11-11-2014

Transforming the Premier Perspective® hospital database to the OMOP Common Data Model

Rupa Makadia
Janssen Research & Development, rmakadia@its.jnj.com

Patrick B. Ryan
Janssen Research & Development, pryan4@its.jnj.com

Recommended Citation

Makadia, Rupa and Ryan, Patrick B. (2014) "Transforming the Premier Perspective® hospital database to the OMOP Common Data Model," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*: Vol. 2: Iss. 1, Article 15.
DOI: http://dx.doi.org/10.13063/2327-9214.1110
Available at: http://repository.academyhealth.org/egems/vol2/iss1/15

The Electronic Data Methods (EDM) Forum is supported by the Agency for Healthcare Research and Quality (AHRQ), Grant 1U18HS022789-01. eGEMs publications do not reflect the official views of AHRQ or the United States Department of Health and Human Services.

Transforming the Premier Perspective® hospital database to the OMOP Common Data Model

Abstract

Background: The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) has been implemented on various claims and electronic health record (EHR) databases, but has not been applied to a hospital transactional database. This study addresses the implementation of the OMOP CDM on the U.S. Premier Hospital database.

Methods: We designed and implemented an extract, transform, load (ETL) process to convert the Premier hospital database into the OMOP CDM. Standard charge codes in Premier were mapped between the OMOP version 4.0 Vocabulary and standard charge descriptions. Visit logic was added to impute the visit dates. We tested the conversion by replicating a published study using the raw and transformed databases. The Premier hospital database was compared to a claims database, in regard to prevalence of disease.

Findings: The data transformed into the CDM resulted in 1% of the data being discarded due to data errors in the raw data. A total of 91.4% of Premier standard charge codes were mapped successfully to a standard vocabulary. The results of the replication study resulted in a similar distribution of patient characteristics. The comparison to the claims data yields notable similarities and differences amongst conditions represented in both databases.

Discussion: The transformation of the Premier database into the OMOP CDM version 4.0 adds value in conducting analyses due to successful mapping of the drugs and procedures. The addition of visit logic gives ordinality to drugs and procedures that wasn't present prior to the transformation. Comparing conditions in Premier against a claims database can provide an understanding about Premier's potential use in pharmacoepidemiology studies that are traditionally conducted via claims databases.

Conclusion/Next steps: The conversion of the Premier database into the OMOP CDM 4.0 was completed successfully. The next steps include refinement of vocabularies and mappings and continual maintenance of the transformed CDM.

Acknowledgements

We would like to acknowledge Chris Knoll for his help with the fuzzy string matching, as well as Amy Matcho, Martijn Schuemie, Paul Stang and Erica Voss and for their review.

Keywords

Informatics, Data Use and Quality, Research Networks

Disciplines

Health Services Research

Transforming the Premier Perspective® Hospital Database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model

Rupa Makadia, MS; Patrick B. Ryan, PhD[i]

Abstract

Background: The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) has been implemented on various claims and electronic health record (EHR) databases, but has not been applied to a hospital transactional database. This study addresses the implementation of the OMOP CDM on the U.S. Premier Hospital database.

Methods: We designed and implemented an extract, transform, load (ETL) process to convert the Premier hospital database into the OMOP CDM. Standard charge codes in Premier were mapped between the OMOP version 4.0 Vocabulary and standard charge descriptions. Visit logic was added to impute the visit dates. We tested the conversion by replicating a published study using the raw and transformed databases. The Premier hospital database was compared to a claims database, in regard to prevalence of disease.

Findings: The data transformed into the CDM resulted in 1% of the data being discarded due to data errors in the raw data. A total of 91.4% of Premier standard charge codes were mapped successfully to a standard vocabulary. The results of the replication study resulted in a similar distribution of patient characteristics. The comparison to the claims data yields notable similarities and differences amongst conditions represented in both databases.

Discussion: The transformation of the Premier database into the OMOP CDM version 4.0 adds value in conducting analyses due to successful mapping of the drugs and procedures. The addition of visit logic gives ordinality to drugs and procedures that wasn't present prior to the transformation. Comparing conditions in Premier against a claims database can provide an understanding about Premier's potential use in pharmacoepidemiology studies that are traditionally conducted via claims databases.

Conclusion and Next Steps: The conversion of the Premier database into the OMOP CDM 4.0 was completed successfully. The next steps include refinement of vocabularies and mappings and continual maintenance of the transformed CDM.

Objectives

The project aim was to assess the feasibility and utility of converting the Premier Perspective database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). We describe some challenges in the transformation of the CDM that are due to the unique data structure. A replication study was conducted to assess the data quality of the transformation, and in-patient conditions from the transformed Premier CDM and another claims database that had been transformed into the CDM were compared in order to highlight their similarities and differences.

Background

Observational data is becoming more widely available in the form of electronic health record (EHR) data, insurance claims data, and hospital data, leading to an increase in observational research in outcomes research and pharmacoepidemiology due to the greater availability of data.[1] These data are collected during the course of health care processes, either for reimbursement or for clinical care. With many retrospective databases available in the market, data structure becomes an important consideration in conducting analyses in multiple databases.[2] A logistical challenge in conducting observational studies is developing sufficient technical expertise to adequately work with the data format. The native data structure of disparate data sources can be quite varied, and some can be inapt to handle intricate analysis due to tangled table structures, missing data, free text, and lack of consistent data definitions.[3,4] OMOP was a public-private partnership managed by the Foundation for the National Institute of Health that conducted methodological research to evaluate and establish scientific best practices in the analysis of observational data.[5] As part of its work, OMOP created an infrastructure to house different types of data (mainly EHR and claims) and established a CDM that can be used to map data into a common format.[2] The original focus of the OMOP CDM was drug

[i]Janssen Research & Development

safety surveillance, but it has since been expanded to accommodate analytic use cases for comparative effectiveness, health economics, and quality of care.[2] The CDM accommodates common definitions for visits, patients, and observations where business rules can be applied consistently throughout the database; and multiple analyses across different databases can be completed without changing programming code.[2]

Mapping data to the OMOP CDM version 4.0 allows researchers to conduct analyses with standard definitions, and with a common data format. Conversion also allows for single code sets to be developed that can be run on all data that use the CDM.[2] Transforming data into a CDM is not without its concerns, mainly the quality of the transformation and potential loss of data. Thus careful evaluation of the transformation process and results are critical to our understanding of the utility of this approach.

In the last few years, many organizations have converted their data into the OMOP CDM. One example is The Health Improvement Network (THIN) database from the United Kingdom, an EHR database where the native data structure had many caveats that were addressed via the conversion to the CDM.[6] The biggest challenge in this conversion was the extensive mapping of medical and drug codes. The study revealed that conversion was not useful for epidemiologic studies due to mapping discrepancies in their data, but with revised and updated mappings the transformation of THIN into the OMOP CDM would be valuable in epidemiological research. Truven MarketScan claims data have been converted successfully to the CDM, and multiple analyses have been performed on the data to validate the transformation.[2] The CDM transformation has been useful to other types of databases including EHR databases. These conversions are documented on the OMOP website, along with the code and implementation for analytic methods that can be executed on data converted to the OMOP CDM.[7] However, the world of health data does not stop at inpatient- and outpatient claims and EHRs.

Not all observational data sources are equally suited to answer all clinical research questions. Each data source carries with it a unique set of limitations that need to be considered when evaluating the utility of the data source for a particular analysis. Administrative claims data capture billing records that can be used to assess longitudinal patterns in disease history and treatment utilization, but often lack the specificity of clinical detail, such as laboratory results, that may be needed for some questions. EHRs often provide rich clinical information, but much of the value is trapped within unstructured free text, and longitudinal capture of out-of-system health encounters is limited. Public health surveys often offer the best nationally representative data, but are commonly limited to only cross-sectional analyses. Hospital databases can provide great depth of activities within an inpatient encounter, which can be useful for monitoring the use of products and services immediately proximal to a specific inpatient procedure, but do not provide sufficient capture of out-of-hospital events to allow for longer-term assessment of medical interventions. Because the strengths of each database support important and complementary use cases, an effective analytics strategy needs to be able to accommodate these disparate sources to support a comprehensive research portfolio. The use of a CDM to standardize the structure and content of these disparate databases is one potential path forward, but there have been limited published examples of where the OMOP CDM has been applied beyond administrative claims and EHR databases. This paper addresses the transformation of the United States hospital billing system database, Premier Perspective, into the OMOP CDM with respect to the quality of the transformation and usability.

The Premier database contains information on hospital visits (both inpatient and outpatient encounters) for a person and records events during the visit—such as drug administration, implantation of a device, or even simply the use of a bandage—as a transaction. This database provides robust and detailed information for a patient's visit in any Premier hospital, information that cannot be captured from a claims database and that is inconsistently captured in EHRs. Claims often do not provide line level detail that occurs in the hospital, and this information can be left to the provider to record in an EHR system. However, this database provides a rich source of information to answer questions related to inpatient administration of drugs and possible outcomes, patients that have multiple surgical procedures and adverse events that could be captured around the visit or subsequent visits.[8] Due to Premier capturing only hospital episodes, longitudinal data before and after hospital episodes is not available. Information about encounters that occurred previous to and following the hospital stay are not captured. Additionally, laboratory test results aren't available, prescription information is as prescribed, and duration and use are unknown after the patient leaves the hospital.

We describe some challenges that are apparent in the transformation of the CDM due to the unique data structure of the Premier database.[8] We conducted a replication study to assess the data quality of the transformation and compared inpatient conditions and the transformed Premier CDM with another claims database that had been transformed into the CDM to highlight the similarities and differences between both and to assess the utility of the CDM.

Materials and Methods

Data Set

The Premier Perspective database contains data from over 467 hospitals—which includes teaching- and nonteaching hospitals—and has over 95 million encounters. These hospitals provide care to a largely urban population, and are believed to be broadly representative of the United States hospital experience.[9] The database is available as hospital discharge files that are date stamped records of all billable items—including therapeutic and diagnostic procedures, medication, and laboratory usage—which are all linked to an encounter.[9]

Data include transactions from both outpatient and inpatient visits. The database is visit oriented and uses a visit key to link all information about a visit together—including cost, diagnosis, and procedures that occurred within the visit. The four main tables contain information on diagnosis, procedures, visits, demographics, and billing. Information about drugs and procedures are captured in the billing table and are identified by a Premier-generated standard billing code.

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)

The OMOP CDM is a patient centric data model that has drug, procedure, and condition information. Additionally, information on providers, payers, care sites, and death can be populated for each person. There are 18 tables in the CDM; more information for each table can be found on the OMOP website, including specifications for each data element.[7]

Transformation of Premier Perspective Data

We transformed all data from the Premier database. Patients were excluded if they had multiple genders recorded, and if the years of birth varied more than two years from the admission date, to adhere to CDM specifications of having unique values for each patient. The data fields in the original Premier database that had the same conceptual meaning as those in the OMOP CDM documentation were transferred in their raw format—such as gender and race. As described in the previous section, coding systems (e.g., diagnosis codes) used in a database are mapped to standard vocabularies in OMOP. In the case of Premier, code mapping tables for values of International Conference for the Ninth Revision of the International Classification of Diseases (ICD-9) codes, and procedural codes in Common Procedural Terminology (CPT) or Healthcare Common Procedural Coding System (HCPCS) were mapped to their appropriate vocabulary. Additional detailed information about the tables and fields used can be found in the extract, transform, and load (ETL) document posted on the OMOP website (see additional online supplement).[10] One of the biggest transformational changes that occurred in the conversion was around the implementation of visit logic for each patient.

Dates of Service

Premier provides admission and discharge dates, in month and year format, for each visit. It is difficult to conduct analyses across the period of a stay in order to establish the ordinality of procedures or drugs that were administered during the stay. Premier includes the service day of each record in the billing table, and includes a variable that designates the order of the visits. An algorithm was developed to impute a start date and end date for each visit by calculating the maximum number of service days from the billing records. The maximum numbers of service days was used instead of the length of stay variable, which is not captured for outpatient visits. The logic assumes that the first visit for every patient begins on the first day of the month. Each subsequent visit begins on the previous visit's discharge date plus one. Patients who have visits that extend to the end of the month and year will be assigned an admission- and discharge date as the last day of the month.

Another important part of the algorithm is for those patients who are admitted one month and discharged in the subsequent month. Here, the discharge date is the first of the month, and the admission date is calculated by counting backwards from the discharge. The intention of the logic is not to re-identify the actual dates, but instead to preserve the logical temporal sequence of events using a date format that is amenable to standardized research queries.

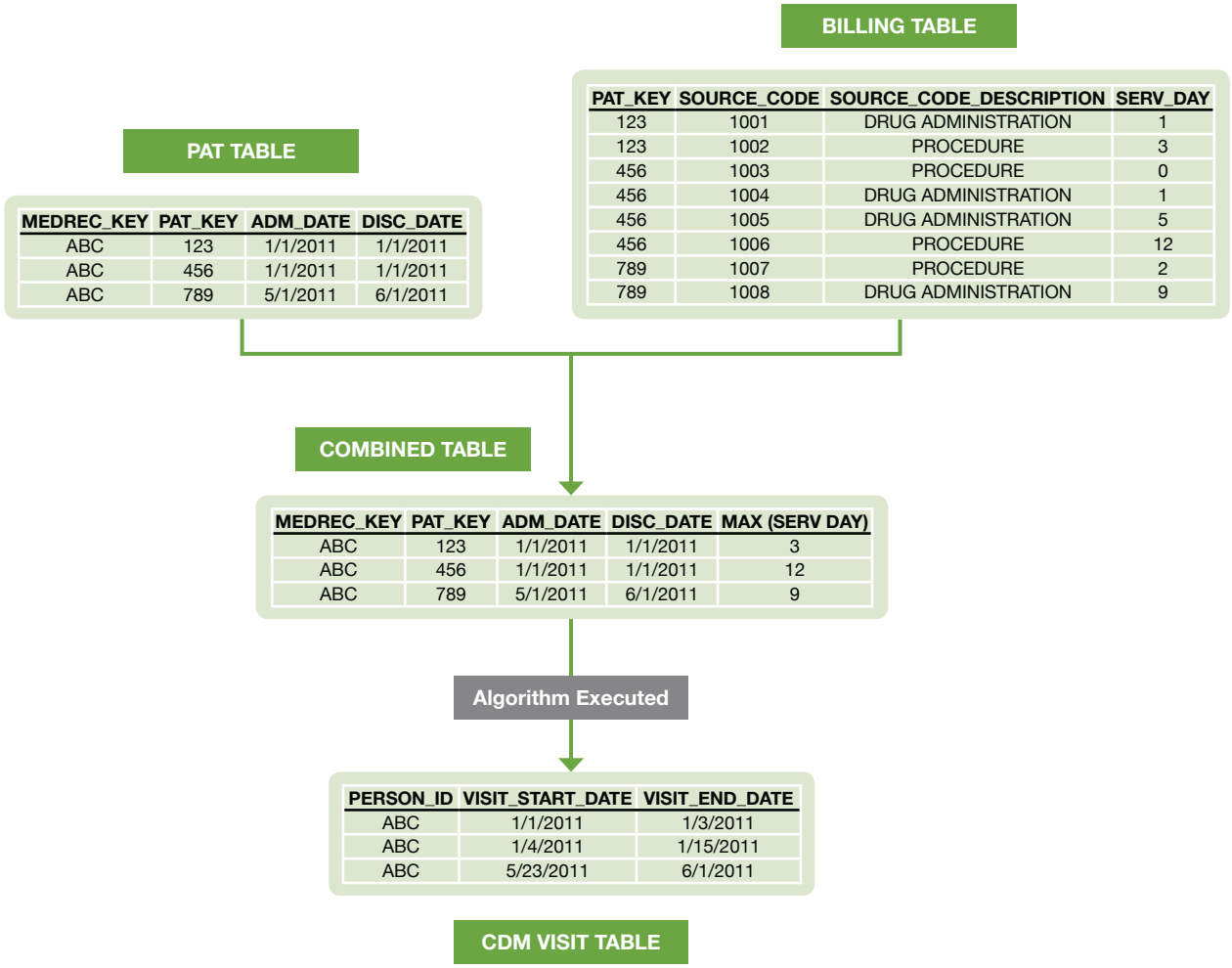
The visit table was created first, and then used to create the entries for drug exposures and procedures. The following figure describes how the visit logic is implemented for a sample patient (Figure 1). The admission date and discharge date for each patient are obtained from the patient table, and linked with the billing table by visit ID. The maximum number of service days for each visit is captured. The admission date, discharge date, and maximum number of service days for each visit are aggregated into a table. The algorithm is then applied to the aggregated table to create the start and end dates for each visit.

The date information that is derived from the visit table drives the date information in the ETL process. The drug start dates and procedure start dates are determined from the service day that the event occurred, which falls within the visit start and end dates. The observation periods are determined from the transformed visit information. If visits are within 30 days of one another they create a singular observation period. For full details on the visit implementation logic, see the additional online supplement.[10]

Standardizing Billing Codes

Another important component of the transformation involved the mapping of the billing table into standard concepts. The data were categorized within the billing tables according to hospital department. For the purpose of creating the algorithm CDM, the department header was used to separate the drug administrations that would be added to drug tables from other billable items that would be added to the procedure table. Each billed item had an associated internal Premier standard charge code and a charge code description. The description for each standard charge code was a free-text field, which was then mapped into a standard vocabulary concept by using a fuzzy string matching algorithm. Two input data sets were created for drugs and procedures. One data set contained the raw text descriptions from Premier and their associated standard charge codes. The other data set included the appropriate OMOP vocabulary (RxNorm for drugs, SNOMED for procedures). The fuzzy string matching program ignored common words such as "tablet" or "liquid' in both data sets. An abbreviations file was also created to expand words in the Premier database to their long form—such as "VL," which stands for "Vial," or "NaCl," which stands for "sodium chloride"— in order to successfully match more terms. The fuzzy string program tokenized each of the words in both input data sets. Next, the algorithm ignores common words, and translated the abbreviations into their full forms.

Figure 1. Example of Implementation of Visit Logic in CDM Transformation for a Sample Patient



The words on both sides were compared by the number of letters and order to determine if they were the same or similar. A score was assigned based on the number of matched components. A score ranged from 0 to 1,000—with 1,000 being an exact match, and a score of 500 meaning that only 50 percent of the token matched. Through a scan of the data, a determination was made that any score lower than 400 (40 percent of terms matched) needed to be reviewed manually for validity or assigned a concept of 0 (no matching concept). Codes were ordered by the most frequent occurrences in the data, and the top 1,000 codes were reviewed for accuracy. The drug codes that did not map accurately due to incorrect ingredients were then mapped to zero (no matching concept).

Assessment of Transformation

To assess the transformation, a replication study of a published observational study that utilizes the Premier database was conducted. The study was replicated in the raw Premier data schema and the transformed data. The second study demonstrated the utility of the CDM across the two transformed CDMs. The study highlighted the breadth of conditions that are represented within Premier and the Optum Clinformatics (Optum) commercial claims database in order to gain insights about the conditions represented and the potential use cases for the transformed CDM.

Replication study. We evaluated the transformation by completing a replication study of a prior Premier study in both the raw data schema and the transformed CDM. The study "Estimating pediatric inpatient medication use in the United States" by Lasky et al. used the Premier database to develop national estimates of the pediatric inpatient medication use.[11]

The paper compares the Premier database to national inpatient estimates in regards to demographics and inpatient drug use.[11] Patients were selected if their age at the time of their admission was less than or equal to the age of 18 in an inpatient setting. Patients were selected if they had an inpatient admission in 2008. Demographic data were collected for gender, age, payer information, length of stay, admission source, discharge status, bed size, and hospital location. The top 10 drugs that were reported from the Lasky paper were string searched in the billing table within the raw data schema and were searched using the vocabulary from the transformed CDM.[11] The proportions were calculated by taking the number of visits that utilized each drug ingredient over the total number of visits per 100 patients.

Comparing Premier to claims data. To assess the utility of the transformation, we evaluated the diseases with inpatients in

Premier, compared them with inpatient stays in Optum. The Optum database is a commercial claims database that contains data about over 36.2 million privately insured patients, with data from October 1, 2005 to December 31, 2012 in the United States. Enrollment, utilization, pharmacy records, and inpatient data are collected from the standard UB-92 claim form. In this analysis the OMOP CDM version of the Optum database was used. Specific transformation details about Optum can be found in its ETL document.[12] When the Premier database was transformed into the CDM, conducting the same analysis in both databases required only one set of codes.

Premier and Optum databases have conditions coded in ICD-9 diagnosis codes, and in following the OMOP CDM specifications, the ICD-9 codes are mapped to Systematized Nomenclature of Medicine—Clinical Terms (SNOMED) concepts.[7] Inclusion criteria included patients that had a valid inpatient visit in 2011 and enrolled in a commercial plan. All SNOMED concepts associated to those visits were captured. In addition, the patient characteristics for each patient were collected: gender, and the age at the time of admission.

Because the condition concepts are represented as SNOMED concepts, the granular nature of the concepts makes it difficult to make an effective comparison. Grouping terms to higher level terms was necessary. The OMOP vocabulary provides a map from SNOMED concepts to Medical Dictionary for Regulatory Activities (MedDRA) concepts, thus each SNOMED concept was mapped into a MedDRA preferred term.[3] Each MedDRA term was represented as a proportion of visits per 1,000 patients within each term over the total number of visits, stratified by age deciles and gender. MedDRA concepts that had at least one occurrence of a condition in both databases were kept. Pregnancy codes were excluded from the analysis due to differences in coding practices. A scatter plot of the proportions in CDM Optum was plotted on the x-axis, and the CDM Premier proportions were plotted on the y-axis on a log-log scale. The scatter plots were stratified by age deciles, and each gender was plotted on the same graph.

Findings

Mapping Premier into CDM Premier

The total number of patients that were transformed into the OMOP CDM was 88,202,612; the number of patients in the Premier database was 88,892,294. A 0.78 percent reduction of the data was due to exclusion criteria that were applied in Premier: all persons must have only one unique gender; and they could not have two year of birth records with values greater than two years apart.

There were 55,470 unique standard charge codes in the data, and of those 44,346 (which represent 91.4 percent of the data) could be mapped to a concept in the OMOP vocabulary. Of the remaining 11,124 codes, they were representative of Premier's administrative and billing codes and could not be logically mapped to an appropriate vocabulary; thus, they were mapped to zero. (Table 1A.)

Of the 44,346 codes, 6,647 codes were mapped to zero or no matching concept, which represent about 12.7 percent of the data. Manual review of the top 1,000 codes, which account for 72.9 percent of the total data, were assessed for accuracy. The total number of codes that were mapped correctly was 838 codes, and 28 codes needed to be remapped to an appropriate concept (Table 1B).

Table 1A. Statistics for the Standard Charge Code Mapping to Concept Domains

Statistic Value	Number	Percent of Codes	Number of Bill Records	Percent of Data
Total number of codes	55,470	100.00%	9,604,632,594	100.00%
Total number of codes (mappable)	44,346	79.90%	8,781,445,153	91.40%
Unmapped administrative codes (mapped to zero)	11,124	20.05%	823,187,441	8.60%

Table 1B. Statistics for the Manual Review of Standard Charge Code

Statistic	Number	Percentage
Codes mapped correctly	838	83.80%
Codes initially mapped incorrectly and remapped upon manual review	28	2.80%
Unmapped codes (mapped to zero)	134	13.40%
Total codes reviewed	1,000	100.00%

Replication Study Results

Demographic summaries were calculated for the raw data and the CDM data. The distribution for males among raw Premier is 49.9 percent compared to 50.2 percent in the transformed CDM (Table 2). The discharge status among patients who were classified as routine was 65.0 percent in the raw data compared with 65.6 percent in the transformed CDM (Table 2). The demographic distributions between the raw data and transformed CDM for payer, hospital status, location, and bed size had very similar results.

The 10 drugs that were observed for patients in the study were acetaminophen, albuterol, ampicillin, ceftriaxone, fentanyl, gentamicin, ibuprofen, lidocaine, morphine, and ondansetron. Obtaining drug information from the raw Premier data produced a rate of 15.0 per 100 patients as compared to 19.6 per 100 patients in the transformed CDM for acetaminophen (Table 3). For albuterol, ampicillin, fentanyl, ibuprofen, lidocaine, morphine, and ondansetron, the CDM rate was lower than the raw data except for the case of acetaminophen. The rate stayed the same for gentamicin and ceftriaxone between the raw data and the transformed CDM. The use of the vocabulary caused the rate per 100 patients to increase due to the inclusion of more standard charge codes (Table 3).

Table 2. Demographic Information from Replication Study[11]

Demographic	Study Results 2008	Premier 2008	CDM Premier 2008
Sex			
Male	50.9%	49.9%	50.2%
Source of admission			
Routine including births and other sources	72.0%	65.0%	65.6%
Other hospital or health care facility	13.7%	19.1%	18.6%
Emergency department	14.8%	15.9%	15.7%
Discharge status			
Routine including births and other sources	94.0%	97.4%	97.4%
Died	0.4%	0.3%	0.3%
Other	5.6%	2.2%	2.2%
Payer			
Medicare/Medicaid/other government payer	45.9%	46.1%	45.8%
Private insurance	46.2%	51.9%	52.0%
Other	7.9%	2.1%	2.4%
Mean length of stay			
Days	3.7	4.0	3.8
Region			
Midwest	18.7%	19.7%	19.7%
Northeast	14.3%	18.4%	18.4%
South	48.7%	43.3%	43.3%
West	18.3%	18.6%	18.7%
Teaching status			
Teaching hospital	41.5%	44.6%	42.9%
Urban versus rural			
Urban	89.2%	90.5%	90.7%
Bed size			
Small	10.3%	13.9%	16.7%
Medium	17.9%	16.5%	15.70%
Large	71.8%	69.6%	67.6%

Table 3. Inpatient Drug Use in the Pediatric Population in 2008 from Lasky et al., Native Database and Transformed Database

Drug Name	Study Results per 100 Patients[11]	Premier 2008 per 100 Patients	CDM Premier 2008 per 100 Patients
Acetaminophen	14.7	15.0	19.6
Albuterol	5.1	6.3	5.1
Ampicillin	8	9	8.3
Ceftriaxone	5.6	5.7	5.7
Fentanyl	6.6	8.2	7.6
Gentamicin	6.6	6.8	6.8
Ibuprofen	6.3	7.5	7.1
Lidocaine	11	15	14.7
Morphine	6.2	7.3	7.1
Ondansetron	6.2	7.1	6.9

Claims Database Comparison

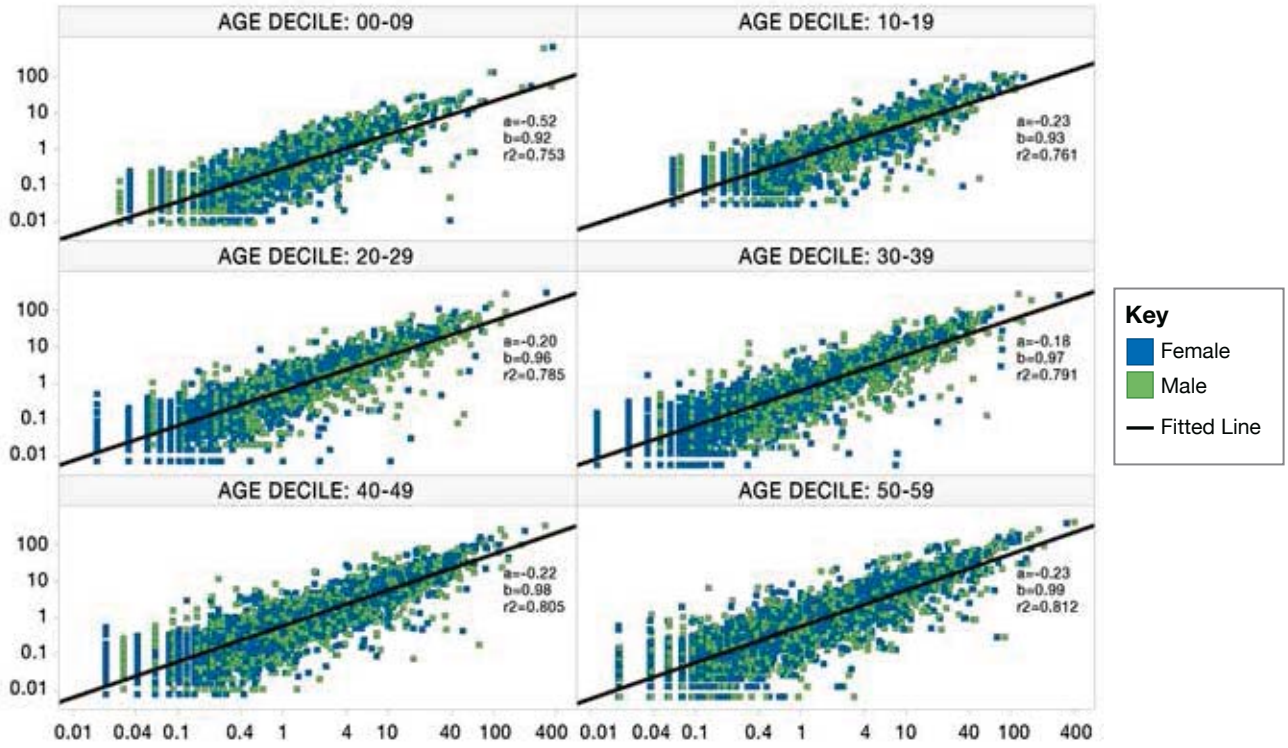
The proportions for inpatient conditions in the commercial population during 2011 from CDM Premier and CDM Optum—stratified by age and gender—are displayed in Figure 2. The graphical representation shows the distribution of the proportions of each MedDRA concept for each database. Pregnancy codes were excluded from the analysis. The goal of comparing inpatient visits within both databases is to give an overall perspective of the similarities and differences between the conditions represented in both databases. The correlation coefficient (r-squared) is calculated for each age decile, and as age increases the r-squared values decrease (Figure 2).

Many observations have similar proportions between the two databases; there are some MedDRA terms in which the proportions differ greatly between databases. A snapshot of the conditions that have the greatest absolute difference and a list of selected chronic conditions are displayed in Table 4 for the age decile 30–39. The

Table 4. Proportions of Selected Conditions per 1,000 Patients Ages 30–39 in CDM Optum and CDM Premier for Patients with Commercial Plans by Gender

MedDRA Term	Proportion of Patients in Optum	Proportion of Patients in Premier	Absolute Difference	Gender
Codes with the highest absolute difference				
Hepatitis immunization	104.84	0.05	104.79	Male
Abdominal pain	128.87	43.96	84.91	Male
Procedural pain	80.98	3.01	77.97	Female
Caesarean section	83.20	7.10	76.09	Female
Perineal laceration	286.93	211.67	75.26	Female
Distribution of selected codes				
Coronary artery bypass	2.05	2.20	0.15	Male
Coronary artery bypass	0.24	0.65	0.40	Female
Crohn's disease	16.10	10.36	5.74	Male
Crohn's disease	5.58	5.85	0.27	Female
Heart transplant	0.64	0.27	0.37	Male
Heart transplant	0.13	0.04	0.08	Female
Hyperlipidemia	7.49	21.81	14.32	Female
Hyperlipidemia	51.07	72.63	21.56	Male
Kidney infection	0.23	0.03	0.20	Male
Kidney infection	0.12	0.03	0.09	Female
Liver disorder	28.05	15.87	12.17	Male
Liver disorder	8.90	8.76	0.14	Female

Figure 2. Proportions of Patients (per 1,000) from CDM Optum and CDM Premier A



Note: Inpatient conditions for commercially insured patients in 2011 are plotted on a log-log scale. The x-axis represents proportions in CDM Premier; and the y-axis represents proportions in CDM Optum. Each box represents a 10-year age decile and is stratified by gender (pink=female, blue=male). The outliers in the 0–9 age decile are due to single birth events that occur in Premier, while the Optum information contains additional inpatient encounters for the 0–9 age decile.

conditions with the greatest absolute difference among patients ages 30–39 are hepatitis immunization, abdominal pain, procedural pain, caesarean section, and perineal laceration. A selected number of common conditions are included: heart transplant, kidney infection, Crohn's disease, liver disease, coronary heart bypass, and hyperlipidemia. Hyperlipidemia in males ages 30–39 occurs more frequently in CDM Premier than CDM Optum, and the highest absolute difference is at 21.56.

Discussion

Transforming the Premier Perspective database has many advantages in comparison to the raw data format. These include data standardizing, adding logic to visits to make subsequent analysis more convenient, and mapping data that could only be utilized by string manipulations. Data communities are increasing as observational data become more popular, and adding the Premier Perspective database brings a unique data resource that is engaged around hospitalizations. More importantly, the transformation has enabled researchers to look into granular details around hospital procedures regarding devices and adverse events during inpatient encounters. The OMOP has shown that implementing a CDM can help standardize data. This data model has been able to capture all relevant information without losing critical information from the source. In total, only about 0.78 percent of data were lost due to raw data discrepancies and not due to the transformation itself.

Overall the ETL transformation required a time investment up front to understand the data and conduct the transformation. The vocabulary mapping is a combination of using automated-processes and manual review to create the standard charge code mapping. The manual effort invested for the mapping will hold for all subsequent data refreshes, the only review that will be necessary will be for additional codes that have been added to the raw data schema. While initial ETL development took considerable resources, subsequent refreshes of the CDM are able to occur in a streamlined fashion with quarterly data updates generally taking less than one week of analyst- and computing time to be able to release the updated CDM version.

The replication of the demographics table from the Lasky et al. study is in concordance with the published results. Although there may be some differences in the versions of Premier used between the published study and our work, the overall rates are similar for both the raw data schema and the transformed CDM, which provides evidence that the transformation was completed without major data anomalies. The Lasky et al. study compares Premier to national inpatient estimates but does not provide additional insight into the rates of outcomes. The data for 0.78 percent of the population were removed in the CDM, which accounts for the slight discrepancies within the raw data and transformed data. Traditional claims and EHR systems create a story with the information that is presented by following a patient through time. This chronological aspect was missing in the Premier data, and the addition of the visit imputation logic was necessary to create a complete picture for a patient.

Premier can be used for assessment of quality of care in hospitals, particularly in instances where quality metrics rely only on information contained within the inpatient encounter. There have been many published studies using Premier database that look to capture the relationship between an event such as a procedure or condition and any potential adverse events that occur during the hospital study. A recent paper published by Oderda et al. explores the effect of opioid-related adverse events in patients who have surgery.[13] The objective was to look for patients in Premier with a selected list of procedures that are known to have postoperative pain, opioid use, and adverse events associated with the use of the drug. The length of stay (LOS), costs, and readmission rates were recorded between groups that experienced an adverse event and those that did not.[13] Rothberg et al. studied chronic obstructive pulmonary disease (COPD) patients and treatment failures with and without antibiotic use during hospitalizations.[14] Treatment failures are defined by the initiation of mechanical ventilation on the second hospital stay.[14]

Hospital data offer the opportunity for complementary data to address different research questions of interest. For example, studies that address acute clinical events require detailed information about services rendered in the hospital that cannot be easily studied in a private-payer claims database due to having only summarized encounter data available. The studies cited above offer utility for using hospital data. The addition of using standard vocabularies and definitions can provide consistency in defining patients and outcome definitions for these types of studies. By transforming the data into the OMOP CDM, the data become standardized and provide the ability to use standard tools that are available within OMOP.

One of the biggest drawbacks in using the data in its native format is the mappings to standard vocabularies that need to occur in order to conduct any type of study. Prior to this study any information about drugs administered in the hospital was subject to string searches in the billing table. We were able to map 79 percent of codes to a valid concept in the OMOP Standard Vocabulary, which represents over 90 percent of the total data. The manual review of the top 1,000 codes that occur the most frequently in the data resulted in some concepts being mapped to zero, some concepts requiring mapping by hand, and the remainder being mapped correctly. An important inference is that 866 codes can be mapped to a correct concept, with only 134 mapping to a concept of zero. The review did eliminate the misclassification that could have occurred from the mapping algorithm for these 1,000 codes. The manual remapping effort for the standard charge codes is a one-time effort, and the changes are incorporated each time the data are updated. In the future, more codes can be reviewed to improve the accuracy of the mapping of standard charge codes to the OMOP Standard Vocabulary.

Drugs identified in the study were searched using string searches for the raw data, and for the CDM, using the standardized vocabularies. The discrepancies in the rates between the raw data and the CDM are due to the strategies used to identify drug use.

Acetaminophen is often abbreviated as "APAP" or included in medications as an ingredient—but that is not always reflected in the name. Thus, the rate of acetaminophen was higher using the standardized vocabularies. Albuterol was lower than in the transformed data as compared to the raw data schema due to the fact that string searches were utilized in the raw data and the drug levalbuterol contains the drug name albuterol. Thus using the native format would not capture all possible drug encounters. Therefore, use of the standardized vocabularies may provide a better and more accurate route to identify drugs within the data. As further exploration, the mapping algorithm can be developed more to increase the mapping quality and quantity of standard charge codes to a standard vocabulary.

Traditionally Premier has been used for studies that involve health outcomes, comparative effectiveness, and hospital cost. Often researchers can look past the use of Premier for many studies because the populations that are represented in Premier are not similar to claims or the data are difficult to use for these studies. The comparison between CDM Optum and CDM Premier shows us that the data represented overall between both databases are similar for some conditions but have some inherent differences as well. The r-squared values for each age decile are above .70 except for the 0 to 9 age group. A possible explanation for this is that in Premier most of the encounters represented for this group at the birth event—care before and after—is conducted elsewhere, thus they are not captured in the database, contributing to the disproportions for the conditions represented in this age group.

Despite the fact that Premier does not have longitudinal data for each patient, the granularity of information captured during each visit can be greater than a traditional claims record. Conditions that are common in a hospital setting, such as coronary artery bypass, are very similar among both databases. The reasoning behind this is that more complex procedures are one-time visits in a hospital, and thus should be similarly distributed in both databases. The transformation of Premier into the CDM does provide a platform to conduct analysis between databases as such to assess feasibility for research in the database, but more importantly it adds standardization to the data. Prior to the transformation, this type of comparison would be very difficult to conduct.

The Premier Perspective data source has shown some challenges in the transformation process, but the inclusion of this data resource within the data community can be highly informative. Because Premier includes privately and publicly insured patients and uninsured patients, it may be more representative of hospital care in the United States than is a private-payer claims database. The transformed data can aid in the missions of many data networks by providing hospital data that can reflect the granularity of adverse events in hospitals and the use of devices.

A drawback to hospital data is that longitudinal data capture is often insufficient, because health care encounters outside of the hospital network are not captured or linked. The future of data and data collaboration relies on the ability to access linked data. Premier has partnered with Optum to create data that link claims data to the hospital data that are provided. In order to use the linked data effectively, transformation of the raw data into the common data format is necessary. Without the conversion data, the conveniences of having one table with all patient demographics, or drug usage in one table, are not available to conduct studies.

Each data source has its intrinsic qualities and applications, but by converting into a common format, the definitions can be applied consistently and results can be interpreted correctly. The OMOP CDM has provided a platform to take data sources and convert them into a useful research tool. As future versions of the OMOP CDM are released, the work that was conducted for this version can be adapted and enhanced to accommodate other versions. This is the stepping stone to creating a successful data network to use and encourage better patient care.

Conclusion

This work demonstrates that the Premier Perspective database can be converted to the OMOP CDM. While the conversion can involve some loss of data, it results in selection of data deemed to be of sufficient quality for research purposes, as we demonstrate. For our analysis, and the use cases involved, we do not consider the extent of information loss to be an impediment for routine use of the Premier OMOP CDM. The data have become more standardized through the adding of the appropriate vocabularies to the billing data. Comparing the transformed CDM of a hospital database to claims CDM demonstrates that standardizing data shows us similarities and differences that were not evident prior to the transformation.

Acknowledgements

We would like to acknowledge Chris Knoll for his help with the fuzzy string matching, as well as Amy Matcho, Martijn Schuemie, Paul Stang and Erica Voss and for their review.

References

1. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. The New England journal of medicine. 2000 Jun 22;342(25):1878-86. PubMed PMID: 10861324. Epub 2000/06/22. eng.
2. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. Journal of the American Medical Informatics Association : JAMIA. 2012 Jan-Feb;19(1):54-60. PubMed PMID: 22037893. Pubmed Central PMCID: PMC3240764. Epub 2011/11/01. eng.
3. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. Journal of biomedical informatics. 2012 Aug;45(4):689-96. PubMed PMID: 22683994. Epub 2012/06/12. eng.
4. Kim W, Choi B-J, Hong E-K, Kim S-K, Lee D. A Taxonomy of Dirty Data. Data Mining and Knowledge Discovery. 2003 2003/01/01;7(1):81-99. English.

5. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Annals of internal medicine. 2010 Nov 2;153(9):600-6. PubMed PMID: 21041580. Epub 2010/11/03. eng.
6. Zhou X, Murugesan S, Bhullar H, Liu Q, Cai B, Wentworth C, et al. An evaluation of the THIN database in the OMOP Common Data Model for active drug safety surveillance. Drug safety : an international journal of medical toxicology and drug experience. 2013 Feb;36(2):119-34. PubMed PMID: 23329543. Epub 2013/01/19. eng.
7. Foundation R-U. Observational Medical Outcomes Partnership [cited 2013]. Available from: http://omop.org/.
8. Premier Research Services. Premier Perspective Database. Charlotte, N.C.2013.
9. Lindenauer PK, Pekow P, Wang K, Mamidi DK, Gutierrez B, Benjamin EM. Perioperative beta-blocker therapy and mortality after major noncardiac surgery. The New England journal of medicine. 2005 Jul 28;353(4):349-61. PubMed PMID: 16049209. Epub 2005/07/29. eng.
10. Makadia R. OMOP Common Data Model (CDM V4.0) ETL Mapping Specification Premier. 2012.
11. Lasky T, Ernst FR, Greenspan J, Wang S, Gonzalez L. Estimating pediatric inpatient medication use in the United States. Pharmacoepidemiology and drug safety. 2011 Jan;20(1):76-82. PubMed PMID: 21182155. Epub 2010/12/25. eng.
12. Voss E. MQ. OMOP Common Data Model (CDM V4.0) ETL Mapping Specification Optum Clinformatics 2012.
13. Oderda GM, Gan TJ, Johnson BH, Robinson SB. Effect of opioid-related adverse events on outcomes in selected surgical patients. Journal of pain & palliative care pharmacotherapy. 2013 Mar;27(1):62-70. PubMed PMID: 23302094. Epub 2013/01/11. eng.
14. Rothberg MB, Pekow PS, Lahti M, Brody O, Skiest DJ, Lindenauer PK. Antibiotic therapy and treatment failure in patients hospitalized for acute exacerbations of chronic obstructive pulmonary disease. JAMA : the journal of the American Medical Association. 2010 May 26;303(20):2035-42. PubMed PMID: 20501925. Epub 2010/05/27. eng.