made for elimination of the epidemic by 2030; yet major HCV cascade of care (CoC) barriers exist. We secured CTSA pilot funding to obtain preliminary data for an innovative clinical trial utilizing big data modeling toward HCV elimination. METHODS/STUDY POPULATION: Our pilot work has developed a coordinated, real-time clinical data management process across 3 major CTSA affiliated hospital systems (MedStar Health, Emory-Grady, and UT-Southwestern), and additional data will be obtained from a pragmatic clinical trial. Electronic medical records data will be mapped to the OHDSI model, securely transmitted to Oak Ridge National Laboratory, Knoxville, TN and exposed to integrated data, analytics, modeling and simulation (IDAMS). RESULTS/ANTICIPATED RESULTS: Our U01 CTSA application proposes that HCV-IDAMS will model modifications to the established HCV CoC at community and population levels and thus simulate future outcomes. As data volume increases, system knowledge will expand and recursive applications of IDAMS will increase the accuracy of our models. This will reveal real-world reactions contingent upon population dynamics and composition, geographies, and local applications of the HCV CoC. DISCUSSION/SIGNIFICANCE OF IMPACT: Only an innovative, integrated approach harnessing pragmatic clinical data, big data and supercomputing power can create a realistic model toward HCV elimination.

2356

openSESAME: a "search engine" for discovering drug-disease connections by leveraging publicly available high-throughput experimental data

Adam C. Gower, Avrum Spira and Marc E. Lenburg

OBJECTIVES/SPECIFIC AIMS: Microarray technology has produced large volumes of gene expression data profiling differences in gene expression in a vast array of conditions, much of which is publicly available. Methods to query these data for similarities in patterns of gene regulation are limited to comparisons between preannotated groups. In response, we developed openSESAME to find experiments where a set of genes is similarly coregulated without regard to experimental design. An important application of open-SESAME is drug repositioning: if a pattern associated with disease is reversed by a given drug, the drug might target disease-related processes. METHODS/STUDY POPULATION: Experiments from the Gene Expression Omnibus (GEO) were normalized, signature-association (SA) scores computed for each sample, experiments assigned enrichment scores, and ANOVAs used to assign significance to experimental variables automatically extracted from GEO. SA scores were also generated for hundreds of publicly available signatures, and pairwise correlations used to create a relevance network. RESULTS/ANTICIPATED RESULTS: Using signatures of estrogen and p63, we recovered relevant experimental variables, and with the network approach, we recovered previously reported associations between disease states and/or drug treatments. DISCUSSION/SIGNIFICANCE OF IMPACT: openSESAME has the potential to illuminate "dark data" and discover novel relationships between drugs and diseases on the basis of common patterns of differential gene expression.

2378

A scientometric analysis of CTSA collaboration and impact

Kristi Holmes, Ehsan Mohammadi, Karen Gutzman, Pamela Shaw and Donald Lloyd-Jones

OBJECTIVES/SPECIFIC AIMS: Translational science supports the continuum of activities from early-stage bench research to implementation of discoveries for better and faster treatments to more patients. Past studies have attempted to clarify our understanding of the spectrum of translational research by categorizing the activities into stages ranging from T0 to T4 using explanatory definitions. Unfortunately, this approach is often vague and relies on a process of manual classification and binning of research publications into predetermined categories. This study aims to provide a big-picture analysis of clinical and translational science (CTS) based on an in-depth analysis of the entire corpus of publications resulting from research funded by Clinical and Translational Science Awards (CTSA) U54 awards (through 2016). METHODS/STUDY POPULATION: We harvested bibliographic metadata from all papers that cited any of the U54 award numbers since the inception of the CTSA program to the most recent award announcement. Natural language processing techniques were used to create term co-occurrence networks based on English-language textual data. Relevant and nonrelevant terms were distinguished algorithmically and processed accordingly to provide the clustered visualization. RESULTS/ANTICIPATED RESULTS: With this approach, we uncovered 6 natural clustered areas of emphasis of published CTS research, the evolution of specific concepts through time, and gained a better understanding of their relative impact as demonstrated by citations. We performed additional analyses including discipline-specific impact assessment; identification of categories of excellence relating to both productivity and citations; characteristics of collaborative networks such as organizational, industry, and international collaborations and network dynamics; and resulting global impact of the CTSA program. DISCUSSION/SIGNIFICANCE OF IMPACT: Ultimately we gained a clearer understanding of the CTSA program, its evolution through scholarly publications, and key areas of impact of the program using computational, data-driven evaluation methods.

2412

Predicting response to hemodynamic interventions in the ICU using recurrent neural networks

Julian Genkins and Thomas A. Lasko

OBJECTIVES/SPECIFIC AIMS: Our goal is to explore the value of learning algorithms to improve both the efficiency and accuracy of a clinician undertaking the cognitive task of selecting the best resuscitative intervention for a hemodynamically unstable patient in the ICU. Machine learning is an ideal discipline to solve this problem. The ICU is a data rich environment, however there is significant uncertainty regarding the interdependency of this data. Experts consistently struggle to develop deterministic models of the underlying forces driving hemodynamic perturbations and intervention responsiveness. Machine learning, especially deep learning, assumes no correlation between inputs. Deep architectures disentangle these high-level relationships through exposure to abundant, diverse data sets such as those used in this project, obviating the need to manually explore confounding interactions. METHODS/STUDY POPULATION: We are using the "Medical Information Mart for Intensive Care" (MIMIC-III) database for this project. MIMIC-III is a large, single-center database comprising information relating to patients admitted to critical care units at Beth Israel Deaconess Medical Center, a large tertiary care hospital, from 2001 to 2012. It contains data associated with 38,597 distinct adult patients and 53,423 distinct hospital admissions for those patients, with a mean of 4579 charted observations and 380 laboratory measurements available for each hospital admission. Classes of data in the MIMIC-III are varied and include billing, intervention, laboratory, medication, and physiologic data among others. In addition to training an RNN in the task of predicting hemodynamic states, we will also attempt to train 2 additional models on the same data—a multidimensional linear regression and a nonsequence-oriented deep neural network. For each of these models we will measure accuracy using root mean squared error (RMSE) and mean absolute error (MAE) to provide scale-dependent measurements of accuracy. RESULTS/ANTICIPATED RESULTS: Our results will be reported in 2 primary categories: numerical accuracy of the RNN model and applicability, utility, and accuracy in a live clinical setting. The use of RNNs in biomedical informatics, and in general, is a relatively new phenomenon. This means that the body of literature which could provide a basis for our expected results is limited. Because of this we have chosen staged goals in assessing our model. First, we hope to achieve a model that reliably predicts the direction of response. Being able to answer only the question of how a patient will respond—will they move toward or away from our therapeutic goal—is as good as existing prediction methods. It is well established in the literature that, by almost any metric, ~50% of hemodynamically unstable patients respond to a fluid challenge. If we are within 10% of this average (40%–60% respond), then we can be confident in the accuracy of our model in predicting direction. Upon achieving this, we will then measure accurate prediction of response magnitude. To this affect, we hope to achieve an RMSE <10% between our test data and corresponding predicted output before proceeding further. In addition to numeric accuracy, we acknowledge that a plan for practical, clinical validation is needed before utilizing this tool in a clinical environment. Such validation will require 3 separate components. First, numeric accuracy will need to be determined again as compared with prospective data on actual patients in the ICU. This step is critical to prove that no information leakage from target data back to input data occurred during training. Second, there must be a comparison to existing prediction methods, such as the passive leg raise in combination with measurement of cardiac output to predict volume responsiveness. Finally, we must measure the cost to the clinician of implementing our model in an ICU, specifically how it impacts their time to accomplish the task of selecting an intervention for the hemodynamically unstable patient. However, these tasks are beyond the scope of this project and will be left for later investigations. DISCUSSION/SIGNIFICANCE OF IMPACT: If we are successful, this study will provide the first step toward a data-driven model for predicting patient responsiveness to a given hemodynamic intervention or collection of interventions. As compared with current

https://doi.org/10.1017/cts.2017.73 Published online by Cambridge University Press

best practice maneuvers, this model will not require manipulation of the patient, have less rigid criteria for reliable interpretation, and not require as specific of a technical skillset to interpret. Furthermore, it will include many common categories of resuscitative therapies (eg, vasopressors, inotropes, fluids) and will allow effects of a combination of interventions to be predicted while making no assumptions of interdependence between said interventions. This study will also contribute a novel process of sequence prediction using RNNs by incorporating an element of context on top of the sequential data in every training step. An RNN learning the sequence of hemodynamic data comprising a patient's hemodynamic state would, alone, fit into the realm of sequence prediction. Our innovation is the addition of treatment information with each temporal division of the hemodynamic data. The result is an RNN that combines the task of sequence prediction with sequence translation, the 2 major use cases for RNN learning algorithms.

2413

Immune stress biomarkers correlate to violence and internalization of violence in African American young adults

Latifa Jackson, Max Shestov, Forough Saadatmand and Joseph Wright
Georgetown - Howard Universities, Washington, DC, USA

OBJECTIVES/SPECIFIC AIMS: Allostatic load, the chronic stress-induced wear and tear on the body, has a cumulative deleterious effect in individuals over their lifetime. Recent studies have suggested that socio-economic status, psychological determinants, and biomedical health cumulatively contribute to allostatic load in young adults. Although these finding individually suggest that African American children may be particularly susceptible to the effects of allostatic loading due to racially-based discrimination and economic instability, few studies have shown the effect of exposure to violence on the allostatic load carried by young African Americans. METHODS/STUDY POPULATION: The Biological and Social Correlates of Drug Use in African American Emerging Adults (BADU) data set is composed of young African Americans (n = 557 individuals) living in the Washington, DC area, collected from 2010 to 2012. Study participants were sought equally between males and females (n = 283, n = 274, respectively). This data set provides a rich source of information on the behavioral, mental, and physical health of African American young adults (18–25 year olds) living in the Washington, DC area. Analysis of 6 biomedical markers were measured in BADU study participants: C-reactive protein, cortisol, Epstein-Barr virus IgG, IgE, IgA, and IgM, known to be markers of immune stress and allostatic load. Naive Bayes was used to identify participant responses that were correlated to elevated stress biomarker levels. RESULTS/ANTICIPATED RESULTS: Violence was most closely correlated to elevated EBVVCA IgM and IgE levels. Elevated IgE levels correlated to increased experience of familial violence and sexual abuse; familial drug abuse and depression; violence and community violence. Cortisol is positively correlated to reported emotional state ($R = 0.072$) and perceived individual discrimination ($R = 0.059$). DISCUSSION/SIGNIFICANCE OF IMPACT: Allostatic load appears to be high in individuals who self-report exposure to violence. Both perceived mental health and violence were correlated to elevated stress biomarkers. When Epstein-Barr virus viral capsid antigen IgM was compared with violence features characterized in the data set, we found that internalization of environmental stressors were most strongly correlated to elevated allostatic load markers. This work suggests that internalization of experienced violence may be as important as the actual violence experience.

2416

A machine learning pipeline to predict acute kidney injury (AKI) in patients without AKI in their most recent hospitalization

Samuel Weisenthal, Samuel J. Weisenthal, Caroline Quill, Jiebo Luo, Henry Kautz, Samir Farooq and Martin Zand

OBJECTIVES/SPECIFIC AIMS: Our objective was to develop and evaluate a machine learning pipeline that uses electronic health record (EHR) data to predict acute kidney injury (AKI) during rehospitalization for patients who did not have an AKI episode in their most recent hospitalization. METHODS/STUDY POPULATION: The protocol under which this study falls was given exempt status by our institutional review board. The fully deidentified data set, containing all adult hospital admissions during a 2-year period, is a combination of administrative, laboratory, and pharmaceutical information. The administrative data set includes International Classification of Diseases, 9th Revision (ICD-9) diagnosis and procedure codes, Current Procedural Terminology, 4th Edition (CPT-4) procedure codes, diagnosis-related grouping (DRG) codes, locations visited in the hospital, discharge disposition, insurance, marital status, gender, age, ethnicity, and total length of stay. The laboratory data set includes bicarbonate, chloride, calcium, anion gap, phosphate, glomerular filtration rate, creatinine, urea nitrogen, albumin, total protein, liver function enzymes, and hemoglobin A1c. The pharmacy data set includes, for each medication, a description, pharmacologic class and subclass, and therapeutic class. Data preprocessing was performed using Python library Pandas (McKinney, 2011). Top-level binary representation (Singh, 2015) was used for diagnosis and procedure codes. Categorical variables were transformed via 1-hot encoding. Previous admissions were collapsed using rules informed by domain expertise (eg, the most recent age or sum of assigned diagnosis codes were retained as elements in the feature vector). We excluded any patient without at least 1 rehospitalization during the time window. We excluded any admission with or without AKI where AKI was also present in the most recent hospitalization. For comparison, we do not exclude such admissions for an identical experiment in which we considered any AKI event as a positive sample (regardless of AKI presence in the most recent hospitalization). We defined an AKI event as an assignment of any of the acute kidney failure (AKF) ICD-9 codes [584.5, AKF with lesion of tubular necrosis, 584.6, AKF with lesion of renal cortical necrosis, 584.7, AKF with lesion of renal medullary (papillary) necrosis, 584.8, AKF with other specified pathological lesion in kidney, or 584.9, AKF, unspecified]. Since diagnosis codes are believed to be specific but not sensitive for AKI (Waikar, 2006), we supplemented them using creatinine for patients who had laboratory values. Diagnosis was made according to the Kidney Disease: Improving Global Outcomes (KDIGO) Practice Guidelines (AKI defined as a 1.5-fold or greater increase in serum creatinine from baseline within 7 d or 0.3 mg/dL or greater increase in serum creatinine within 48 h). We report preliminary model discrimination via area under the receiver operating characteristic curve (AUC) using k-fold cross validation grouped by patient identifier (to ensure that admissions from the same patient would not appear in the training and validation set). It was confirmed that the prevalence of positive samples in the entire data set was maintained in each fold. Python library Sci-kit Learn (Pedregosa, 2011) was used for pipeline development, which consisted of imputation, scaling, and hyper-parameter tuning for penalized (l1 and l2 norm) logistic regression, random forest, and multilayer perceptron classifiers. All experiments were stored in IPython (Pérez, 2007) notebooks for easy viewing and result reproduction. RESULTS/ANTICIPATED RESULTS: There were 107,036 adult patients that accounted for 199,545 admissions during a 2-year window. Per admission, there were at most 54 ICD-9 diagnoses, 38 ICD-9 procedures, 314 CPT-4 procedures, and 25 hospital locations visited. The admissions were 55% female, the average age was 46 ± standard deviation 20, and average length of stay was 2.5 ± 8.0 days. We excluded 2360 admissions that involved an AKI event that directly followed an admission with an AKI event and 4130 admissions that did not involve an AKI event but directly followed an admission with an AKI event. In total, there were 4561 (5.3%) positive samples (AKI during rehospitalization without AKI in the previous stay) generated by 3699 unique patients and 81,458 negative samples (non-AKI during rehospitalization without AKI in the previous stay) generated by 31,831 unique patients. When using any AKI event as a positive sample (regardless of whether or not AKI was in the most recent stay), the prevalence was 7.3% (6921 positive samples generated by 4395 unique patients and 85,588 negative samples generated by 33,287 unique patients). Best results were achieved with a code precision of 3 digits for which we had a total of 4556 features per patient. Fitted hyper-parameters corresponding to each classifier were logistic regression with l1 penalty C as 2×10^{-3}; logistic regression with l2 penalty C as 1×10^{-6}; random forest number of estimators as 100, maximum depth as 3, minimum samples per leaf as 50, minimum samples per split as 10, and entropy as the splitting criterion; and multilayer perceptron l2 regularization parameter α as 15, architecture as 1 hidden layer with 5 units, and learning rate as 0.001. Five-fold stratified cross validation on the development set yielded AUC for logistic regression with l1 penalty average 0.830 ± 0.006, logistic regression with l2 penalty 0.796 ± 0.007, random forest 0.828 ± 0.007, and multilayer perceptron 0.841 ± 0.005. In an identical experiment for which an AKI event was considered a positive sample regardless of AKI presence in the most recent stay, we had 4592 features per sample with the same code precision. Five-fold stratified cross validation on the development set with identical settings for the hyper-parameters yielded AUC for logistic regression with l1 penalty average 0.850 ± 0.004, logistic regression with l2 penalty 0.819 ± 0.006, random forest 0.853 ± 0.004, and multilayer perceptron 0.853 ± 0.006. DISCUSSION/SIGNIFICANCE OF IMPACT: Our objective was to investigate the feasibility of using machine learning methods on EHR data to provide a personalized risk assessment for "unexpected" AKI in rehospitalized patients. Preliminary model discrimination was good, suggesting that this approach is feasible. Such a model could aid clinicians to recognize AKI risk in unsuspicious patients. The authors recognize several limitations. Since our data set corresponds to a time-window sample, patients with high frequency of hospital utilization are likely overrepresented. Similarly, our data set contains records from only 1 hospital

https://doi.org/10.1017/cts.2017.73 Published online by Cambridge University Press