Spontaneous reports (SRs) of suspected adverse events associated with biopharmaceuticals and other medical products are key sources for identifying potential new drug hazards. They are required by most regulatory agencies as well as large drug distribution projects, e.g., HIV/TB programs. The WHO collects these pharmacovigilance data from >100 countries. Many databases conform to WHO's Council for International Organizations of Medical Sciences (CIOMS) standard data fields. B.R.I.D.G.E. TO DATA (www.bridgetodata.org), an international resource of database profiles, can serve as a template and complement the CIOMS effort. **OBJECTIVES:** Analyze SR database profiles and classify use of data fields. **METHODS:** We identified databases profiled in B.R.I.D.G.E. collecting SR data using two search criteria: *Database Type=Spontaneous reporting systems*; and *Database Source=Spontaneous reports*. Twenty of 198 profiles matched ≥1 criteria; frequency of use of the 75 data fields were compared. Based on use frequency, fields were categorized into Group 1 (G1 - consensus in use of field among the set), Group 2 (G2 - use by ≥50% databases), or Group 3 (G3 - use by <50% databases). **RESULTS:** Of the 75 data fields, 53 (71%) were frequently used among databases with SR data: 21 (28%) were classified into G1 (e.g., Diagnosis data, Drug data), and 32 (43%) into G2 (e.g., Death recorded, Physician specialty). Fields utilized less frequently (n=22; 29%) comprised G3 (e.g., Medical Record Access, Linkage Capabilities). Analysis of G1 revealed that a majority of SR databases are funded by government agencies, capture OTC & prescription drug use in inpatient & outpatient settings; however, diagnosis data are heterogeneously coded. Of the 25 fields on the CIOMS reporting form, 17 corresponded to G1, and 8 to G2. **CONCLUSIONS:** In this analysis, B.R.I.D.G.E. served as a screening tool to categorize data fields used in SR databases and successfully identified additional fields to complement the CIOMS effort.

PRM69

WHAT IS THE OPTIMAL SEARCH ENGINE FOR RESULTS FROM EMBASE AND MEDLINE: OVID OR EMBASE.COM?

Fortier KJ, Kiss N, Tongbram V

Oxford Outcomes, Morristown, NJ, USA

OBJECTIVES: Ovid and Embase.com are two search engines that are commonly used for searching the Medline and Embase databases. The standard of practice for conducting a search for a systematic literature review seems to support the use of Ovid. Our objective was to discern the advantages and disadvantages of both the Embase.com and Ovid search engines. We sought to compare the two databases in the following ways: results of a search, and ease of searching. **METHODS:** We conducted several searches to see if there were any differences in results. We analyzed the variation in results to determine why the disparities existed. We wrote a step-by-step search guide and highlighted the differences/difficulties encountered. Using a set of written directions, we also asked researchers to conduct searches using both interfaces and rate ease of use on a scale of 0 through 5. Research was also conducted through websites and help desks to ascertain the disparities in content between the two search engines. **RESULTS:** There were differences in the search results between Ovid and Embase.com using the same search criteria. The coding of articles sometimes differed, but there were very few, if any, relevant articles that were missed by either engine. In terms of ease of use, Embase.com had lower scores indicating greater ease of use in searching, exporting into a reference manager, saving searches, and recalling data compared to Ovid. In the research assessing the differences in content through the company websites, we found that Embase.com alleged to contain articles published further in the past, include more conference abstracts, and update Emtree terms more frequently than did Ovid. **CONCLUSIONS:** There is no notable difference between Ovid and Embase.com in terms of search results when searching Embase and Medline. However, Embase.com was rated as easier to use in several domains compared to Ovid.

PRM70

STATE-LEVEL, POPULATION-BASED CANCER REGISTRY AND ADMINISTRATIVE DATA LINKAGES IN THE UNITED STATES: A REVIEW OF THE LITERATURE

Foley K[1], Miller J[1], Bradley CJ[2]

[1]*Truven Health Analytics, Cambridge, MA, USA,* [2]*Virginia Commonwealth University, Richmond, VA, USA*

OBJECTIVES: Observational studies using administrative claims data are increasingly used for oncology outcomes research. The only population-based administrative data inclusive of disease stage and death from registries, however, are the SEER-Medicare data which exclude the under 65 population. Many state cancer registry and claims data linkages exist, but little is known about them. We conducted a literature review to identify and describe U.S. state-level, population-based cancer data linkages. **METHODS:** PubMed was searched for all years using the keywords "registry," "cancer OR oncology," and "claims OR administrative." Inclusion criteria were U.S. state cancer registries and administrative claims data linkages, exclusive of the SEER-Medicare database. **RESULTS:** A total of 687 abstracts were identified; 74 articles met the inclusion criteria. These papers represented 21 state cancer registries and linkages to Medicaid, Medicare, commercial insurance claims, state-level hospital discharge files, state-level screening program data, or state health care agency administrative files. Most linked Medicaid (N=12) and Medicare (N=9) data to assess quality of cancer care for specific types of cancer. Eighty percent of states specifically linked data on breast-cancer, 40% linked colorectal cancer, 25% linked on cervical, prostate or lung, and 40% were inclusive of all cancers. Only seven states (35%) linked commercial claims data or hospital discharge data (representing the under 65 population). **CONCLUSIONS:** Significant work is being conducted at the state level to generate cancer data, yet the under 65 population remains under-represented among linked cancer lives. Although fragmented, the existing linkages provide a foundation on which to build a more comprehensive oncology data system. Maintaining and expanding these data are critical for quality assessments, and disparity, outcomes and comparative effectiveness research. Barriers related to proprietary data and sustainable funding for linkages must be addressed before a data infrastructure that is broadly representative of the US national population can be generated.

PRM71

IDENTIFYING CHRONIC KIDNEY DISEASE STAGES USING PATIENT PRESCRIPTION INFORMATION

Cai Y[1], Han Y[2], Jiao X[3], Mu G[1]

[1]*IMS Health, Plymouth Meeting, PA, USA,* [2]*IPSEN Biopharmaceuticals, Inc., Basking Ridge, NJ, USA,* [3]*IMS Health, Alexandria, VA, USA*

OBJECTIVES: It's important for health care policy makers, payers and drug manufacturers to identify Chronic Kidney Disease (CDK) patient stages in order to evaluate the prevalence, economic burden or market opportunities. The office based medical claims data, which contains ICD-9 diagnosis and treatment information of CDK stages, has very limited coverage. To identify more CDK patients, we built and compared varies of statistical models and machine learning algorithms to project CKD stages using prescription database. **METHODS:** The model data contained year 2011 patient level CKD stage indications, longitudinal drug therapies, days of supplies, titration rates, Demographic characteristics, payment type, and physician specialties, etc. The data were randomly divided into a training set (66.7%) and a validation set (33.3%). The classification models used were Logistic regression, linear/quadratic Discriminant, Classification and Regression Tree (CART), C4.5 decision tree, Logit Boost classification tree, Bayes learning networks, Support Vector Machine and Neural networks. Bagging and boosting techniques were also tested to improve the precision. **RESULTS:** Logistic regression showed the best classification accuracy out of all models. The overall correct classification rate was 66.8%, Kappa Statistics 0.35, F-measure 0.63 and receiver operating characteristic (ROC) area 0.75. Among all CDK stages, CDK stage 1-3 were predicted most accurately with True Positive Rate (TPR) 93% and 65% precision. ESRD was moderately identified with TPR 46% and precision 76%. CDK 4 was most difficult to identify, with TPR 9% and precision 40%. The bagging and boosting improved decision trees showed comparable results. **CONCLUSIONS:** The modern machine learning algorithms were proved to be more accurate in many cases but for the CKD dataset the classical logistic model worked out better. The logistic model could fairly classify severe CKD stages (4-5) patients from lower stages (1-3) using prescription information but it had hard time to separate CKD stage 4 from 5.

PRM72

THE IMPACT OF CIGARETTE SMOKING ON MORTALITY FOR DIABETIC PATIENTS: A PROPENSITY SCORE MATCHING APPROACH

Cai B, McAdam Marx C, Nelson RE

University of Utah, Salt Lake City, UT, USA

OBJECTIVES: Confounding issues have not been accommodated adequately in the public health literature of smoking. This study examines the impact of smoking on mortality for patients with diabetes using a propensity score matching approach. **METHODS:** We employed the National Health Interview Survey (NHIS) data from 1997 to 2001. Patients who were older than 18 and have been diagnosed with diabetes are included in the study. The study outcome was all cause mortality, where death records obtained from National Death Index were available through Dec. 31, 2006. Propensity score matching (PSM) was performed based upon patients' smoking status. Time to event model was performed on the PS matched samples. **RESULTS:** There were 165,057 patients older than 18 participated NHIS survey between 1997 and 2001, of which, 10,194 (6.18%) and had prior diabetes. Our estimation results suggest stop smoking for less than 5 years was associated with a 67% greater risk of death than having never smoked; stop smoking for more than 5 years did not have a different risk of mortality than those who never smoked. In addition, smoking 16-34 cigarettes per day was associated with 52% greater risk of death, and smoking more than 34 cigarettes per day more than doubles the risk relative to the never smokers. **CONCLUSIONS:** This study indicates smoking reduced survival significantly for patients with diabetes; smoking cessation for more than 5 years prolonged survival. It provides important evidence to decision makers who attempt to advocate smoking cessation programs to patients with diabetes.

RESEARCH ON METHODS – Modeling Methods

PRM73

IDENTIFICATION OF PATIENTS AT HIGH RISK FOR SUBSEQUENT DEVELOPMENT OF ATRIAL FIBRILLATION OR STROKE USING A CLAIMS DATABASE

Reynolds MR[1], Hunter TD[2], Mollenkopf SA[3], Turakhia MP[4]

[1]*Harvard Clinical Research Institute, Boston, MA, USA,* [2]*S2 Statistical Solutions, Inc., Cincinnati, OH, USA,* [3]*Medtronic, Inc., Mounds View, MN, USA,* [4]*Stanford University School of Medicine, Palo Alto, CA, USA*

OBJECTIVES: To evaluate the impact of epidemiologically established risk factors, both singular and multiple, on the rates of AF and stroke in a large claims database. **METHODS:** Using patients in the Truven Health MarketScan® Commercial and Medicare Supplemental Databases with enrollment during the entirety of 2007 (baseline), six conditions were considered as risk factors for stroke and AF: age (grouped as <65, 65-74, 75+), heart failure (HF), hypertension (HTN), diabetes, renal insufficiency, and coronary artery disease. A single categorical variable was created reflecting each patient's exact combination of these conditions. Using this variable as the single predictor, Cox proportional