;48 (6):S114–S120.
3. Madigan D, Stang PE, Berlin JA, *et al*. A systematic statistical approach to evaluating evidence from observational studies. *Ann Rev Stat Appl.* 2014;1(1):11–39.
4. Psaty BM, Furberg CD. COX-2 inhibitors–lessons in drug safety. *N Engl J Med.* 2005;352 (11):1133–1135.
5. Reisinger SJ, Ryan PB, O'Hara DJ, *et al*. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *JAMIA* 2010;17 (6):652–662.
6. National Research Council. *The Future of Drug Safety: Promoting and Protecting the Health of the Public.* Washington, DC: The National Academies Press; 2007.
7. Food and Drug Administration Amendments Act of 2007. US Government Printing Office Website. http://www.gpo.gov/fdsys/pkg/PLAW-110publ85/html/PLAW-110publ85.htm. Accessed June 12, 2014.
8. FDA's Sentinel Initiative. US Food and Drug Administration Website. http://www.fda.gov/safety/FDAsSentinelInitiative/ucm2007250.htm. Accessed June 12, 2014.
9. Observational Medical Outcomes Partnership (OMOP). Observational Medical Outcomes Partnership Website. http://omop.org. Accessed June 12, 2014.
10. Observational Health Data Sciences and Informatics (OHDSI) Website. http://www.ohdsi.org. Accessed June 12, 2014.
11. Coloma PM, Schuemie MJ, Trifiro G, *et al*. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol Drug Safety.* 2011;20(1):1–11.
12. EU-ADR. EU-Adverse Drug Reactions Website. http://www.euadr-project.org. Accessed June 12, 2014.
13. OMOP Common Data Model (CDM). Observational Medical Outcomes Partnership Website. http://omop.org/CDM. Accessed June 13, 2014.
14. OMOP Vocabularies. Observational Medical Outcomes Partnership Website. http://omop.org/Vocabularies. Accessed June 13, 2014.
15. Stang PE, Ryan PB, Racoosin JA, *et al*. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med.* 2010; 153(9):600–606.
16. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *JAMIA.* 2012;19(1):54–60.
17. Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. *Med Care.* 2012;50 Suppl:S60-S67.
18. Madigan D, Ryan PB, Schuemie M, *et al*. Evaluating the impact of database heterogeneity on observational study results. *Am J Epidemiol.* 2013;178(4):645–651.
19. Madigan D, Ryan PB, Schuemie M. Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies. *Ther Adv Drug Safety.* 2013;4(2):53–62.
20. Ogunyemi OI, Meeker D, Kim HE, Ashish N, Farzaneh S, Boxwala A. Identifying appropriate reference data models for comparative effectiveness research (CER) studies based on data from clinical information systems. *Med Care.* 2013; 51(8 Suppl 3):S45-S52.
21. Ryan PB, Madigan D, Stang PE, Overhage JM, Racoosin JA, Hartzema AG. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Stat Med.* 2012;31(30):4401–4415.
22. Ryan PB, Stang PE, Overhage JM, *et al*. A comparison of the empirical performance of methods for a risk identification system. *Drug Safety.* 2013;36(Suppl 1):S143–S158.
23. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. *J Biomed Inform.* 2012;45(4):689–696.

24. *WhiteRabbit* [computer application]. GitHub Website. https://github.com/OHDSI/WhiteRabbit. Accessed June 13, 2014
25. Matcho A. OMOP Common Data Model (CDM, Version 4.0): Clinical Practice Research Datalink *(CPRD) Mapping Specification*. Washington, DC: Reagan-Udall Foundation for the FDA; Year of Publication: 2014. http://75.101.131.161/download/loadfile.php?docname=CPRD%20ETL. Accessed June 23, 2014.
26. Makadia R. *OMOP Common Data Model (CDM, Version 4.0): ETL Mapping Specification Premier*. Washington, DC: Reagan-Udall Foundation for the FDA; Year of Publication: 2014. http://75.101.131.161/download/loadfile.php?docname=Premier%20ETL. Accessed June 23, 2014.
27. Ma Q, Voss E. *Janssen Research & Development, Pharmaceutical Companies of Johnson & Johnson Common Data Model (CDM, Version 4.0) ETL Mapping Specification for Optum* Washington, DC: Reagan-Udall Foundation for the FDA; Year of Publication: 2014. http://75.101.131.161/download/loadfile.php?docname=OPTUM%20ETL. Accessed June 23, 2014.
28. Ma Q, Voss E. *Johnson & Johnson Common Data Model (CDM, Version 4.0) ETL Mapping Specification for TRUVEN (CCAE and MDCR)*. Washington, DC: Reagan-Udall Foundation for the FDA; Year of Publication: 2014. http://75.101.131.161/download/loadfile.php?docname=CCAE%2FMDCR%20SAS%20ETL%20by%20Janssen. Accessed June 23, 2014.
29. Ma Q, Voss E. *Johnson & Johnson Common Data Model (CDM, Version, 4.0) ETL Mapping Specification for TRUVEN (MDCD)*. Washington, DC: Reagan-Udall Foundation for the FDA; Year of Publication: 2014. http://75.101.131.161/download/loadfile.php?docname=MDCD%20SAS%20ETL%20by%20Janssen. Accessed June 23, 2014.
30. Ryan C, Joshua JG, Jennifer N, *et al*; US Food and Drug Administration. Mini-Sentinel Surveillance Plan: Mini-Sentinel prospective routine observational monitoring program tools (prompt): rivaroxaban surveillance. http://www.mini-sentinel.org/work_products/Assessments/Mini-Sentinel_PROMPT_Rivaroxaban-Surveillance-Plan.pdf. Accessed June 23, 2014.
31. Matcho A, Ryan PB, Fife D, Reich C. Fidelity assessment of a clinical practice research datalink conversion to the OMOP Common Data Model. *Drug Safety.* 2014;37(11):945–959.
32. Makadia R, Ryan P. Transforming the Premier Perspective hospital database into the OMOP Common Data Model. *eGEMs (Generating Evidence & Methods to improve patient outcomes).* 2014;2(1)Article 15.
33. Weinstein RB, Schuemie MJ; Observation Medical Outcomes Partnership. Seasonality in acute liver Injury in healthcare claims data. http://omop.org/sites/default/files/34_Weinstein_ALI%20seasonality.pdf. Accessed June 23, 2014.
34. Fife D, Zhu V, Voss E, Levy-Clarke G, Ryan P. Exposure to oral fluoroquinolones and the risk of retinal detachment: retrospective analyses of two large healthcare databases. *Drug Safety.* 2014;37(3):171–182.
35. Cepeda MS, Fife D, Ma Q, Ryan PB. Comparison of the risks of opioid abuse or dependence between tapentadol and oxycodone: results from a cohort study. *J Pain.* 2013;14(10):1227–1241.
36. Knoll CA, DeFalco FJ, Ryan PB; Observation Medical Outcomes Partnership. Applications of the OMOP Common Data Model for clinical trial feasibility assessment. http://omop.org/sites/default/files/04_Knoll_OMOP%20DM%20for%20Clincial%20Trial%20Feasibility%20Assessment.pdf. Accessed June 23, 2014.

AUTHOR AFFILIATION

[1]Epidemiology Analytics, Janssen Research & Development, Titusville, New Jersey, USA

[2]Medical Informatics, Janssen Research & Development, Titusville, New Jersey, USA