individuals of African genetic ancestry. We will approach this goal by completing the following objectives: (i) localize a genetic signal that accounts for the significantly increased risk for primary open-angle glaucoma in African Americans and (ii) utilize electronic health records (EHR) data to expand our understanding of risk to incorporate endophenotypes of glaucoma and other clinically recorded variables that may influence disease risk. METHODS/STUDY POPULATION: We will genotype at least 200 available African American samples with glaucoma on the Illumina Infinium® Expanded Multi-Ethnic Genotyping Array (MEGAEX) and perform admixture mapping. We will then access EHR data to expand our analysis beyond glaucoma to encompass other relevant risk modifiers captured in the clinical record. RESULTS/ANTICIPATED RESULTS: We anticipate localizing a genetic signal or signals that may account for the increased POAG risk in African Americans. Our calculations indicate that we have ~81% power to detect association at a LOD score of 2 and a risk ratio of 2. Thus, we are well-powered to detect a true signal at this modest level of association DISCUSSION/SIGNIFICANCE OF IMPACT: This project will not only help to achieve precision medicine by filling in the gaps in knowledge regarding glaucoma in African Americans, but it will also address health disparities and aid in the realization of the full potential of "big data" so that all of these elements can be incorporated into a better understanding of health disparities.

2140

Estimating microscopic structures of glomeruli in renal pathology

Pinaki Sarder, Rabi Yacoub and John E. Tomaszewski
University at Buffalo, State University of New York, Buffalo, NY, USA

OBJECTIVES/SPECIFIC AIMS: (i) Digitally quantify pathologically relevant glomerular microcompartmental structures in murine renal tissue histopathology images. (ii) Digitally model disease trajectory in a mouse model of diabetic nephropathy (DN). METHODS/STUDY POPULATION: We have developed a computational pipeline for glomerular structural compartmentalization based on Gabor filtering and multiresolution community detection (MCD). The MCD method employs improved, efficient optimization of a Potts model Hamiltonian, adopted from theoretical physics, modeling interacting electron spins. The method is parameter-free and capable of simultaneously selecting relevant structure at all biologically relevant scales. It can segment glomerular compartments from a large image containing hundreds of glomeruli in seconds for quantification—which is not possible manually. We will analyze the performance of our computational pipeline in healthy and streptozotocin induced DN mice using renal tissue images, and model the structural distributions of automatically quantified glomerular features as a function of DN progression. The performance of this structural-disease model will be compared with existing visual quantification methods used by pathologists in the clinic. RESULTS/ANTICIPATED RESULTS: Computational modeling will reveal digital biomarkers for early proteinuria in DN, able to predict disease trajectory with greater precision and accuracy than manual inspection alone. DISCUSSION/SIGNIFICANCE OF IMPACT: Automated detection of microscopic structural changes in renal tissue will eventually lead to objective, standardized diagnosis, reflecting cost savings for DN through discovery of digital biomarkers hidden within numerical structural distributions. This computational study will pave the path for the creation of new digital tools which provide clinicians invaluable quantitative information about expected patient disease trajectory, enabling earlier clinical predictions and development of early therapeutic interventions for kidney diseases.

2166

Semantic characterization of clinical trial descriptions from ClincalTrials.gov and patient notes from MIMIC-III

Jianyin Shao, Ram Gouripeddi and Julio C. Facelli

OBJECTIVES/SPECIFIC AIMS: This poster presents a detailed characterization of the distribution of semantic concepts used in the text describing eligibility criteria of clinical trials reported to ClincalTrials.gov and patient notes from MIMIC-III. The final goal of this study is to find a minimal set of semantic concepts that can describe clinical trials and patients for efficient computational matching of clinical trial descriptions to potential participants at large scale. METHODS/STUDY POPULATION: We downloaded the free text describing the eligibility criteria of all clinical trials reported to ClinicalTrials.gov as of July 28, 2015, ~195,000 trials and ~2,000,000 clinical notes from MIMIC-III. Using MetaMap 2014 we extracted UMLS concepts (CUIs) from the collected text. We calculated the frequency of presence of the semantic concepts in the texts describing the clinical trials eligibility criteria and patient notes. RESULTS/ANTICIPATED RESULTS: The results show a classical power distribution, $Y=2^{10}\ X^{(-2.043)}$, $R^2=0.9599$, for clinical trial eligibility criteria and $Y=5^{13}\ X^{(-2.684)}$, $R^2=0.9477$ for MIMIC patient notes, where Y represents the number of documents in which a concept appears and X is the cardinal order the concept ordered from more to less frequent. From this distribution, it is possible to realize that from the over, 100,000 concepts in UMLS, there are only ~60,000 and 50,000 concepts that appear in less than 10 clinical trial eligibility descriptions and MIMIC-III patient clinical notes, respectively. This indicates that it would be possible to describe clinical trials and patient notes with a relatively small number of concepts, making the search space for matching patients to clinical trials a relatively small sub-space of the overall UMLS search space. DISCUSSION/SIGNIFICANCE OF IMPACT: Our results showing that the concepts used to describe clinical trial eligibility criteria and patient clinical notes follow a power distribution can lead to tractable computational approaches to automatically match patients to clinical trials at large scale by considerably reducing the search space. While automatic patient matching is not the panacea for improving clinical trial recruitment, better low cost computational preselection processes can allow the limited human resources assigned to patient recruitment to be redirected to the most promising targets for recruitment.

2182

Developing a corpus for natural language processing to identify bleeding complications among intensive care unit patients

Rashmee Shah, Benjamin Steinberg, Brian Bucher, Alec Chapman, Donald Lloyd-Jones, Matthew Rondina and Wendy Chapman
The University of Utah School of Medicine, Salt Lake City, UT, USA

OBJECTIVES/SPECIFIC AIMS: An accurate method to identify bleeding in large populations does not exist. Our goal was to explore bleeding representation in clinical text in order to develop a natural language processing (NLP) approach to automatically identify bleeding events from clinical notes. METHODS/STUDY POPULATION: We used publicly available notes for ICU patients at high risk of bleeding (n = 98,586 notes). Two physicians reviewed randomly selected notes and annotated all direct references to bleeding as "bleeding present" (BP) or "bleeding absent" (BA). Annotations were made at the mention level (if 1 specific sentence/phrase indicated BP or BA) and note level (if overall note indicated BP or BA). A third physician adjudicated discordant annotations. RESULTS/ANTICIPATED RESULTS: In 120 randomly selected notes, bleeding was mentioned 406 times with 76 distinct words. Inter-annotator agreement was 89% by the last batch of 30 notes. In total, 10 terms accounted for 65% of all bleeding mentions. We aggregated these results into 16 common stems (eg, "hemorr" for hemorrhagic and hemorrhage), which accounted for 90% of all 406 mentions. Of all 120 notes, 60% were classified as BP. The median number of stems was 5 (IQR 2, 9) in BP Versus 0 (IQR 0, 1) in BA notes. Zero bleeding mentions in a note was associated with BA (OR 28, 95% CI 6.5, 127). With 40 true negatives and 2 false negatives, the negative predictive value (NPV) of zero bleeding mentions was 95%. DISCUSSION/SIGNIFICANCE OF IMPACT: Few bleeding-related terms are used in clinical practice. Absence of these terms has a high NPV for the absence of bleeding. These results suggest that a high throughput, rules-based NLP tool to identify bleeding is feasible.

2204

Evaluations of physiologic perturbations and their relationship with length of stay in neonatal hypoxic-ischemic encephalopathy

Susan Slattery, Lei Liu, Haitao Chai, William Grobman, Jennie Duggan, Doug Downey and Karna Murthy

OBJECTIVES/SPECIFIC AIMS: Neonatal hypoxic-ischemic encephalopathy (HIE) is frequently accompanied with physiologic perturbations and organ dysfunction. Markers of these perturbations and their associations with length of stay (LOS) are uncertain. To estimate the association between changes in selected physiologic and/or laboratory values with LOS in newborns with HIE. METHODS/STUDY POPULATION: Using the Children's Hospitals Neonatal Database (CHND), we identified neonates with HIE at our center born ≥36 weeks' gestation from 2010 to 2016. Those with major congenital anomalies were omitted. Infants uniformly received therapeutic hypothermia for 72 hours unless death occurred sooner. Inpatient vital signs and selected laboratory markers were collected from our institution's health informatics,

https://doi.org/10.1017/cts.2017.61 Published online by Cambridge University Press

electronic data warehouse (EDW) and then matched to records in CHND. With severity of HIE, gender, and confirmed seizures, each marker's association with LOS was calculated using multivariable Cox proportional hazards regression equations. These analyses were stratified by mortality. Candidate markers were vital signs, pulse oximetry, creatinine, acidosis (pH), international normalized ratio (INR), and supplemental oxygen (FiO_2). RESULTS/ANTICIPATED RESULTS: There were 66 eligible infants (38 males) and 1741 patient-days identified; Severe HIE (48%) and mortality (n = 21, 32%) were common. Overall, the median length of stay (mLOS) was 20.5 days (25th–75th centile: 10–31 days), although shorter for nonsurvivors [nonsurvivors mLOS = 8 days (5, 20); survivors mLOS = 24 days (14, 31), $p < 0.001$). Median birthweight and gestational age were 3.3 kg and 39.4 weeks' gestation, respectively. In survivors (n = 45, 1290 days), regression analyses demonstrated that none of the selected parameters were associated with LOS. Among nonsurvivors (n = 21, 451 days), diastolic blood pressure changes [hazard ratio (HR) = 0.93, 95% confidence interval (CI) = 0.88, 0.97, $p = 0.04$] was related to longer time of survival; conversely, temperature (HR = 2.0, 95% CI = 1.24, 3.26, $p = 0.005$) was related to shorter survival. Creatinine, pH, INR, FiO_2, or other vital signs were unrelated to time-to-death in nonsurvivors. DISCUSSION/SIGNIFICANCE OF IMPACT: In a pilot study of neonatal HIE, changes in physiologic values were related to duration of survival in nonsurvivors, while neither physiologic nor laboratory values were related to survivors' mLOS. These results both exemplify novel uses for disease-specific, exposure-outcome relationships using EDWs and incorporates required functionalities of required software patches to extract, clean, and report from clinical information captured in electronic health records. We anticipate that text mining with techniques such as natural language processing will augment associations and/or predictions of short-term outcomes.

2240

High-throughput phenotyping and the increased risk of OSA in Rosacia patients

Peter Elkin, Sarah Mullin, Sanjay Sethi, Shyamashree Sinha and Animesh Sinha

University at Buffalo, State University of New York, Buffalo, NY, USA

OBJECTIVES/SPECIFIC AIMS: To create a new semantically correct high-throughput phenotyping (HTP) platform. To demonstrate the utility of the HTP platform for observational research and can allow clinical investigators to perform studies in 5 minutes. To demonstrate the improved accuracy of observational research using this platform when compared with traditional observational research methods. To demonstrate that patients who have Roseacea are at increased risk of having obstructive sleep apnea (OSA). METHODS/STUDY POPULATION: This population is a set of 212,343 patients in the outpatient setting cared for in the Buffalo area over a 6-year period. All records for these patients were included in the study. Structured data was imported into an OMOP (OHDSI) database and all of the notes and reports were parsed by our HTP system which produces SNOMED CT codes. Each code is designated as a positive, negative or uncertain assertion and compositional expressions are automatically generated. We store the codified data 750,000,000 codes in Berkley DB, a NOSQL database, and we keep the compositional graphs in both Neo4J and in GraphDB (a triple store). Labs are coded in LOINC and drugs using RxNorm. We have developed a Web interface in .Net named BMI Search, which allows real-time query by subject matter experts. We analyzed the accuracy of structured Versus unstructured data by identifiying NVAF cases with ICD9 codes and then looked for any additional cases based on the SNOMED CT encodings of the clinical record. This was validated by 2 clinical human review of a set of 300 randomly selected cases. Separately we ran a study to determine the relative risk of OSA with and without Rosacea using the data set described above. We compared the rates using a Pearson χ^2 test. RESULTS/ANTICIPATED RESULTS: We are able to parse 7,000,000 records in an hour and a half on 1 node with 4 CPUs. This yielded 750,000,000 SNOMED CT codes. The HTP data set yielded 1849 cases using ICD9 codes and another 873 using the HTP-NLU data, leading to a final data set of 2722 cases from our population of 212,343 patients. In total, 580 patients had Rosacea;5443 patients had OSA without Rosacea and 51 patients had OSA with Rosacea. Patients with Rosaca had an 8.8% risk of OSA whereas patients without Rosacia only had a 2.6% risk of OSA. This was highly statistically significant with a $p < 0.0001$ (Pearson χ^2 test). The number needed to test was only 12. DISCUSSION/SIGNIFICANCE OF IMPACT: HTP can change how we do observational research and can lead to more accurate and more prolific investigation. This rapid turn around is part of what is necessary for both precision medicine and to create a learning health system. Patients with Rosacea are at increased risk of and should be screened for OSA.

2246

Characterization of resistant hypertension in a statewide electronic health record-based database (OneFlorida)

Caitrin W. McDonough, William R. Hogan, Betsy Shenkman and Rhonda M. Cooper-DeHoff

OBJECTIVES/SPECIFIC AIMS: Our objective is to create a Resistant Hypertension (RHTN) computable phenotype from electronic health record (EHR)-based data, and to determine the characteristics associated with RHTN within a large, diverse, EHR-based database. METHODS/STUDY POPULATION: The OneFlorida Clinical Research Consortium includes 10 unique health care systems providing care for approximately half of the state (48%, ~10 million). OneFlorida houses a Data Trust which contains longitudinal EHR data and claims data from these providers in a common format, the PCORnet common data model v3.0. For the current project, data from 5 health care systems were considered. All of the adult hypertension (HTN) patients with a HTN diagnosis from an outpatient encounter were extracted from the OneFlorida Data Trust. Additional data such as demographics, prescribing, and vitals information were also extracted. The RHTN computable phenotype was created by constructing a drug exposure variable that took into consideration the number of antihypertensive medications an individual was prescribed at any point in time over the course of the OneFlorida dataset. RHTN was defined as any blood pressure requiring four or more antihypertensive drugs, or uncontrolled blood pressure (≥140/90) on 3 antihypertensive drugs. RHTN cases had to meet the definition criteria twice during the data period, at least 30 days apart. All data extraction, computation phenotype coding, and statistical analyses were conducted using SQL or SAS. RESULTS/ANTICIPATED RESULTS: Our preliminary results show that there were n = 342,026 adults with a HTN diagnosis from an outpatient visit in the data set. After the RHTN computable phenotype was constructed, n = 11,670 RHTN cases were identified from the n = 130,901 HTN individuals with all of the required variables in the data set (8.9% RHTN prevalence). In all, 55% of RHTN cases were Black or African American, compared with the total HTN population (25% Black/African American). RHTN cases also had more prescriptions for loop diuretics, centrally acting agents, α-blockers, and vasodilators compared with the total HTN population. Not surprisingly, the RHTN cases had 26% of the antihypertensive prescriptions in the data set, and the RHTN cases had fewer blood pressure readings that were in control (only 49.4% of readings <140/90). DISCUSSION/SIGNIFICANCE OF IMPACT: Overall, our preliminary data shows that it is possible to create the very complicated computable phenotype of RHTN within the OneFlorida Data Trust. We found that the RHTN prevalence in OneFlorida is 8.9% which is consistent with previous studies from NHANES. Although promising, these results require further validation of the computable phenotype and replication in other similar data sets in order to ascertain their true meaning. Once validated, the experience gained from this computable phenotype can be applied to many other phenotypes.

2278

Identifying causative mutations in Treacher Collins syndrome using iobio

Alistair N. Ward, Matt Velinder, Chase Miller, Tony Di Sera, Yi Qiao, Dave Viskochil and Gabor Marth

The University of Utah School of Medicine, Salt Lake City, UT, USA

OBJECTIVES/SPECIFIC AIMS: The objective of the study was 2-fold; to identify potentially deleterious alleles in a child with Treacher Collins syndrome, and; to demonstrate the value of the iobio analysis platform for intuitively and rapidly analyzing genomic data. METHODS/STUDY POPULATION: We used the iobio suite of web-based applications to analyze quality metrics for the sequencing data and called variants for the proband and his parents. We then visually interrogated variants in genes potentially associated with the syndrome in real-time, using the intuitive gene.iobio application. We sought high impact variants that demonstrated a predicted impact on the protein function, and were simultaneously at low allele frequency in the general human population. Variants were also compared against the ClinVar database of known mutations to identify variants that have already been associated with this, or related syndromes in the literature or clinical studies. Finally, the gene.iobio tool allows users to interrogate the primary sequencing data to ensure that no variants had been missed by the primary variant calling pipeline. This analysis pipeline was performed using intuitive web-based apps in real time, and consequently represents a system that is available to users that traditionally are excluded from these analyses. RESULTS/ANTICIPATED RESULTS: The iobio suite was used to rapidly assess data quality and interrogate genetic variants for a child with

https://doi.org/10.1017/cts.2017.61 Published online by Cambridge University Press