the lowest proportion males (54%). Mean age was lowest in the hallucinogen (24 years) and characteristic, and lowest was 18. Over half the amphetamine cohort resided in the western and southwestern US. Major psychiatric comorbidity was present for 44% in the sedative cohort, compared to 15% in the alcohol cohort. Median total post-index costs were lowest for the cannabinoids (39) and highest was 18. Over half the sedative cohort size ($16,628). CONCLUSIONS: Patients with claims carrying diagnoses indicative of AUD/SUD are a highly heterogeneous group in terms of drug use, demographic and psychiatric comorbidities and health care utilization. Measures combining AUD and SUD patients may be more reflective of alcohol than other substances, because AUD patients predominate.

PRM5 EMPIRICAL INVESTIGATION OF TECHNIQUES FOR HANDLING MISSING COST-TO-CHARGE RATIO (CCR) IN NATIONWIDE INPATIENT SAMPLE (NIS) Kay et al. 1,2
1Novartis Pharmaceuticals Corporation, East Hanover, NJ, USA, 2DKM Consulting Inc., Forham Park, NJ, USA

OBJECTIVES: Researchers use the Healthcare Cost and Utilization Project (HCUP) Cost-Related National Disease Policy Studies to create a data set that contains all charges in the NIS sample ($11,118 for 2010-2011). We investigated the impact on cost estimates using four techniques for handling missing CCRs. METHODS: This is a cross-sectional study using nationally representative hospitalization data in the US in 2011. We investigated four techniques including: 1) complete case analysis, 2(a) using a reweighting technique recommended by HCUP, 2(b) reweighting by adjustment cells, and 3) using hotdeck imputation by adjustment cells. We used five disease cohorts to represent different sample sizes: we created a disease cohort data set with non-missing CCRs as a benchmark complete data set. We picked one state at a time and set their CCRs to missing to create a data set with missing CCR data. We compared the CCR-adjusted cost estimates (mean and total) from both complete data sets. To evaluate the quality of an estimator, we assessed bias, variance, and the mean squared error (MSE) of an estimator. RESULTS: The range of missing rates was 4% to 40%. The results showed that using both the benchmark data set with non-missing CCRs and the benchmark data set with missing CCRs is the best estimation technique. The complete case analysis underestimated total costs and became worse as missing rates increased. CONCLUSIONS: In NIS data, to handle missing CCRs, when missing data is <4%, researchers will obtain similar mean cost estimates using complete case analysis, the reweighting technique recommended by HCUP; or reweighting by adjustment cells. However, if total cost is the estimate of interest, researchers should adopt either reweighting techniques in order to avoid underestimation.

PRM6 DEVELOPMENT OF A SPATIALLY-ENABLED PUBLIC-USE DATABASE FOR END-STAGE RENAL DISEASE POLICY STUDIES Stephens M1, Maione-Dowling B1, Brotherton SA2, Giffin MD2
1Prima Health Analytics, Weymouth, MA, USA, 2InnoPeritus, Geneva, Switzerland

OBJECTIVES: While patients with end-stage renal disease (ESRD) represent only 1 percent of Medicare beneficiaries, this population accounts for 8 percent of total Medicare spending. The disproportionate cost of treating ESRD has resulted in a high level of monitoring and policy focus. A number of public databases currently exist for studying ESRD in the US, but these public databases are not integrated for policy analysis of this program. Objective was to develop an integrated ESRD "data warehouse" with GIS capabilities from publicly-available dialysis patient and provider data. METHODS: Publicly available population statistics were linked to a geocoding geocoder and linked to spatial data files and general population data obtained from Census Bureau websites. Validation and outlier handling rules were created for mast data and were applied either during the build/update process or at the point of analysis. Longitudinal datasets were created for trend analysis. Data are updated quarterly to keep the database current. RESULTS: The ESRD provider database contains over 6000 cost, utilization, quality, demographic and geographic variables, covering 100% of current Medicare ESRD providers (N=6288 facilities). Approximately 9% are hospital-based units and 91% are free-standing facilities. Depending on data source, the database covers between 6 and 14 calendar years (2003-14). Quality of the data varies, but for most applications, outlier and missing observations are less than 10%. CONCLUSIONS: Public datasets of ESRD-related information can be integrated with general population data and a GIS to support high-quality and cost-effective economic and clinical policy studies that would not otherwise be feasible.

PRM7 PERFORMANCE OF NLP TOOL DESIGNED TO IDENTIFY AND EXTRACT BIOLOGIC DRUG INFUSION DATA FROM CLINICAL NOTES Leng J1, Lu CC1, Cannon G2, Teng CC1, Zhou X1, He T1, Harrison DJ1, Shah N3, Sauer BC1
1Departments of Internal Medicine, University of Utah, Salt Lake City, UT, USA, 2VA Salt Lake City Health Care System, Salt Lake City, UT, USA, 3Amgen, Inc, Thousand Oaks, CA, USA

OBJECTIVES: Infusions of outpatient medications including biologic Disease Modifying Anti-Rheumatic Drugs (DMARDs) administered at Veterans Health Administration sites are well established in the electronic medical record but data are not consistently entered into the pharmacy dispensing or nurse administration structured data sources. Although CPT codes can be used to identify major infusion events but inconsistent coding does not allow estimation of the administered dose. To address this, we developed Natural Language Processing (NLP) software to identify potential infusion notes. We used the NLP software to extract documented dosage information, and standardized results. METHODS: Trained reviewers compared the NLP extractions to source documents and judged whether the software correctly extracted and standardized data. The software contains a display window allowing reviewers to directly assess the NLP extraction. NLP was run on all notes in the VA Health System. The CCI adjusted cost was$10,789 - 35% and warfarin $19,964 - 65% patients matched comparably at 8,093 and 3,897 patients, respectively. The overall accuracy rate was 94.0% [95%CI: 85.5-100%] 33,077 notes were correctly extracted, 1,950 failed to extract infusion data and 163 contained an incorrect extraction. The range of the lower bound of 95%CI for the 255 titles was 78.8-100%. 247 (96.5%) titles had a lower bound >80%. CONCLUSIONS: The NLP software demonstrated acceptable extraction but requires refinement before use for approximately 97% of note titles, suggesting that clinical notes are a reliable data source to identify biologic DMARD infusion data when coding is inconsistent.

PRM8 BUILDING A BRIDGE: ICD-9-CM TO ICD-10-CM MAPPING CHALLENGES AND SOLUTIONS Sears J1, Yong J1, Yu S1, Zarcosky V1
1Optum, Eden Prairie, MN, USA, 2Takeda Pharmaceuticals International Inc., Deerfield, IL, USA

OBJECTIVES: The mandated implementation of ICD-10-CM (International Classification of Diseases, Tenth Revision, Clinical Modification) codes in 2014 challenges health care systems to accurately map existing ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) codes to these new codes, in order to preserve consistency and validity of medical codes. The usage of central authorization Centers for Medicare & Medicaid Services (CMS) General Equivalence Mappings (GEMs), is hindered by their limited definitions of medical conditions. Additional mappings for conditions could be complicated. Instead of a simple one-to-one mapping, a one-to-many or a many-to-one mapping might be needed. A systematic approach is needed to overcome the limitations and complexities. METHODS: Nineteen Charlson Comorbidity Index (CCI) medical conditions plus 21 other conditions were chosen for mapping. CMS’ 2013 GEMs were used. When no GEMs existed, other public and proprietary mapping systems were utilized. When necessary, expertise in coding and clinical knowledge was leveraged to improve the mapping. RESULTS: The 1646 unique ICD-9-CM codes required to fully define the 40 medical conditions, mapped to 3742 unique ICD-10-CM codes. A total of 6355 code-to-code mappings were obtained. 1135 (17.9%) had a one-to-one match, 3636 (57.2%) had a many-to-one match (ICD-10-CM less granular), 425 (6.7%) had a one-to-many match (ICD-10-CM granular), and 1150 (18.2%) had a one-to-many match (ICD-10-CM more granular) and 1150 (18.2%) had a more complex mapping. Conclusions: GEMs provides a strict and limited mapping of many conditions necessitating additional work to fully define these conditions. When ICD-10-CM codes are in effect, validation studies are needed to verify the accuracy of the proposed mappings, particularly for commonly used conditions such as CCI. The CCI’s reliability and specificity, particularly, must be maintained with the new ICD-10-CM codes in order to continue to be useful as an important research tool.

PRM9 USE OF A COMMON DATA MODEL TO FACILITATE RAPID ANALYTICS SUPPORTING HEALTH OUTCOMES RESEARCH Kim H1, Joo S2, Atansit D3, Morrison J1, Reisinger S2, Germscheid L2, Murray R2
1Bristol-Myers Squibb Company, Pennington, NJ, USA, 2UBC, Harrisburg, PA

OBJECTIVES: Building a research using observational data is complex and costly due to massive and disparate formats across observational databases. Therefore, there is a need to facilitate consistent and efficient analysis. The objective of this study was to test the feasibility of, and efficiency of how disparate sources reflect "known" population characteristics, utilizing standardized patient selection, analyses and visualization software on data that has been transformed into a CDM. METHODS: Using CEWorks® software, AF patients treated with warfarin or NOA were selected from multiple Administrative Claims and EHR databases previously transformed into a CDM format. Rates of selected disease states were calculated using a CEWorks analysis module, and then compared to results published in a recent study. Results across disparate databases were imported into a visualization tool for further comparison among data sources. RESULTS: Preliminary results indicate that rates of selected disease state across disparate databases are similar to those published in a previous study. Total number and percentage reported in the study for NOAC [10,789 - 35%] and warfarin [19,964 - 65%] patients matched comparably at 8,093 [29%] NOAC patients and 30,133 [65%] warfarin patients in CDM version. Different data sources also show similar prevalence rates on selected disease state, although such similarities show some gaps introduced by region and other demographic variable. CONCLUSIONS: Use of a CDM enables rapid data analysis to be performed across multiple data sources, enabling meaningful comparisons across disparate data. Results are easily linked to data visualization tools for further analysis. This study has far reaching implications for data scientists using multiple, large data sources, and further study is in need to verify its practicality.

PRM40 ACCURACY OF A NATURAL LANGUAGE PROCESSING SOFTWARE DESIGNED TO COMPUTER AVERAGE WEEKLY DOSE FROM NATIVE MEDICATION SCHEDULE Lu C.C1, Leng J1, Cannon G2, Zhou X1, Harrison DJ1, Shah N3, Sauer BC1
1Optum, Eden Prairie, MN, USA, 2Takeda Pharmaceuticals International Inc., Deerfield, IL, USA, 3Amgen, Inc, Thousand Oaks, CA, USA

OBJECTIVES: We designed a software tool to estimate average daily dose from a native medication schedule. The software calculates drug interactions and side effects are well established in the electronic medical record but data are not consistently entered into the pharmacy dispensing or nurse administration structured data sources. Although CPT codes can be used to identify major infusion events but inconsistent coding does not allow estimation of the administered dose. To address this, we developed Natural Language Processing (NLP)