National Library of Medicine

Annual Training Meeting

June 22-24, 2020

Oregon Health & Science University
Portland, OR

OHSU

TABLE OF CONTENTS

EBook Acknowledgements

When OHSU volunteered to host the 2020 NLM Informatics Trainees Conference, we never envisioned the Covid-19 pandemic and the subsequent need to convert the conference to a virtual event. This first-ever Virtual NLM Informatics Trainees Conference would not have happened without the work of many people.

We first wish to thank the staff of OHSU who planned initially for an in-person event but then transitioned to planning a virtual meeting. They include:

Andrea Ilg
Kate Fultz-Hollis
Lynne Schwabe
Diane Doctor
Virginia Lankes
Lauren Ludwig
Monica Garlough
Brent Solheim
Dave Heron

We also want to thank the Student Program Committee, who developed the conference program:

Lily Cook, OHSU
Meena Mishra, OHSU
Sarah Mullin, Buffalo
Tony Sun, Columbia
Peter Hong, Harvard
Garret Eickelberg, Northwestern
Brandan Dunham, Pittsubrgh
Rebecca Rivera, Regenstrief
Varuna Chander. Baylor
Rocky Aikens, Stanford
Nicole Lopez. UCSD
Aaron Bohlmann, North Carolina
YiFan Wu, Washington
Janette Vazquez, Utah
Marily Barron, Vanderbilt
Chris Magnano, Wisconsin
David Chartash, Yale

Finally, we want to thank the NLM leadership and staff for their support and assistance:

Patricia Flatley Brennan
Valerie Florance
Christine Ireland
Alan VanBiervliet

William Hersh, MD, FACP, FACMI, FAMIA
Program Director, OHSU NLM T15 Biomedical Informatics Training Grant
Professor and Chair, Department of Medical Informatics & Clinical Epidemiology
School of Medicine, Oregon Health & Science University

2020 National Library of Medicine Training Meeting

Oregon Health & Science University

AGENDA

Monday, June 22, 2020

7:00 – 8:00 Conference Opening and Welcome

Dr. Sharon Anderson, Oregon Health & Science University

Dr. Bill Hersh, Oregon Health & Science University

Dr. Patricia Brennan, National Library of Medicine

Dr. Valerie Florance, National Library of Medicine

8:00 – 10:30 Plenary Session
Clinical and Population Health Informatics
Moderator, Dr. Larry Hunter, University of Colorado

Quantifying the Scope and Work of Secure Messaging Among Breast Cancer Care Team Members Bryan D. Steitz, Vanderbilt University

Provider Adherence to Syphilis Testing Guidelines Among Stillbirth Cases
Yenling A. Ho, Indiana University

Utility of EHR to Replicate Cochrane Systematic Reviews
Jason A. Thomas, University of Washington

Predictive Modeling of Bacterial Infections and Antibiotic Therapy Needs in Critically Ill Adults
Garrett Eickelberg, Northwestern University

Designs: Extending Experimental Principles to Observational Studies
Rachael C. Aikens, Stanford University

PheKnowLator: An Optimized Library for Automated Construction of Large Scale Heterogeneous Biomedical Knowledge Graphs

Tiffany J. Callahan, University of Colorado

mHealth for Mental Health: Acceptance and Design of an App to Help African American Women Manage Anxiety and Depression
Terika McCall, University of North Carolina at Chapel Hill

Predicting Preventable Hospital Readmissions with Causal Machine Learning: From Data to Decisions
Ben J. Marafino, Stanford University

Family History and Phenotype Heritability in Type I Diabetes
Erin M. Tallon, University of Missouri

Automated Conversational Health Coaching
Elliot G. Mitchell, Columbia University

10:30 – Noon Parallel Focus Sessions
Focus Session 1
Text Extraction – Summarization
Moderator, Dr. Harry Hochheiser, University of Pittsburgh

Improving Physical Activity Among Prostate Cancer Survivors and Their Loved Ones Through a Peerbased Digital Walking Program
Regina Casanova-Perez, University of Washington

Extracting Housing Stability Concepts from Clinical Text Notes
Andrew K. Teng, University of Washington

Developing an Efficient Model for Identifying and Extracting Social Determinants of Health from Clinical Notes
Rachel Stemerman, University of North Carolina at Chapel Hill

Literature-Based Discovery of Drug Repurposing for Rare Diseases
Daniel N. Sosa, Stanford University

Learning to Identify Synonymous Terms Across Biomedical Terminologies
Vinh Nguyen, National Library of Medicine

Focus Session 2
HIE – Interoperability
Moderator, Dr. Peter Elkin, University at Buffalo

Identifying Acute Respiratory Distress Syndrome in De-Identified EHR Data
V. Eric Kerchberger, Vanderbilt University

Representation of Medical Knowledge Towards Generalizable Prediction
Sarah Mullin, University at Buffalo

Informatics of Early Intervention Referrals for Children with Disabilities
Ben Sanders, Oregon Health & Science University

Health Information Exchange Use During Dental Visits
Heather L. Taylor, Indiana University

A SMART on FHIR and CDS-Hooks Enabled Approach to the Exchange and Review of Genomic Test Results
Michael Watkins, The University of Utah

Focus Session 3
Bioinformatics – Genomics
Moderator, Dr. Sylvia Katina Plevritis, Stanford University

Vitessce: Visual Integration Tool for Exploration of Spatial Single-Cell Experiments
Trevor Manz, Harvard University

Integration of Large-Effect Expression Variants Enhances Polygenic Risk Prediction
Craig Smail, Stanford University

Identifying Protein-Metabolite Networks Associated with COPD Phenotypes
Emily Mastej, University of Colorado

Alternative Splicing Event Signatures of Drug Response in AML
Julian Egger, Oregon Health & Science University

Heritable Components Explain Variation in Acute Kidney Injury Recovery Rate
Kathleen LaRow Brown, Columbia University

Investigating Somatic Mosaicism in Blood for Cardiovascular Disease Risk
Varuna Chander, Baylor College of Medicine

Noon – 12:45 Poster Pitch
Moderator, Dr. Karen Eden, Oregon Health & Science University

1:00 – 2:00 Funding your Research: NLM R01 and SBIR/STTR Grants
Dr. Alan VanBiervliet, National Library of Medicine/NIH

AGENDA

Tuesday, June 23, 2020

7:00 – 9:00 Plenary Session
Translational Bioinformatics
Moderator, Dr. Mark Craven, University of Wisconsin-Madison

Identifying MHC Class I Binding Peptides Using an Ultradense Peptide Array
Amelia K. Haj, University of Wisconsin - Madison

Elucidating Bio-Molecular Mechanisms of Disease for Novel Pathogenic Variants Using Structural Bioinformatics Approaches
Rolando Hernandez, The University of Utah

Prognostic Value of Imaging-Based Estimates of Glioma Pathology Pre-and Post-Surgery
Evan D. H. Gates, University of Texas, MD Anderson Cancer Center

An Empirical Bayes Approach to Image Normalization in Highly Multiplexed Cellular Imaging Data
Coleman R. Harris, Vanderbilt University

Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings
David Chang, Yale University

Prediction of Drug Co-Prescription Induced Adverse Events Using CANDO
Zachary Falls, University at Buffalo

A Reward System Polygenic Risk Score for Predicting Obesity and Substance Use Disorders
Kristen M. Stevens, Oregon Health & Science University

Multi-region Expression Profiling of Archived Breast Ductal Carcinoma In Situ
Adam Officer, University of California, San Diego

Genome-Wide Detection of Epistasis in Antibiotic Resistance *M. Tuberculosis*
Anna G. Green, Harvard Medical School

9:00 – 9:45 Poster Sessions
Poster Session 1
Machine Learning in Clinical Environments

Ethnicity-Associated Microbiome and Metabolome Differences Persist During a Short, Plant-based Diet
Robert H. Markowitz, Vanderbilt University

Clostridioides Difficile Infection Among Privately Insured Patients in the United States
Jessica El Halabi, Harvard University

Exploring How Temporal Representation Affects Deep Longitudinal Models
Matthew C. Lenert, Vanderbilt University

Fully Automatic Detection of REM Sleep Without Atonia
Daniel Yaeger, Oregon Health & Science University

Identifying Modifiable Predictors of Patient Outcomes after Intracerebral Hemorrhage with Machine Learning
Andrew N. Hall, Northwestern University

Improving Prediction of Survival for Extremely Premature Infants Born at 23 to 29 Weeks Gestational Age in the Neonatal Intensive Care Unit
Angie Li, University at Buffalo

Nonparametric Deep Survival Analysis: Regularization & Missingness
Shreyas Bhave, Columbia University

Optimizing Machine Learning Models for Clinical Application
Collin Engstrom, University of Wisconsin-Madison

Predict Late Patients in Pediatric Ophthalmology Clinic Using Machine Learning
Wei-Chun Lin, Oregon Health & Science University

Spatio-Temporal Analysis of the Effects of Air Pollution on COPD
Janette Vazquez, The University of Utah

Using High-Dimensional Pharmacogenomics Data to Predict Effective Antidepressant Treatment in Major Depressive Disorder Patients
Lauren M. Rost, University of Pittsburgh

Poster Session 1
Literature Mining, Computational Phenotyping and Exploratory Analyses

A Biomedical Use-Case Library and Voice Enabled Exploratory Search for Developing Intelligent User Interfaces
Michael Segundo Ortiz, University of North Carolina at Chapel Hill

Geostatistical Visualization of Ecological Interactions in Tumors

Hunter B. Boyce, Stanford University

Identifying Candidate Antibiotic Pairs for Alternating Treatment Strategies
Andrew Beckley, University of Pittsburgh

Incorporating Electronic Health Record and Genetic Data to Improve Coronary Artery Disease Risk Prediction
Harsh Patel, University of Washington

Data-driven Refinement of Gene Sets for Analysis of Disease Gene Expression
Alexander Wenzel, University of California, San Diego

Leveraging the Electronic Health Record to Evaluate the Effect of Anti-Hypertensive Medications on Mortality in COVID-19 Patients
Zachary Strasser, Harvard University

Unsupervised Literature Tagging of Computational Neuroscience Literature, Towards Question Answering
Evan Cudone, Yale University

9:45 – 10:30 Poster Session
Poster Session 2
Genomics and Bioinformatics

An Ensemble Approach to Study Tumor Evolution Using Multiple Samples
Carlos C. Vera Recio, University of Texas MD Anderson Cancer Center

Assessing the Influence of Regulatory Landscape Complexity on Gene Expression
Mary Lauren Benton, Vanderbilt University

Correlated Mutation Analysis Related to HIV-1 Drug Resistance
Skyler T. Kramer, University of Missouri

Inferring Signaling Pathways with Probabilistic Programming
David Merrell, University of Wisconsin - Madison

Enhanced Biological Interpretability of Single-Nucleus RNA-Seq Via Computational Correction of Systematic Biases
John Chamberlin, The University of Utah

Generalized Analysis of Environmental, Host, and Geographic Specificity in Microbiome Data
John L. Darcy, Colorado University

Inference of Novel Disease-Gene Associations Through Network Communities
Jennifer Asmussen, Baylor College of Medicine

Known Variant Aggregation to Aid Exome Variant Interpretation
David W. Sant, The University of Utah

MRP: Rare-Variant Analysis Via Bayesian Model Comparison Prioritizes Strong Risk and Protective Effects Across Biomarkers and Diseases
Guhan Ram Venkataraman, Stanford University

Sequencing Families Reveals Common Deletions that Contribute to Autism Risk
Kelly Paskov, Stanford University

Integrated Omics Modeling of Transcriptional Regulation in Medulloblastoma
Owen S. Chapman, University of California, San Diego

Zero-Inflated Random Forests for Genetic Regulatory Network Estimation in Single Cell RNA-Seq Data
Daniel Conn, University of Wisconsin - Madison

Review of Herbal Natural Products for Pain Management in Biomedical Classification Systems
Termeh Feinberg, Department of Veterans Affairs

Poster Session 2
Clinical Informatics

Computational Drug Repurposing Validation Strategies: A Review
Malvika Pillai, University of North Carolina at Chapel Hill

Early Evaluation of Cancer Prevention and Survivorship Care TeleECHO (Extension for Community Health Care Outcomes) Program
Zheng Milgrom, Regenstrief

EHR Utilization and Fragmentation: Sequence Analysis of Clinical Workflows
Amanda J. Moy, Columbia University

New Clinical Decision Support System for Veterans
Melissa Resnick, University at Buffalo

Syphilis Testing Adherence Among Women with Livebirth Deliveries: 2014-2016
Opeyemi C. Ojo, Regenstrief

The Spectrum of Recognition and Engagement of Chronic Kidney Disease Care of a Single Healthcare System
YiFan Wu, University of Washington

Ventilated Patients, Thromboprophylaxis and Major ICU Stay Related Outcomes: An Analysis of MIMIC-III Clinical Database
Mrigendra M. Bastola, National Institutes of Health

10:00 – 11:30 Director's Meeting

10:30 – Noon Career Panel – Preparing for Professions in Informatics
Virginia Lankes, OHSU DMICE Career Development Specialist

Dr. Dana Womack – Academia
A Learning Health Systems Science K Scholar in the OHSU-Kaiser Permanente Northwest Center of Excellence and K12 Training Program and Assistant Professor in the Department of Medical Informatics & Clinical Epidemiology and School of Nursing at Oregon Health & Science University in Portland, Oregon

Dr. Mitzi Boardman – Government/Consulting
Dentist and fellowship-trained informaticist supporting federal, enterprise electronic health record (EHR) implementations as well as contributing to the development of a research-based COVID-19 website and application.

Dr. Steven Kassakian – Operational Clinical Informatics
Associate Chief Health Information Officer at OHSU. His interest is in general internal medicine and particularly how health information technology can improve outcomes, provider efficiency and patient engagement.

Dr. Nathan Lazar – Industry
Senior data scientist, leading a small team of researchers at Recursion Pharmaceuticals.
Our work involves processing huge amounts of cellular imaging data from high-throughput drug screens and computationally searching for potential treatments for rare genetic diseases.

AGENDA

Wednesday, June 24, 2020

7:00 – 8:30 Parallel Focus Session
Focus Session 4
Machine Deep Learning – Graphical Models
Moderator, Dr. Tsung-Ting Kuo, University of California, San Diego

Mechanism Generalization Using Mechanism Centrality
Harrison Piekle-Lombardo, University of Colorado

A Bayesian Approach to Local Causal Discovery in the Presence of Latent Confounding

Bryan Andrews, University of Pittsburgh

Lopsided Bootstrap Loss Enables Segmentation Networks to Learn from Annotations with Multiple False Negatives
Darvin Yi, Stanford University

Decoding Cancer Evolution via Clinical Phylogenetics and Machine Learning
J. Nick Fisk, Yale University

Identifying Novel, Precancer-Associated Stem Cell Subtypes Through Regulatory Network Inference and Machine Learning
Bob Chen, Vanderbilt University

Predictive Modeling of Post-Operative Discharge to Skilled Nursing Facilities (SNF)
Nicole E. Lopez, University of California, San Diego

Applying Open-Set Learning Methods to Cancer Treatment Predictions
Alexander Cao, Northwestern University

Focus Session 5
Data Science – Population Health
Moderator, Dr. Chi-Ren Shyu, University of Missouri

Automatic Discovery of Complex Interactions Between Multiple Risk Factors
William Baskett, University of Missouri

Interactive Visualization for Medical Images in Jupyter Notebooks
Brian Pollack, University of Pittsburgh

Exploring Administrative Time-of-Day Efficacy for Seasonal Influenza Vaccinations
Darwin Y. Fu, Vanderbilt University

Evaluating Electronic Health Record Use Among Academic Ophthalmologists
Sally L. Baxter, University of California, San Diego

Investigating the Link Between Obstructive Sleep Apnea and Air Quality: What Can We Learn from CPAP Device Data?
Jay P. Kitt, The University of Utah

Patient Perspectives on Using Conversational Agents for Hypertension Medication Self-Management
Ashley C. Griffin, University of North Carolina at Chapel Hill

Characterizing Patterns of Disease Diagnosis Across Men and Women
Tony Y. Sun, Columbia University

Focus Session 6
Pharmacoinformatics
Moderator, Dr. Lydia Kavraki, Rice University

Toward *in vivo* Molecule Effect Prediction with Deep Learning and Fast Proteomics
Jesse G. Meyer, University of Wisconsin - Madison

Leveraging Partitioning Statistics for Scalable Microbial Identification
Pavan K. Kota, Rice University

cando.py: Software for Drug Candidate Prediction with Application to SARS-CoV-2
William Mangione, University at Buffalo

Identifying Neoantigens From Noncanonical Transcription and Translation Events at Complex Structural Variants in Human Cancers
Vinayak Viswanadham, Harvard University

Investigating the Antibiotic Biosynthetic Potential of the ADT Microbiome
Reed M. Stubbendieck, University of Wisconsin - Madison

The Unreasonable Effectiveness of Naïve Bayes for Predicting Combinatorial Drug Effects
Hannah A. Burkhardt, University of Washington

8:30 – 10:00 Parallel Open Mic Session

Open Mic Session 1
Data Science and Bioinformatics
Moderator, Dr. Karen Eilbeck, The University of Utah

Dissecting the Tumor Ecosystem Mediating Metastatic Breast Cancer
Emily K. Stroup, Northwestern University

Deep Generative Graph Neural Networks for de Novo Drug Design and Optimization
Benjamin Kaufman, University of Wisconsin - Madison

Imputation of Sparse Metabolomic Data
Elin Shaddox, University of Colorado

Modeling Histone Modifications at a Single-Cell Level Using EpiTOF
Laurynas Kalesinskas, Stanford University

Diversity in Pharmacogenomics: A Review Article
Roderick Gladney, University of North Carolina at Chapel Hill

Multidrug Resistant Organism Carriage in Wisconsin Children
Ashley Kates, University of Wisconsin

Synteny Maps Lead to High Quality Ortholog Predictions
Nicholas P. Cooley, University of Pittsburgh

A Grammar of Patient-Centered Communication Documentation
David Chartash, Yale University

MREC: A Fast Framework for Aligning Single Cell Molecular Data
Mathieu Carrière, Columbia University

Determining Optimal Patient Selection for Adjuvant Chemotherapy after Surgical Resection in Patients with Stage II and III Colorectal Cancer Using Computational Modeling
Justin J. Hummel, University of Missouri

Automated Node-Positive Classification and Feature Extraction Using Tumor Micro-Environment Imaging
Gautam Machiraju, Stanford University

Open Mic Session 2
Clinical and Public Health Informatics
Moderator, Dr. Robert McDougal, Yale University

Improved Physician Follow-Up Using Post-Handoff Report Outcomes (PHaROs)
Cindy C. Reynolds, University of California, San Diego

A Convex Framework for Optimal Resource Allocation and Influence Maximization
Michael Liou, University of Wisconsin - Madison

EHR-Based CDS Implemented to Deliver Public Health Information to Prevent and Manage Disease Outbreak Following the Inception of Meaningful Use – A Scoping Review
Jean Frédo Louis, The University of Utah

Learning Tasks of Pediatric Providers from Electronic Health Record Audit Logs
Barrett Jones, Vanderbilt University

Examining Organizational Characteristics and Perceptions of Clinical Event Notification Services in Health Care Settings
Kevin Wiley, Jr., Indiana University

Delays in Care and Differential Presentation of Acute Myocardial Infarction
Harry Reyes Nieva, Columbia University

Thinking Fast and Slow: Visualization of Financial Conflict of Interest Disclosure in Clinical Trial Publications
Alex Rich, University of North Carolina at Chapel Hill

Private Insurance Market Participation by Psychiatrists in Massachusetts
Nicole M. Benson, Harvard University

Impact of a Clinical Decision Support Tool in the Evaluation of Acute Pulmonary Embolism
Keaton Morgan, The University of Utah

Health Information Exchange in Primary Care: Use Among Teams
Nate C. Apathy, Regenstrief

Yes, We Closed the Tap! Now What?
Meenakshi Mishra, Oregon Health & Science University

A Proposal for the Relationship Between Physical Therapy and Opioid Use for US Veterans
Lindsey Brown-Taylor, Department of Veterans Affairs

Diagnostic Test Follow-Up Support: Incomplete Test Notifications, Purge of Non-Essential Notifications, and Flexible Secondary Pool Assignment
Peter Hong, Harvard University

Open Mic Session 3
Machine Learning
Moderator, Dr. Noemie Elhadad, Columbia University

Stabilized Image Segmentation in the Presence of Noise
Jonas A. Actor, Rice University

Decomposition of Clinical Disparities with Machine Learning
Noah Hammarlund, University of Washington

Electronic Health Record Phenotyping of Patients with Drug-Induced Renal Injury
Zaid K. Yousif, University of California, San Diego

Developing and Testing Clinical Decision Support for Neonatal Ventilator Management
Lindsey A. Knake, Vanderbilt University

Acceleration Signals in Determining Gait-Related Difficulties and the Motor Skill of Walking in Older Adults
Pritika Dasgupta, University of Pittsburgh

Predicting Total Daily Dose of Insulin for Optimal Blood Glucose Control in Hospitalized Patients
Minh Nguyen, Stanford University

Predicting Suicidal Ideation and Attempt in Children Aged 9-10
Gareth Harman, Oregon Health & Science University

Clinical Decision Support for Predicting Postpartum Depression Using Machine Learning
Houda Benlhabib, University of Washington

Developing a Predictive Model for Endometriosis Diagnosis
Amber Kiser, The University of Utah

A Bayesian Hidden Markov Model for Assessing Seizure Risk
Emily T. Wang, Rice University

How to Capture, Characterize, and Classify What We Don't Know
Mayla R. Boguslav, Colorado University

An Interpretable Electronic Health Record (EHR) Phenotype for Treatment Resistance in Depression (TRD)
James T. Brown, Vanderbilt University

10:00 – 11:00 Ada Lovelace Computational Health Lecture

AI in the Age of COVID-19: Computational Tools for the Classification, Prediction, and Characterization of a Pandemic
John H. Holmes, PhD, FACE, FACMI, FIAHSI
Department of Biostatistics, Epidemiology, and Informatics
University of Pennsylvania Perelman School of Medicine

This lecture will be live-streamed globally at videocast.nih.gov – currently in 'Future Events' section
The talk with be archived and available on the NIH Video Casting and will be in "Past Events' section the day after the talk.

11:00 – Noon Closing Session

Award Ceremony

ABSTRACTS
Monday, June 22, 2020

Plenary Session 1
Clinical and Population Health Informatics

Quantifying the Scope and Work of Secure Messaging Among Breast Cancer Care Team Members

Authors: Bryan D Steitz, Kim M Unertl, Mia A Levy, Vanderbilt University

Abstract: Asynchronous messaging is an integral but potentially inefficient means of communication in clinical settings. The varied acuity of medical needs contained within any given message often require providers to constantly manage their inbox throughout the day. In this study, we utilize the data stored in electronic health records (EHRs) to investigate the work incurred by EHR-based asynchronous messaging. We combined EHR access logs and secure messaging logs into sessions of activity, defined as any sequence of events that occur within fifteen minutes of any other event by the same employee about the same patient. We studied a cohort of 9,761 patients at Vanderbilt University Medical Center (VUMC) who had an appointment with a breast medical or surgical oncologist. These patients were the subject of 430,857 message threads, which involved 7,194 VUMC employees over a three-year period. Through a cluster analysis, we identified 114 employees who were highly involved in breast cancer treatment. These employees performed messaging actions in 516,325 (37.5%) sessions, averaging 29.8 messaging sessions per day. Messaging sessions lasted an average of 1.1 minutes longer than sessions that did not include messaging. Cancer providers performed messaging actions in 15 sessions per day on days when they did not have clinical responsibilities. We found that many messaging tasks are a result of triage work, in which providers view messages in their inbox, but do not send a subsequent message. Our results suggest that EHR-based asynchronous messaging is a primary work product that is integral to delivering and coordinating care across team members of all roles. By better understanding how asynchronous messaging contributes to EHR work, we can begin to create and evaluate informatics initiatives to systematically identify situations in which we can improve employee workload by reducing unnecessary interruptions.

Statement of Significance: Recent studies have related asynchronous messaging work to professional burnout among the healthcare workforce. This type of messaging is integral to coordinating care, but can be interruptive to clinical workflow and lead to unnecessary triage work. This study is one of the first to investigate the electronic work of asynchronous communication on care team members. Understanding work patterns associated with asynchronous messaging by combining EHR data sources can help to systematically identify and alleviate unnecessary interruptions.

Keywords: EHR Work, Communication, Workflow
E-mail of First Presenter: bryan.steitz@vanderbilt.edu

Provider Adherence to Syphilis Testing Guidelines Among Stillbirth Cases

Authors: Yenling A Ho, Katie Allen, Brian E Dixon, Indiana University

Abstract: Objective: The Centers for Disease Control and Prevention (CDC) has recommended that all women who have a stillbirth delivery have a syphilis test after delivery. Our study is to evaluate adherence to CDC guidelines for syphilis screening among women with a stillbirth delivery and determine the validity of using ICD-9-CM and ICD-10-CM codes to identify cases of stillbirth.
Methods: We utilized data recorded in electronic health records for women who gave birth between January 1, 2014 and December 31, 2016. Patients were included if they were 18-44 years old and possessed an ICD-9-CM or ICD-10-CM diagnosis of stillbirth. To evaluate syphilis screening, we estimated the proportion of women who received syphilis testing within 300 before stillbirth, within 30 days after their stillbirth delivery, and women who received syphilis testing both before and after stillbirth delivery. To validate ICD-9-CM and ICD-10-CM codes, we calculated positive predictive values based on the gestational age and signs of life at time of delivery.
Results: Among confirmed stillbirth cases, 51.4% had any syphilis testing conducted, 31.4% had testing before their stillbirth delivery, 42.9% had testing after the delivery, and only 22.9% had testing before and after delivery. For the validity of ICD-9-CM and ICD-10-CM codes, positive predictive values were 69.0%, and 50%, respectively.
Conclusion: A majority of women with a stillbirth delivery do not receive syphilis screening adherent to CDC guidelines. Stillbirth ICD codes do not accurately identify cases of stillbirth.

Keywords: STD, Epidemiology, Validation Study
E-mail of First Presenter: yeho@iu.edu

Utility of the EHR to Replicate Cochrane Systematic Reviews

Authors: Jason A Thomas, Adam B Wilcox; University of Washington Department of Biomedical Informatics and Medical Education, Seattle WA USA; UW Medicine, Seattle, WA USA

Abstract: To examine the link between clinical trials and structured Real-World-Data (RWD) found in Electronic Health Records (EHR), we conducted a review of the Cochrane Database of Systematic Reviews (CDSR) to evaluate the abundance of potential findings that could be replicated through mining the EHR. A single reviewer (JAT) assessed the most recent 2-4 reviews for 24 randomly selected Cochrane Review Groups and Topics (e.g. *Common Mental Disorders*) to determine whether or not review findings: (1) had ≥low evidence, (2) could be mined from an EHR, (3) could be mined from *structured* data in an EHR, (4) were pediatric specific. A subset was reviewed by a second reviewer (ABW) for validation. Our cursory review evaluated 50 CDSRs yielding a total of 14 reviews (28%) with 39 suitable findings meeting our inclusion criteria. Suitability required ≥low evidence and ability to be mined from structured data present within an EHR. A total of three CDSRs (6%) of the 50 reviewed were excluded solely due to the need for analysis of unstructured data. Of the 15 suitable CDSRs, seven (46%) were conducted on solely pediatric (≤18 years old) populations. Our findings report a higher feasibility of replicating clinical trials using RWD than a recently published review (Bartlett et. al, 2019) on a different population which found that 15% of clinical trials could feasibly be replicated from mining the EHR.

Statement of Significance: The importance of using Real-World-Data to inform clinical decision making has increased since the FDA incorporated doing so into its 2019 Strategic plan. Clinical trials are time-consuming, expensive, and are conducted under conditions with populations that do not necessarily

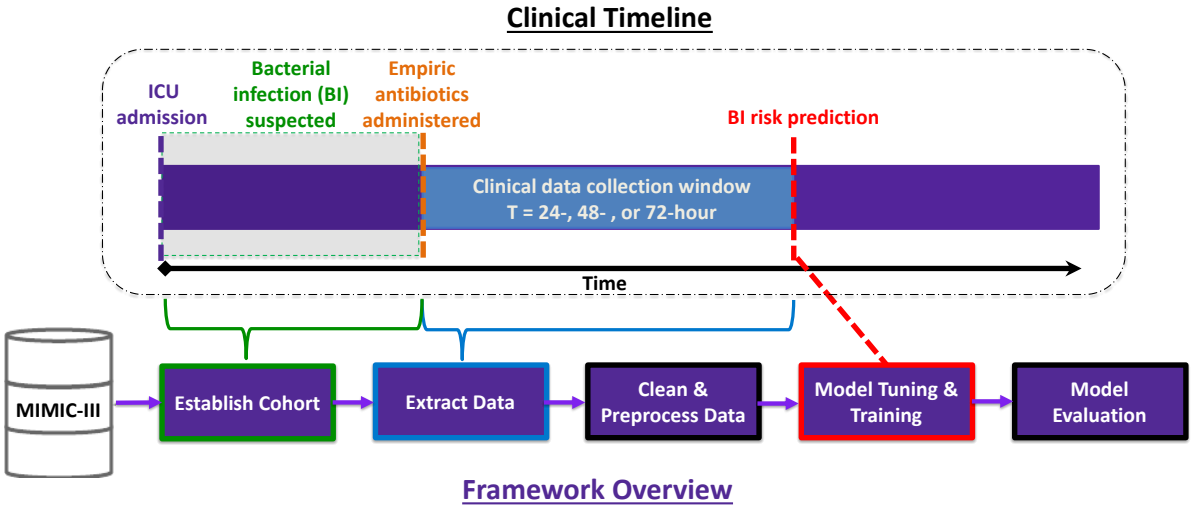
represent real world conditions and patients in practice. Real World Data found in Electronic Health Records may augment clinical trials and provide long-term post market surveillance. Thus, it is important to quantify the opportunity for replicating clinical trials using observational EHR data and identify areas for improvement in data capture to increase the utility of RWD for this purpose.

Keywords: Electronic Health Records, Real World Data, Systematic Review
E-mail of First Presenter: thomasjt@uw.edu

Predictive Modeling of Bacterial Infections and Antibiotic Therapy Needs in Critically Ill Adults
Author: Garrett Eickelberg, Northwestern University

Abstract: Prolonged antibiotic regimens in the intensive care unit (ICU) are associated with adverse patient outcomes and antimicrobial resistance. Bacterial infections (BI) are both common and deadly in ICUs, and as a result, patients with a suspected BI are routinely started on broad-spectrum antibiotics prior to having confirmatory microbiologic culture results, a practice known as empiric antibiotic therapy (EAT). However, EAT guidelines lack consensus and existing methods to quantify patient-level BI risk rely largely on clinical judgement and inaccurate biomarkers or expensive diagnostic tests. As a consequence, patients with low risk of BI often are continued on EAT, exposing them to unnecessary side effects. Augmenting current intuition-based practices with data-driven predictions of BI risk could help inform clinical decisions to shorten the duration of unnecessary EAT and improve patient outcomes. We propose a novel framework to identify ICU patients with low risk of BI as candidates for earlier EAT discontinuation. For this study, patients suspected of having a community-acquired BI were identified in the Medical Information Mart for Intensive Care III (MIMIC-III) dataset and categorized based on microbiologic culture results and EAT duration. Using structured longitudinal data collected up to 24, 48, and 72 hours after starting EAT, our best models identified patients at low risk of BI with AUROCs up to 0.8 and negative predictive values >93%. Overall, these results demonstrate the feasibility of forecasting BI risk in a critical care setting using patient features found in the electronic health record and call for more extensive research in this promising, yet relatively understudied, area.

GRAPHICAL ABSTRACT



Keywords: Critical Care; Prediction Models; Antibiotic Stewardship; Machine Learning; MIMIC; Electronic Health Records
E-mail of First Author: GarrettEickelberg2023@u.northwestern.edu

Designs: Extending Experimental Principles to Observational Studies

Authors: Rachael C. Aikens, Dylan Greaves, Michael Baiocchi

Abstract: In an observational study, researchers are not allowed to dictate which individuals receive a treatment or exposure and which do not. This **Pilot** gives rise to concerns that the results of an observational study may be biased due to confounding factors - characteristics of the individuals in the study that influence both their selection of treatment and their probable outcome. Matching methods seek to account for this self-selection by grouping treated and control individuals with similar baseline characteristics. Methodological studies suggest that using a 'prognostic score' to inform the matching process increases the precision of the effect estimate may reduce sensitivity to bias from unmeasured confounding factors. This talk will discuss the practical implementation of the prognostic score to improve observational study designs, introducing the concept of the 'pilot design' for large observational studies. A common mistake is to believe reserving more data for the analysis phase of a study is always better. This talk suggests how clever use of data in the design phase of large studies can lead to major benefits in the robustness of the study conclusions.

Statement of Significance: Electronic medical record data promises to enable more observational research. However, the most difficult questions to address with an observational study are questions of causality: how does a treatment or exposure effect an outcome? This talk discusses methods to improve causal inference from observational studies of medical data. First, I will explain the prognostic score matching framework, originally proposed by Hansen (2008). Next, I will introduce the 'Fisher-Mill plot', a visualization of the study population in terms of probability of treatment (propensity score) and probable outcome (prognostic score). Finally, I will discuss a 'pilot design' approach and show how it

can enable prognostic score techniques while maintaining careful separation of the design and analysis phases of a study. These techniques outline a new approach for large studies of medical data, in which data is used to cleverly in the design phase to enable more robust inference on questions of causality.

Keywords: causal inference, observational studies, matching, propensity score, prognostic score, Fisher-Mill plots
E-mail of First Presenter: raikens@stanford.edu

PheKnowLator: An Optimized Library for Automated Construction of Large Scale Heterogeneous Biomedical Knowledge Graphs

Authors: Tiffany J Callahan, Ignacio J Tripodi, Jordan M Wyrwa, William A Baumgartner Jr*, Lawrence E Hunter*

Abstract: Knowledge graphs (KGs) facilitate the representation of complex relationships among heterogeneous data types and have been used extensively in biomedical research to model biological phenomena. While many data-driven KG construction methods have been developed, they remain largely unable to construct KGs from multiple disparate data sources, combine KGs created by different systems, and collaborate or share KGs across institutions due to their inability to account for the use of different schemas, standards, and vocabularies. Used extensively in life sciences research, the Semantic Web was created to resolve these types of knowledge integration problems. The Web Ontology Language (OWL) is a Semantic Web standard for a graph-based knowledge representation and reasoning framework. OWL is highly expressive, enabling the integration of heterogeneous data using explicit semantics, and allows for the generation of new knowledge using deductive logic. Unfortunately, existing OWL-based KG construction methods are often built using complicated programs or toolsets, in arcane or difficult to use programming languages and require extensive computational resources. We introduce PheKnowLator (Phenotype Knowledge Translator), a fully automated Python 3 library explicitly designed for optimized construction of semantically-rich, large-scale biomedical KGs from complex heterogeneous data. The PheKnowLator framework provides detailed Jupyter Notebooks and scripts which greatly simplify KG construction, assisting even non-technical users through all steps of the build process. To accommodate a wide range of users and use cases, PheKnowLator has three build types (partial, full, and post-reasoner), can include inverse edges to link nodes, outputs KGs with and without OWL semantics (e.g. OWL-NETS), and generates KGs in several formats (e.g. triple edge lists, OWL API-formatted RDFXML, graph-pickled Networkx MultiDiGraph). To demonstrate PheKnowLator's performance, we constructed a KG of human disease mechanisms using eight open biomedical ontologies, twenty-four linked open datasets, and results from two large-scale, experimentally-derived datasets. Highly performant, PheKnowLator is able to construct a KG with 5.2 million nodes and 31.7 million edges from these sources in mere hours (comparable to runtimes for existing systems containing far fewer attributes, e.g. Hetionets, which contains 47,031 nodes and 2.3 million edges) on a laptop with 16 GB of RAM and a 2.7 GHz processor with 8 cores. Further, using the ELK reasoner we were able to deductively close PheKnowLator KG, verifying it was logically consistent and thus validating the KG construction process. Documentation and code are available at: https://github.com/callahantiff/PheKnowLator.

Keywords: Knowledge-based data science, network science, ontologies.

E-mail of First Presenter: **tiffany.callhan@cuanschutz.edu**

mHealth for Mental Health: Acceptance and Design of an App to Help African American Women Manage Anxiety and Depression

Authors: Terika McCall, MBA, MPH and Saif Khairat, MPH, PhD, University of North Carolina at Chapel Hill

Abstract: ***Background:*** The rates of mental illness among African American women are comparable to the general population, however they significantly underutilize mental health services compared to their white counterparts. Previous studies revealed that telehealth interventions increase access to mental health services and resources, and are effective in reducing anxiety and depression. Approximately 80% of African American women own smartphones. This presents a great opportunity to use mobile technology to help reduce the disparity in mental health service utilization and improve health outcomes for this population.
Aims: The aims of this exploratory study are to: (1) gauge the acceptability of using a smartphone app to help African American women manage anxiety and depression and (2) determine what culturally-tailored content should be included in an app tailored to help African American women manage anxiety and depression.
Methods: Women who identify as African American were eligible to participate in the study. A Web-based questionnaire was launched in November 2019 and closed in January 2020. Focus groups were held at a Durham County library and at the University of North Carolina at Chapel Hill in January 2020.
Results: The results of the survey (N= 395) showed statistically significant associations between age (less than 50 years old vs. 50+) and agreement with the use of a mobile app to communicate with a professional to receive help for managing anxiety and depression (all $p < .05$). Focus group participants (N= 20) voiced that content on how to deal with overt racism and sexism, microagressions, imposter syndrome, relationships, and the "Superwoman complex" should be included in the app.
Conclusions: Younger African American women were more likely to endorse the use of a mobile app to help them manage anxiety and depression. App content should be tailored based on user-centered recommendations to increase adoption and daily use.

Statement of Significance: Taking into account the burden of unmet need and disparity in mental health service utilization among African American women, there is great potential to use mobile technology to deliver services to this population. Over 70% of survey respondents reported accessing mobile apps 4 or more times per day. To increase the likelihood of adoption among this population, the app should be culturally-tailored. A "one-size-fits-all" approach to designing telehealth interventions to help African American women manage anxiety or depression my lead to more options but continued disparity in receiving care.

Keywords: telemedicine, mhealth, mental health, African Americans, women
E-mail of First Presenter: tmccall@unc.edu

Predicting Preventable Hospital Readmissions with Causal Machine Learning: From Data to Decisions

Authors: Ben J Marafino[1], Alejandro Schuler[2], Vincent X Liu[2], Gabriel J. Escobar[2], Mike Baiocchi[3]

[1] Biomedical Informatics Training Program, Stanford University School of Medicine, Stanford, CA 94305.
[2] Division of Research, Kaiser Permanente Northern California, Oakland, CA 94612.
[3] Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA 94305.

Abstract: *Background:* Clinical predictive algorithms are often applied to help target interventions to the patients who might benefit most—which, in many settings, are assumed to be those at high risk. However, this assumption, among others commonly held, may not be strictly correct and may lead to suboptimal decision-making. Indeed, despite taking advantage of recent advances in supervised machine learning, many such algorithms remain, at best, blunt instruments, often being developed and deployed without a full accounting of the causal aspects of the prediction problems they are intended to solve. In light of these limitations, we outline an alternative framework that instead applies causal machine learning to predict the causal effects of interventions directly. These effects are coupled to the predicted costs associated with each of the potential outcomes to guide the targeting of these interventions.
Methods: We apply this framework and estimate its overall impact in a retrospective cohort of hospital discharges from Kaiser Permanente Northern California (KPNC) where a risk-based readmissions prevention intervention began implementation in 2016. With data on 1,539,285 inpatients discharged from 21 KPNC hospitals between 6/1/2010 and 12/31/2018, we estimated the conditional average treatment effects (CATE) of this intervention on 30-day readmission using causal forests.
Results: There exists considerable heterogeneity in the effectiveness of this readmissions prevention intervention, particularly on the predicted risk scale. Notably, treatment effects appeared to decrease as risk increased, suggesting a mismatch between these two quantities. The intervention appeared most effective for patients at moderate risk. Moreover, modeling the impact of this intervention as targeted using CATE estimates suggest that as many as 4,500 readmissions could be prevented annually, compared to the 1,200 currently prevented using risk-based targeting.
Conclusions: Population health management interventions could be targeted more efficiently and potentially achieve greater overall impact if their estimated treatment effects were used to guide their targeting.

Statement of Significance: Many, if not nearly all, clinical predictive algorithms in real-world use, especially for population health management and quality improvement applications, assume that outcome risk correlates with treatment benefit. However, we provide empirical evidence that this assumption may not always be true, and attempt to quantify the extent of this mismatch and its implications for patient outcomes. In addition, our results also have potential implications for real-world implementations of these predictive algorithm-driven interventions.

Keywords: Prediction, heterogeneous treatment effects, population health
E-mail of First Presenter: marafino@stanford.edu

Family history and phenotype heritability in type 1 diabetes

Authors: Erin M. Tallon, Mark A. Clements, Danlu Liu, Katrina Boles, Chi-Ren Shyu

Abstract: Type 1 diabetes (T1D) is a complex, immune-mediated chronic condition associated with numerous comorbidities and significant premature mortality. Heterogeneity in T1D onset and progression has confounded efforts to associate discrete phenotypes with disease genotypes. Contrast patterns that point to significant differences in comorbidity status in individuals with a family history of T1D (familial T1D, or F-T1D) versus sporadic T1D (S-T1D) may reflect different disease trajectories for these subgroups and inform personalized risk prediction related to the development of comorbidities and diabetes-related complications.
Existing subgroup discovery approaches lack explainability; but interpretability in biomedical research is key to identifying smaller homogenous subgroups from large heterogeneous populations. We conducted experiments by applying a customized deep exploratory data mining algorithm that identifies highly contrasted, explainable patterns occurring with a significant difference in prevalence between two subgroups (e.g., F-T1D vs. S-T1D). We used an Apache Spark high performance computing environment to analyze publicly-accessible family history, gender, and medical conditions data from the T1D Exchange Clinic Registry.
We identified numerous phenotypic contrasts between F-T1D (n = 2157) and S-T1D (n = 7263) and used Fisher's exact tests to analyze statistical significance. F-T1D associated with hypertension in 31.7% of cases (684/1473), compared to 19.8% of cases (1439/5824) of S-T1D ($p = < 0.001$). F-T1D was associated with hypertension *and* neuropathy in 10.0% of cases (216/1941), compared to 5.1% of cases (375/6888) in S-T1D ($p < 0.001$). F-T1D consistently and significantly associated with a higher prevalence of diabetes-related complications, including cardiovascular, neurologic, and renal disease.
Complex patterns generated with this algorithm may yield contrast patterns that provide insights about personalized interventions to improve health outcomes. These results are part of an on-going effort to investigate family history as a biomarker of risk and to develop personalized risk prediction profiles for individuals with T1D.

Statement of significance: Type 1 diabetes (T1D) is a heterogenous, immune-associated disease characterized by progressive failure or targeted destruction of insulin-producing beta (β) cells in the pancreas. Global incidence of T1D is increasing, and 1.6 million Americans have been diagnosed with T1D. The disease is characterized by marked genotypic and phenotypic variability, but discrete phenotypes resulting from varied environmental and genetic influences remain largely unknown. Family history, a readily-available genomic tool and data source for predicting risk for a range of health conditions, is a cornerstone for the implementation of precision health. We applied a novel subgroup discovery algorithm to family history and medical conditions data in the T1D Exchange Clinic Registry to identify contrast patterns that indicate significant differences in comorbidities associated with familial versus sporadic T1D. These findings highlight an opportunity to use family history data as a genomic tool to explore phenotype heritability in T1D and other chronic health conditions.

Keywords: type 1 diabetes, subgroup discovery, data mining, family history, phenotype
Email of First Presenter: erin.tallon@mail.missouri.edu

Automated Conversational Health Coaching
Authors: Elliot G Mitchell, Lena Mamykina, Columbia University, New York, NY

Abstract: In-person self-management education and *health coaching* are among the most successful interventions for managing chronic conditions. However, there are not enough coaching practitioners to reach the growing population living with chronic diseases. Because coaching takes the form of a conversation between a coach and a client, conversational agents are well-positioned to deliver coaching interventions, and can reach broader and more diverse populations than their in-person counterparts. However, key questions remain in designing effective conversation agent coaches. We sought to explore the extent to which a scripted conversational agent could serve as a health coach for individuals with type 2 diabetes

We designed *t2.coach* through an iterative, user-centered design process. Following the dominant paradigm for conversational agents in health, t2.coach is a scripted chatbot based on an established protocol for goal-setting. Individuals with type 2 diabetes participated in a series of focus groups (N=23), a two-week study with a "wizard-of-oz" prototype (N=13), and a two-week deployment pilot (N=10). We calculated descriptive usage statistics and analyzed participant debrief interviews with thematic analysis. Participants were relatively engaged with the agent, responding to daily messages about 50% of the time. Participants noted that the daily messages and meal-time reminders helped keep them on a schedule, and promoted a continued awareness of their health goals. While some participants reported mixed reactions to the daily frequency of messages from t2.coach, which would sometimes arrive at inconvenient times, many remarked that they appreciated the consistency and patience of the agent throughout daily messaging, for example when it would continue messaging even after it didn't understand a response.

A scripted, rule-based chatbot can lead to engagement and cultivate a coach-like relationship in the context of self-management. We discuss future directions and implications for incorporating more advanced dialog modeling techniques for health coaching.

Statement of Significance: Chronic diseases like type 2 diabetes are growing in prevalence, and there is an increased recognition of the need to focus on longitudinal health and wellbeing. While in-person coaching can be effective, there are not enough coaching practitioners to reach the growing population with chronic conditions, or provide preventative support to healthy individuals. In addition, there are many barriers to accessing in-person coaching, from availability to transportation to scheduling, especially in low resource communities. Individuals with low socio-economic status and ethnic minorities are disproportionately affected by chronic conditions, and the failure of interventions to reach these communities has the potential to deepen existing health disparities.

Keywords: Chronic disease self-management; conversational agents; chatbots; health coaching
E-mail of First Presenter: egm2143@cumc.columbia.edu

Focus Session 1
Text Extraction - Summarization

Improving Physical Activity Among Prostate Cancer Survivors and Their Loved Ones Through a Peerbased Digital Walking Program

Authors: Regina Casanova-Perez MS, Harsh Patel BS, David J Cronkite MS, Savitha Sangameswaran MS, Andrea L Hartzler PhD, University of Washington

Abstract: Objective: Although regular physical activity (PA) improves fitness and vitality of older adults and reduces cancer progression and mortality, only 30% of prostate cancer survivors meet national PA recommendations. Barriers to exercise programs include cost, limited access, and lack of exercise partners, which limit engagement and program impact. Technology-based programs could improve access and reach, but most programs lack social features many cancer survivors desire. We collaborated with prostate cancer survivors and their loved ones to design and evaluate a digital walking program that includes personalized step count goal tracking with Fitbit, peer support of a self-selected walking buddy, and collective support through a private Facebook group.
Materials and Methods: We employed mixed methods to assess the efficacy, engagement, perceived utility, and social influence of our digital walking program through observing program use over 6-weeks, pre-post surveys, and an exit interview.
Results: After 6 weeks of program use, participants (n=18) significantly increased their average daily step count (mean increase of 2,463 steps/day, $p<0.001$). Although engagement and perceived utility of Fitbit and walking buddies was high, engagement and perceived utility of Facebook were limited. Social influence of the program was driven more by group attraction to the collective task of walking than by interpersonal bonds.
Conclusion: Findings demonstrate the feasibility of prostate cancer survivors and their loved ones to successfully increase their PA with a digital walking program designed to meet their needs with low cost, ubiquitous technologies that improve the reach and impact of PA support.

Keywords: physical activity, prostate cancer, fitness trackers, program evaluation, social support
E-mail of First Presenter: reginacp@uw.edu

Extracting Housing Stability Concepts From Clinical Text Notes

Authors: Andrew K. Teng, Adam B. Wilcox, University of Washington

Abstract: Clinical text notes often contain meta information that may not be explicit within the structured data and can create new domains for analysis. We collaborated with a local hospital that has a housing unstable patient population of about 6.5%. Although there have been previous attempts to classify housing stability from text notes, our approach differs as we utilize open source Python packages to test simple text classification methods that can easily generalizable and implemented. We first extracted structured data from acute care patients over a one year timeframe. Once we gathered a patient list, we then extracted social history from various sources, such as Admit and ED notes, over the

past five years for this list of patients. To verify that social history text was being extracted correctly, we performed manual chart review on a random set of ten patients. Once confirmed, we extracted the clinical text and manually labelled the sentiment of the text, treating each entry independently. We extracted 21,876 social history entries, of which 2,408 were manually reviewed. Due to missing data, only 1,785 rows were manually labelled as "housing stable" and "housing unstable", covering 200 unique patients, of which 71 (35%) are scored homeless, and 1,361 unique encounters. Three different models with two different feature selection methods (bag of words (BOW) and bigrams) were used to classify and predict housing stability for the extracted clinical text. From our preliminary analysis, we found slight variation in the accuracy amongst text classifiers. However, there was high accuracy (87.3% to 92.2%). There were many limitations in our preliminary analysis including social factors not present due to patient condition, multiple copy-forward entries (5.7%), and housing stability notes were complex as homelessness qualities were recorded differently amongst providers (e.g. shorthand).

Statement of Significance: Even though clinical text data are often unstructured and harder to analyze, they can contain significant information that can expose new domains for analysis. Rather than creating our own text classifier or installing special software, we used an open source Python package to apply classification methods on clinical text from emergency department and admission notes to classify social features, specifically housing stability, of acute care patients. We found that these methods can be easily implemented and generalized while yielding meaningful results.

Keywords: Text extraction, text classification, housing stability
E-mail of First Presenter: akteng@uw.edu

Developing an Efficient Model for Identifying and Extracting Social Determinants of Health from Clinical Notes

Authors: Rachel Stemerman MPH, Jane Brice MD, Jamie Arguello PhD, Ashok Krishnamurthy PhD, and Rebecca Kitzmiller PhD. University of North Carolina at Chapel Hill

Abstract: Social determinants of health (SDH) are environmental and behavioral factors that are increasingly recognized for their impact on health outcomes. Understanding how to identify the social determinants of health from electronic health records (EHRs) could provide important insights to understand health, disease outcomes, and guide public policies. We developed an efficient methodology to capture SDH among mental health and substance use disorder patients who frequent the emergency department. Our work addresses several challenges to identifying and extracting SDH characteristics from clinical notes including developing standardized terms, exploring new terms using word embedding equerry expansion, and using semi-supervised learning to accelerate the annotation process by identifying sentences likely to document SDH.

Statement of Significance: According to the Agency for Healthcare Research and Quality (AHRQ), mental health disorders were one of the five most costly conditions in the United States with expenditures at $57.5 billion annually. Population-level inequalities in health care result in $309 billion in losses to the economy annually. Previous studies found that over 30% of the highest-costing users of medical (i.e. non-mental health) services had comorbid mental illness or addiction, a rate 3-5 times higher than the lowest-cost users. Many MHSUD patients with persistent unmet health needs become frequent users of

the ED, defined as four ED visits or more during a 12-month period. The 2014 National Survey on Drug Use and Health, the National Institute of Mental Health reported that one in five adults suffer from a mental illness in a given year and receive suboptimal quality of care due to lack of coordination among the various public and private entities necessary to avoid frequent ED visits. Further, MHSUD related ED visits are more than twice as likely to result in hospital admission when compared with ED visits that do not involve MHSUDs. Despite the high costs associated with this population, more than 72% of hospitals still do not have a dedicated budget to support population health initiatives. Even if hospitals had population health initiatives, two-thirds of hospital EHRs lack screening tools for patient's social and behavioral needs.
Outcomes of this study may lead to the development of clinical decision support applications, linking referral services, and inform public policy as recommended by the American College of Physicians, IOM, and AHRQ.

Keywords: Machine learning, natural language processing
E-mail of First Presenter: rstem15@live.unc.edu

Literature-Based Discovery of Drug Repurposing for Rare Diseases
Authors: Daniel N Sosa, Alexander Derry, Margaret Guo, Eric Wei, Connor Brinton, Russ Altman

Abstract: In the modern big data era, researchers have access to an abundance of biomedical data across many modalities. However, despite this plethora of data (of varying quality), extracting knowledge from data, integrating disparate knowledge, and generating new wisdom that translates to clinical practice remains a perennial challenge. In this work, we advance the idea that hypotheses generated using modern statistical inference methods will be more trustworthy and thus more likely translational if they are corroborated by relevant primary source evidence from literature. Further, by extracting knowledge from literature, this work takes advantage of undiscovered knowledge which may be missed by manual curators of structured databases. To achieve this goal, we break down the task into three components: knowledge extraction, knowledge representation, and knowledge inference. In this talk, I will discuss work on a proof-of-concept end-to-end pipeline for generating drug repurposing hypotheses for rare diseases broadly, and I will discuss proposed future work for advancing this paradigm by improving upon the system's three key components.

Statement of significance: Because rare diseases, by definition, affect relatively few people, pharmaceutical companies lack economic incentive to invest in developing a new drug for these cases. As an alternative, many have turned their attention to drug repurposing --off-label use of drugs that are already in the market and are proven to be safe. In this work, we take a new strategy to discovering repurposing for rare disease by extracting knowledge directly from literature. We extract relational information about interactions between drugs, proteins, genes, and diseases, which can be represented as a knowledge graph. In this framework, the task of repurposing is well-posed as a link prediction task in the network. This research is scalable to all of PubMed and enables the incorporation of undiscovered but useful knowledge for this important task.

Key phrases: literature-based discovery, knowledge graphs, drug repurposing,
E-mail of First Presenter: dnsosa@stanford.edu

Learning to Identify Synonymous Terms Across Biomedical Terminologies

Authors: Vinh Nguyen, Olivier Bodenreider, National Library of Medicine

Abstract: In the UMLS Metathesaurus, synonymous strings from biomedical source vocabularies are assigned the same concept unique identifier (CUI). The 2019AA release of UMLS Metathesaurus comprises approximately 14.6 million biomedical names from 210 source vocabularies grouped into 3.85 million concepts. Given the size of the Metathesaurus, the manual process of inserting new resources to the existing Metathesaurus can be costly, time-consuming, and error-prone as identifying the concept similarity among these atom strings heavily relies on the lexical knowledge, semantic preprocessing, and the expertise of human editors.

This project explores a deep learning (DL) approach to learning the synonymy and non-synonymy between UMLS terms using a Siamese network architecture with a shared model from Long Short-Term Memory and Convolutional Neural Network with BioBERT and BioWordVec. Positive/negative samples are taken from the pairs of terms having the same/different CUI. To automatically identify synonymy among terms, we enrich the terms with information available to the UMLS editors, including the source synonymy (i.e., synonyms provided by source vocabularies), hierarchical context information (terms for parent and child concepts in a source vocabulary), and the high-level semantics (semantic group) of a source vocabulary.

We created a mixed dataset with about 80 million pairs of biomedical terms generated from three different distributions: (1) 25 million pairs of terms with high lexical similarity with Jaccard >= 0.6, (2) 24.3 million pairs with 0 < Jaccard < 0.6, and (3) 25 million randomly selected pairs. We divide this mixed dataset with the ratio of 60% for training, 20% for testing, and 20% for validation.

We compare the performance of this DL approach with a lexical rule-based approach that approximates the current Metathesaurus building process. Our initial experiments show that the DL approach outperforms the lexical rule-based approach in both overall accuracy (88.5% vs. 86.15%) and precision (78.47% vs. 65.86%). While these preliminary results are encouraging, further investigation is needed for developing a deep learning-based process for automatic construction of the ULMS.

Keyword: UMLS Metathesaurus, NLP Deep learning, Text Semantic Similarity

E-mail of First Presenter: vinh.nguyen@nih.gov

Focus Session 2
HIE - Interoperability

Identifying Acute Respiratory Distress Syndrome in De-Identified EHR Data

Authors:
V Eric Kerchberger, Julie A Bastarache, Ciara M Shaver, J Brennan McNeil, Chunxue T Wang, Ashish Tapdiya, Neil S Zheng, Lorraine B Ware, Wei-Qi Wei, Vanderbilt University.

Abstract: Acute respiratory distress syndrome (ARDS) remains an important challenge in the intensive care unit, but our ability to study the biology of this disease has been limited to samples from moderately-sized cohorts or trial populations. Population-based biobanks tied to electronic health records (EHRs) offer the opportunity to acquire larger sample sizes for more powerful studies, however traditional manual approaches to identify ARDS will not efficiently scale to electronic health record (EHR) databases containing millions of records. We are developing high-throughput methods to identify ARDS patients using a combination of structured and unstructured EHR data elements that follow the Observational Medical Outcomes Partnership (OMOP) Common Data Model.
Using previously published data and domain expertise, we developed a decision tree classifier using a combination of administrative billing codes (ICD-9-CM, ICD-10-CM, and CPT), laboratory studies, and chest radiograph (CXR) report text to capture key elements of the Berlin Criteria, the internationally accepted research definition of ARDS. We assessed classification performance using in 125 randomly chosen subjects (50 classifier cases, 75 classifier controls) from the VUMC de-identified research EHR. The current classifier PPV was 0.66 (95% CI: 0.51 – 0.78), sensitivity 0.94 (0.79 – 0.99), and balanced F_1 score 0.78 (0.67 – 0.89,) compared with physician review as the reference standard. Manual review of misclassified subjects suggest that false-positive CXR reports are the major cause of misclassification, primarily due to common radiographic ARDS mimics including atelectasis (lung or lobar collapse) and cardiogenic pulmonary edema. Ongoing work is directed at improving CXR report interpretation using natural language processing techniques, and incorporating concepts capturing cardiovascular disease risk.

Statement of Significance: Reliable identification of ARDS cases in de-identified EHR databases may be feasible with a framework that uses both structured and unstructured EHR data. Current challenges are focused on improving chest radiograph report interpretation and differentiating cardiogenic causes of pulmonary edema from ARDS. These findings will guide future studies aimed at identifying ARDS in EHR databases.

Keywords: De-identified Electronic Health Records, Phenotyping, Critical Care
E-mail of First Presenter: vern.e.kerchberger@vumc.org

Representation of Medical Knowledge Towards Generalizable Prediction

Authors: Sarah Mullin, Frank LeHouiller, Werner Ceusters, and Peter Elkin, University at Buffalo

Abstract: Electronic health record (EHR) data can provide useful information for prediction algorithms in populations not likely to enroll in clinical trials, precision-based clinical decision support, and as a basis for prospective research. To use a prediction model for future clinical practice and implementation, a model must be generalizable. However, models often fail in different environments than the training environment due to lack of observations or a limited patient sample [1]. A Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) representation embedding is proposed to facilitate generalizability, utilizing both the terminological structure of SNOMED CT and learned relationships from annotated text. Embedding node and edge graph structure is difficult for medical terminologies. Previous work, like Node2vec [2], maximizes the likelihood of preserving network neighborhoods of nodes for homogeneous networks, while edge2vec [3]

extends node2vec to learn node embeddings for heterogeneous networks taking into account edge semantics. Our embedding methodology builds off of these existing graphs embedding strategies, using different sources of terminological and medical knowledge. First, the embedding uses hierarchical *is a* relations present in the SNOMED CT terminology. In addition, statistically and clinically relevant relations, such as *diagnosed_by*, *has_treatment*, *risk_factor,* and *causes,* found by entity annotation and relation extractions in 2,009,834 Pubmed/Medline citations and 1,013 MedlinePlus Health Topic Summaries, are incorporated. These relations are used to build directed adjacency matrices between the SNOMED CT entities for disorders and clinical findings, procedures, events, morphologic abnormalities, observable entities, and organisms. Finally, 53,432 MIMIC-III patients and 96,681WNY outpatients are used to learn edge weights of the directed adjacency matrices. The final model is a joint variational graph autoencoder with five relation motivated adjacency matrices as input, which are combined with a shared loss function to have a final overall embedding that can be used in subsequent prediction models.

Statement of Significance:
The resulting embedded representation can be useful in terms of transfer learning, where an embedding trained on multiple sources can be used to learn an embedding over a smaller corpus of patient data. Using a pre-initialized embedding matrix has been shown to increase model predictive accuracy and increase generalizability of the algorithm in different samples. Finally, our embedding has the added benefit that the pre-initialized embedding allows you to fine-tune weights supported by known medical and ontological information.

Keywords: embedded representation, ontology, deep learning, SNOMED CT
E-mail of First Presenter: sarahmul@buffalo.edu

1. Toll, D., et al., *Validation, updating and impact of clinical prediction rules: a review.* Journal of clinical epidemiology, 2008. **61**(11): p. 1085-1094.
2. Grover, A. and J. Leskovec. *node2vec: Scalable feature learning for networks*. in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 2016.
3. Gao, Z., et al., *edge2vec: Representation learning using edge semantics for biomedical knowledge discovery.* BMC bioinformatics, 2019. **20**(1): p. 306.

Informatics of Early Intervention referrals for Children with Disabilities

Authors: Ben Sanders, Joan Ash, Katie Zuckerman, and Paul N. Gorman, Oregon Health & Science University

Abstract: This study describes the current state of referral of children to Early Intervention, including what information is exchanged and the barriers and facilitators of high quality referrals. Our mixed-method approach includes: 1) document analysis of states' EI referral forms; 2) survey of state EI leaders on their referral methods; and 3) qualitative analysis of interviews with a subset of responding EI leaders.
At the time of writing, EI forms were collected from 23 states, 30 states are represented in the survey, and interviews are ongoing. Referral forms usually had fields for family contact information, reason for referral, and parental release of information. Healthcare or other professionals are the apparent intended users, and family perspectives are usually absent. On the survey, EI agency leaders estimate that 5 to 100% of their referrals are received using a standardized form, and that healthcare is the most common source (51%, SD 19), with families second. They frequently receive referrals by phone (mean proportion 38%), fax (28%), email (18%), and other electronic (14%). Most states (93%) use state-wide EI data systems which do not or cannot use Health Level

Seven messaging standards (88%). Few data systems receive structured electronic referral data (11%) or metadata (11%). Interview themes include a move towards centralized data management featuring web-based referral entry, the importance of verbal communication between EI staff and the family, and plans to expand role-based security and digitize authentication for parental consent.
Despite current EI agency transitions to centralized information systems, achieving interoperability with medical information systems is not a near-term goal. Although traditional referral forms aren't oriented towards parents, verbal communication with the child's caregiver is a critical component for EI referral staff. This suggests opportunities for process improvement using mobile technology to reach parents.

Statement of Significance:
Early identification and treatment (Early Intervention, EI) of the nearly 1 in 10 children with developmental delay can maximize their life success and pay dividends to society. But up to half of children who would benefit do not receive needed services. Siloed information systems contribute to this problem, widen health disparities, and reinforce cultural gaps between professional EI settings and referring parties in primary care. Effective solutions will require informatics insight with careful consideration of the uses and meanings of information involved.

Keywords: Interoperability, Developmental disabilities, Pediatrics, Health Services, Care Coordination
E-mail of First Presenter: sanderbe@ohsu.edu

Health Information Exchange Use During Dental Visits
Authors: Heather L Taylor, Nate C Apathy, & Joshua R. Vest, Indiana University

Abstract: Dental and medical providers require similar patient demographic and clinical information for the management of a mutual patient. Despite an overlap in information needs, medical and dental data are created and stored in multiple records and locations. Electronic health information exchange (HIE) can connect siloed professionals and bridge gaps in health data spread across various providers. One form of HIE – query-based HIE – is particularly well suited to meet missing information needs of dental providers, as it does not generally require implementation of new software within practice settings. Enabling exchange via query-based HIE may provide critical information at the point of care and reduce information discrepancies in health records. The purpose of this study is to characterize query-based HIE use during dental visits at two Federally Qualified Health Centers (FQHCs) that provided on-site dental services. First, we determine the proportion of dental visits for which providers accessed the HIE. Next, patient and visit characteristics associated with query-based HIE use during dental visits are examined. Last, among dental visits with HIE use, the aspects of the HIE that are accessed most frequently are described. Findings from this study extend the body of research examining HIE use by studying a less explored area of the care continuum – dental.

Statement of Significance (Up to 150 words): Recent federal policies, such as the Health Information Technology for Economic and Clinical (HITECH) act and the 21st Century Cures Act, were implemented to incentivize electronic health record adoption and motivate exchange of electronic health information between all health providers, including dentists. Given such federal pressures and the potential value of shared medical and dental data, it is important to determine whether dental providers use HIE and what aspects of an HIE are valuable to a dental provider.

Keywords: Dental Informatics, Health Information exchange, HIT policy
E-mail of First Presenter: hhavens@iu.edu

A SMART on FHIR and CDS-Hooks Enabled Approach to the Exchange and Review of Genomic Test Results
Authors: Michael Watkins, Karen Eilbeck, The University of Utah

Abstract: The reports of genomic laboratory tests are a common source of clinical genomic data. However, these reports are complex and vary in structure depending on the reporting laboratory. They are also are typically stored as PDF files. This system of creating and exchanging genomic data points is antiquated and presents significant downstream limitations in data utility and reuse. Emerging clinical and technical standards have presented a new paradigm of exchanging laboratory result data and are particularly promising for genomic tests. New FHIR genomic reporting profiles make it possible to represent these test results as computable resources rather than as static PDF fields. SMART on FHIR also makes it possible to exchange them in an interactive environment that is open and compatible with relevant CDS services. CDS-Hooks can also allow these test results to be computationally pulled into all relevant future clinical scenarios rather than requiring a future clinician to open and review an old PDF file. This new approach to exchange and review of genomic test results is a valuable pilot demonstration of how to harness these various interoperability solutions.
Interoperability should be one of the most central objectives for any clinical informatics innovation. However, these new standards can be technically complex and difficult to approach. This project provides a straightforward use-case that is easy to follow and whose scope covers several relevant subtopics within these standards. It also emphasizes the importance of an informatics-based approach of creating and exchanging standardized data, which all sectors of healthcare should prioritize.

Keywords: Clinical genomics, SMART on FHIR, CDS Hooks, interoperability
E-mail of First Presenter: michael.watkins@utah.edu

Focus Session 3
Bioinformatics and Genomics

Vitessce: Visual Integration Tool for Exploration of Spatial SIngle-Cell Experiments
Authors: Trevor Manz, Chuck MacCallum, Ilan Gold, Peter V. Kharchenko, Nils Gehlenborg, Harvard University

Abstract: Vitessce (visual integration tool for exploration of spatial single-cell experiments) is an open-source, web-based viewer for spatial single-cell -omics data. With the rise of single-cell methods, consortium projects like NIH Human BioMolecular Atlas Program (HuBMAP) and the Human Cell Atlas are leading efforts integrate these technologies, creating complex datasets and new challenges for visualization. Specifically, these data include methods which detail both relationships in 2D/3D physical space as well as relationships in high-dimensional space (i.e. gene expression).
Vitessce (http://vitessce.io/) is designed with a modular architecture so that it can be used as a standalone viewer while offering the ability to reuse components in other projects, such as the HuBMAP Data Portal. We recently built a multimodal image viewer for data generated by our collaborators in the Spraggins' Lab at the Vanderbilt University BIOmolecular Multimodal Imaging Center (BIOMIC). The group has developed

experimental methods which fuse data from different imaging technologies on the same tissues, including IMS and high-resolution multiplex immunofluorescence microscopy.
The datasets collected span a range of spatial resolutions and will allow for the construction of comprehensive molecular atlases in human tissues. Current imaging viewers are limited in their ability to view these modalities in context. Since Vitessce is modular by design, we are developing components which support interactions to index these modalities, allowing users to select ranges of mass-to-charge (m/z) ratios specific for the IMS data as well as selecting for different biomarkers used in microscopy. We combine these selections into a single view or separate linked pan-and-zoom views for displaying these data in context. Using segmentation information derived from biomarkers in targeted microscopy, we can partition the IMS data into subsets and explore which proteins, lipids, or other biomolecules are associated with specific cell and tissue types in an untargeted manner.

Statement of Significance: Many of the current visualization tools in this space require substantial server-side resources or for large datasets to be stored locally. We've leveraged recent advances in web-based technologies, including GPU programming and HTTP/2, to enable performant real-time computation and rendering on the client. In addition, we've designed Vitessce to work directly with remote cloud storage, such as Amazon S3 and Google Cloud Storage, which enables researchers to readily access large datasets simply by sharing a URL. These design decisions were crucial to create a tool which can provide nuanced biological insight to complex multimodal datasets while offering the ability to embed these findings across the web and share widely.

Keywords: Multimodal imaging, Data Visualization, Single-cell
E-mail of First Presenter: trevor_manz@g.harvard.edu

Integration of Large-Effect Expression Variants Enhances Polygenic Risk Prediction
Authors: Craig Smail; Matthew Aguirre; Manuel Rivas; Stephen Montgomery

Abstract: Polygenic risk scores (PRS) aim to quantify the contribution of multiple genetic loci to a complex trait. However, existing PRS estimate genetic liability using common genetic variants excluding the impact of rare variants. To assess the impact of rare variants on complex traits and PRS predictions, we focused on large-effect expression (outlier) rare variants found in GTEx that were also present as high-quality UK Biobank (UKBB) imputed rare variants. Across multiple UKBB GWAS, we observed larger phenotypic effects for outlier rare variants compared to control variants, increasing with the degree of outlier severity. By integrating polygenic risk score (PRS) weights from publically-available PRS for BMI near outlier genes, we observed large deviations in PRS-predicted phenotype in the UKBB; for example, individuals in PRS decile 1 ("low-risk") but in the top percentile of outlier gene burden, have a rate of severe obesity of 3.89%, whereas the rate is 0.63% using PRS alone. We replicated our findings using data from the NHLBI Trans-Omics for Precision Medicine (TOPMed) biobank. This work demonstrates that multiple PRS will benefit from the inclusion of rare genetic variants.

Statement of significance: Polygenic risk scores are increasingly being used to stratify individuals in to risk groups for a variety of traits and diseases. These scores are constructed using common genetic variants, and ignore the effects from rare, large-effect variants. We show that rare variants associated with gene expression outliers lead to substantial deviations in risk for body mass index, improving the prediction of risk for many individuals.

Keywords: polygenic risk scores; rare variants; RNA-sequencing
E-mail of First Presenter: csmail@stanford.edu

Identifying Protein-Metabolite Networks associated with COPD Phenotypes

Authors: Emily Mastej, MS[1], Lucas Gillenwater, MPH[2], Yonghua Zhuang, PhD[3], Katherine A. Pratte, PhD, MSPH[2], Russell P Bowler, PhD, MD[2,4], Katerina Kechris, PhD[3]

Abstract: Chronic obstructive pulmonary disease (COPD) is a disease in which airflow obstruction in the lung makes it difficult for patients to breathe. Although COPD occurs predominantly in smokers, there are still deficits in our understanding of the additional risk factors in smokers. To gain a deeper understanding of the COPD molecular signatures, we used SmCCNet, a recently developed tool that uses sparse multiple canonical correlation analysis, to integrate proteomic and metabolomic data from the blood of 1008 participants of the COPDGene study to identify novel protein – metabolite networks associated with lung function and emphysema. Our aim was to integrate -omic data through SmCCNet to build interpretable networks that could assist in the discovery of novel biomarkers that may have been overlooked in alternative biomarker discovery methods. We found a protein- metabolite network consisting of 13 proteins and 7 metabolites which had a -0.34 correlation (p-value = 2.5×10^{-28}) to lung function. We also found a network of 13 proteins and 10 metabolites that had a -0.27 correlation (p-value = 2.6×10^{-17}) to percent emphysema. The most heavily weighted edge in the lung function network connected troponin T and phosphocholine, both of which were network hubs. This network also included features such as C-reactive protein and mannose binding protein and complement suggesting a stronger association with inflammation and heart strain. Alternatively, the network associated with percent emphysema had features such as growth hormone receptor, adipokines, amino acids, and lipids suggesting that growth and metabolism may play a more important role in the pathogenesis of COPD. Protein – metabolite networks can provide additional information on the progression of COPD that complements single biomarker or single -omic analyses.

Keywords: Proteomics, Metabolomics
Email of First Presenter: emily.mastej@cuanschutz.edu

Alternative Splicing Event Signatures of Drug Response in AML

Author: Julian Egger, Oregon Health & Science University

Abstract: Abnormal splicing events can promote drug resistance in a variety of cancer types, however, the full extent of genome-wide splicing in therapy-resistant AML is not fully characterized. Further, expression changes of splice variants do not occur independently, but rather in a coordinated fashion in order to maintain proper cellular function. Therefore, in order to understand how aberrant splicing confers drug resistance in AML, alternative splicing needs to be studied on a systems-wide level. This research utilizes a network-based approach to characterize coordinated alternative splicing variation and identify potential splicing event signatures of drug response in AML.

Statement of Significance: This project will provide a systems-level characterization of coordinated splicing changes promoting therapeutic resistance and identify potential splicing event signatures of drug response in AML.

Keywords: Bioinformatics, Genomics, Cancer Biology
E-mail of First Presenter: eggerj@ohsu.edu

Heritable Components Explain Variation in Acute Kidney Injury Recovery Rate
Authors: Kathleen LaRow Brown, Nicholas Tatonetti, Columbia University

Abstract: Acute kidney injury (AKI) impacts an estimated 13.3 million people each year, yet no drugs are available to promote renal recovery. Comparing patients with differential times to recovery could highlight important influencers of the recovery process. Genetic studies specifically could highlight therapeutically relevant molecular pathways, but only if the recovery process is influenced by genetic variation. In this study, we used first of its kind electronic health record (EHR) linked pedigree data to estimate the narrow-sense heritabilities of time to improvement, time to recovery, and rate of recovery. We used previously recorded serum creatinine data to identify patient AKI events and determine renal recovery, and estimated the narrow-sense heritabilities using linked pedigrees. We found that over one third of patient variation in AKI rate of recovery is explained by heritable components. Our study highlights the importance of exploring AKI rate of recovery in future genetic studies with the target of identifying therapeutically relevant molecular pathways.

Statement of Significance: Treatment for acute kidney injury (AKI) recovery remains largely supportive despite improved understanding of renal injury and repair pathways. The diversity of recent clinical trials highlights the lack of consensus on the best mechanisms to target to promote renal recovery. Genetic studies using patients with differential rates of recovery could provide clarity by highlighting therapeutically relevant pathways if rate of recovery is heritable. In this study, we show that heritable components do impact AKI rate of recovery.

Keywords: Heritability, Acute Kidney Injury, Translational Informatics
E-mail of First Presenter: kel2158@cumc.columbia.edu

Investigating Somatic Mosaicism in Blood for Cardiovascular Disease Risk
Authors: Varuna Chander, Aleksander Milosavljevic, Richard Gibbs, Baylor College of Medicine

Abstract: Common diseases constitute a significant health burden, affecting a large segment of the population and the leading cause of death in the United States. Despite this impact, the underlying genetic causes of common diseases like cancer and cardiovascular disease (CVD) have been poorly understood. Advances in sequencing technologies have facilitated the discovery of somatic driver mutations in the genome that cause clonal expansion of blood cells - leading to a surprising discovery - it is associated with significantly increased risk for blood cancer and CVD. This novel biological process is termed Clonal Hematopoiesis of Indeterminate Potential (CHIP). Remarkably, CHIP is common with about 10-20% of 65-year olds having conspicuous clonal dominance.
Recent studies have also shown that large proportion of CHIP individuals have unknown underlying causes. This suggests the possibility of other active mechanisms being missed with a potential role in CHIP. Particularly, the

effect of DNA methylation in CHIP remains currently unexplored considering frequently mutated CHIP genes are involved in DNA methylation dynamics. Moreover, CVD genetics research has focused mostly on inherited genetic variation. However, we hypothesize that non-inherited acquired mutations in the mendelian CVD genes, when mutated in blood cells, may differentially influence disease in the setting of CHIP. Overall, mechanisms that shape clonal expansions in healthy individuals are not entirely understood, or how they ultimately lead to disease.
To advance understanding, we apply computational methods to perform somatic variant detection within the CVD enriched clinical cohorts at the Human Genome Sequencing Center. Using somatic variant callers, we identify recurrently mutated CVD genes with a potential role in CHIP. We further investigate the downstream pathways to understand what alterations are happening in blood cells and disease risk. Furthermore, we propose to investigate whether there exists an interplay between somatic mutations, methylation and inherited mutations between individuals with CHIP.

Statement of Significance: Cardiovascular disease (CVD) is late-onset condition and risk increases with age. Although age is the dominant risk factor, the mechanistic basis for why age predisposes to CVD is not entirely understood. Besides, more than 60% of patients with CVD have either zero or one conventional risk factors. Through our proposed research, we address an important biological link between aging, CVD, cancer, with tremendous potential impact on the public health of aging populations.
Our research will provide novel data on the mechanisms that cause CHIP and is critical for developing strategies relevant to the prevention of both CVD and cancer, the two common causes of human mortality. Further, it will serve as a resource for future studies that aim to elucidate the association between CHIP and common age-related disorders. Ultimately, this knowledge will benefit biomedical scientists, clinicians and patients where insights will inform clinical care in the aging population.

Acknowledgement: This work is supported by the NLM Training Program in Biomedical Informatics and Data Science T15LM007093, Program Director Dr. Lydia Kavraki.

Keywords: Translational bioinformatics, clonal hematopoiesis, somatic mutations, cardiovascular disease risk.
E-mail of First Presenter: Varuna.Chander@bcm.edu

ABSTRACTS
Tuesday, June 23, 2020

Plenary Session 2
Translational Bioinformatics

Identifying MHC Class I Binding Peptides Using an Ultradense Peptide Array

Authors: Amelia K Haj, David A Baker, Meghan E Breitbach, David H O'Connor; University of Wisconsin – Madison

Abstract: Rational vaccine development and evaluation requires identifying epitope-specific CD8 T cell responses, a process that is streamlined by screening pathogen-derived peptides for their ability to bind MHC class I molecules. Computational methods for predicting peptide-MHC binding affinity exist, but these methods rely on relatively small peptide binding datasets generated by conventional in vitro assays. We have developed and tested an ultradense peptide array as a tool for measuring peptide-MHC binding for hundreds of thousands of peptides simultaneously, which can be used to directly identify binding peptides and generate binding motif predictions. To validate the peptide array as a method to identify CD8 T cell epitopes, we assessed the binding of four well-characterized rhesus macaque MHC class I molecules to 61,000 peptides derived from the full proteomes of 82 simian immunodeficiency virus (SIV) and simian-human immunodeficiency virus (SHIV) isolates. For one MHC molecule, Mamu-A1*001, the array identified 64% of the known CD8 T cell epitopes in SIV in the top 192 highest-binding peptides for that virus. We also successfully recapitulated the established binding motif for Mamu-A1*001 using an array covering every possible amino acid substitution for two known epitopes. To eliminate the need for a substitution array, we employed a decision tree approach, using partition models to analyze the full peptide binding sets for each of 16 MHC molecules and identify optimum splits in the data to generate predicted binding motifs for each. These largely reproduced known binding motifs for the MHC molecules that had been previously characterized. We anticipate that this high-throughput identification of peptides that bind class I MHC and rapid characterization of binding motifs will enable more efficient CD8 T cell response profiling for vaccine development, particularly for pathogens with complex proteomes and few defined epitope-specific responses.

Keywords: Immunology, virology, bioinformatics
E-mail of First Presenter: haj@wisc.edu

Elucidating Bio-Molecular Mechanisms of Disease for Novel Pathogenic Variants Using Structural Bioinformatics Approaches

Authors:
Rolando Hernandez, Craig Teerlink, Lisa Cannon-Albright, Christopher Li, Douglas Grossman, and Julio C. Facelli, The University of Utah

Abstract: Innovative in silico approaches to elucidate molecular mechanisms of disease are necessary to overcome the structural knowledge gap. Structural bioinformatics offers a promising way to study the effects of novel mutations on protein structure and function. In this study, four novel mutations in CDKN2A/p16, MEGF6, ERF, and CELF4 were investigated. This research was done in collaboration with the Huntsman Cancer Institute and the Division of Genetic Epidemiology at the University of Utah. Firstly, after sequencing for CDKN2A/p16 mutations in a melanoma patient with a history of four primary melanomas, a novel missense variant (L117P) was found. Secondly, after candidate gene prioritization across 10 high-risk pedigrees, a novel missense variant (C200Y) in MEGF6 showed evidence of segregation with osteoporosis. Thirdly, after candidate gene prioritization across 14 high risk pedigrees, a novel missense variant (P349L) in ERF showed association with bladder cancer. Finally, after candidate gene prioritization across 6 early diagnosed colon cancer cases and 3 other cancer cases, a novel intronic variant (chr18: 34932587 (ref=T, alt=A)) in CELF4 showed association with the phenotype. Here we demonstrate how using 3D protein and RNA structure prediction, as well as related structural bioinformatics techniques such as binding site analysis, computational phenotype prediction, and structure visualization, it is possible to develop promising hypotheses for molecular mechanisms of disease.

Therefore, in this study we demonstrate the utility of structural bioinformatics techniques in translational biomedical research that could be used for future target discovery and patient diagnostics.

Having the capacity to study biomolecular mechanisms of disease using innovative in silico approaches will help elucidate them. Currently, there is a structural knowledge gap that is being addressed using computational approaches and experimental methods, such as cryo-EM. Taking advantage of the availability of refined force fields and high computational power allows for using a structural bioinformatics approach. Thus, we are able to feasibly investigate more subtle hypotheses for biomolecular dysfunction that could be used to inform clinicians and pharmaceutical scientists.

Keywords: Protein structure, RNA structure, Structural bioinformatics, molecular mechanisms
E-mail of First Presenter: Rolando.Hernandez@utah.edu

Prognostic Value of Imaging-Based Estimates of Glioma Pathology Pre- and Post-Surgery

Authors: Evan D H Gates[1,2], Dima Suki[3], Jeffrey S Weinberg[3], David Fuentes[1], and Dawid Schellingerhout[4]
Department of Imaging Physics, University of Texas MD Anderson Cancer Center
The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences
Department of Neurosurgery, University of Texas MD Anderson Cancer Center
Departments of Neuroradiology and Cancer Systems Imaging, University of Texas MD Anderson Cancer Center

Abstract: We developed a machine learning model for predicting proliferative activity within gliomas using MRI and evaluate the potential to improve patient survival by guiding resection.
To do so, we trained a random forest model for predicting the proliferative index (% Ki-67 expression) using spatially correlated (<1 mm) biopsy samples and routine MRI sequences: T1w (pre and post-contrast), T2w, and FLAIR. The final model was applied to two independent patient cohorts with known outcomes: 1) a collection of high-grade glioma cases publicly available through the 2018 BraTS challenge and 2) a group of previously untreated glioma patients from our institution.
Preoperative images were co-registered and normalized using reference tissue intensities. The random forest was applied voxel-wise through the visible tumor volume. For cohort 2), the same image analysis was also performed on postoperative images acquired within two days of surgery. Survival differences based on image features computed from the proliferative index maps were correlated with survival using a Cox model and logrank test.
The Random Forest trained on the four imaging sequences had root-mean-square error of 5.4% for predicting Ki-67. For the 140 cases in cohort 1 and 68 cases in cohort 2, max Ki-67 thresholds of 28.2% and 24.75% optimally divided the cases by survival with hazard ratios of 1.66 and 3.32 respectively (both $p<0.05$). In cohort 2, a high preoperative estimated Ki-67 with a low postoperative maximum Ki-67 (i.e. when high Ki-67 tumor was removed) was favorably prognostic compared to high Ki-67 pre- and post-op ($HR=0.329$, $p<0.05$).
Estimates of glioma proliferation based on routine brain imaging are predictive of survival. Initial results also suggest that patients with highly proliferative regions removed showed better outcomes, supporting the use of the model to guide surgical intervention. Future work will evaluate the potential benefit in the context of other clinical factors.

Statement of Significance: Spatial heterogeneity in gliomas poses a major challenge to glioma treatment. Our model uses spatially specific training data to estimate cellular proliferative activity at each voxel and avoids this

limitation that is often present in image-based analysis. We are able to accurately estimate tissue pathology characteristics using MR imaging alone, which adds value to imaging data. This is reinforced by the correlations between estimates of pathology and survival outcomes. Using two independent cohorts (one publicly available, one from our institution) for the survival analysis shows robustness of the maximum Ki-67 estimate as a prognostic biomarker.
Our data also supports the use of graphical maps of proliferative activity as an interventional aid. Areas of heightened Ki-67 present a clear target for treatment and could supplement existing surgical techniques.

Acknowledgement:
This work is supported by the NLM Training Program in Biomedical Informatics and Data Science T15LM007093, Program Director Dr. Lydia Kavraki.

The data for this work have been obtained through a search of the integrated multidisciplinary Brain and Spine Center Database. The Brain and Spine Center Database was supported in part, by an institutional M. D. Anderson database development grant.

Keywords: Magnetic Resonance Imaging, Machine Learning, Glioblastoma
E-mail of First Presenter: EGates1@mdanderson.org

An Empirical Bayes Approach to Image Normalization in Highly Multiplexed Cellular Imaging Data
Authors: Coleman R Harris, Eliot McKinley, Joseph Roland, Ken Lau, Robert Coffey, Simon Vandekar, Vanderbilt University

Abstract: High-dimensional multiplexed imaging methods can help quantify the heterogeneity of cell populations in healthy and tumorous tissues. This, in turn, can offer insight into tumor progression and potentially improve treatment strategies. However, implicit biases exist in the imaging pipeline because images are distorted by optical effects, slide and batch effects, and instrument variability. The challenge in normalizing multiplexed imaging is compounded by the number of channels and natural tissue variability within each field of view. This can bias downstream analyses by introducing systematic differences in image intensity that impacts inference. In this work, we introduce an image normalization pipeline to reduce systematic variability in multiplexed image intensity. We build on existing methods to reduce batch effects in genetic data, and develop an Empirical Bayes normalization procedure to model and remove variability of the image intensity that is related to slide effects. We demonstrate these methods by analyzing over 1000 multiplexed immunofluorescence (MxIF) images of mouse colon tissue samples to study the reduction in variability. In future work we will evaluate our pipeline by measuring the reduction in slide-to-slide variation, while retaining or maintaining biological variability.

Statement of Significance: As multiplexed imaging methods develop and improve, incorporating normalization into the research pipeline will be vital to ensuring downstream analyses are accurate. The application of our research is closely tied to research and identification of colorectal cancer. Hence, ensuring that images and data are valid is a key component when applying this area of research into clinical practice.

Keywords: Image Normalization, Imaging Data, Big Data
E-mail of First Presenter: coleman.r.harris@vanderbilt.edu

Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings

Authors: David Chang, Daniel Chawla, Cynthia Brandt, Andrew Taylor, Yale University

Abstract: Much of biomedical and healthcare data is encoded in discrete, symbolic form such as text and medical codes. There is a wealth of expert-curated biomedical domain knowledge stored in knowledge bases and ontologies, but the lack of reliable methods for learning knowledge representation has limited their usefulness in machine learning applications. While text-based representation learning has significantly improved in recent years through advances in natural language processing, attempts to learn biomedical concept embeddings so far have been lacking. A recent family of models called knowledge graph embeddings have shown promising results on general domain knowledge graphs, and we explore their capabilities in the biomedical domain. There are also no established benchmarks for comparing concept embeddings or in-depth discussions of best practices. In this paper, we train several state-of-the-art knowledge graph embedding models on the SNOMED-CT knowledge graph, provide a benchmark with comparison to existing methods as well as discussions on best practices, and make a case for the importance of leveraging the multi-relational nature of knowledge graphs for learning biomedical knowledge representation. The embeddings, code, and materials will be made available to the community.

Keywords: Biomedical natural language processing, knowledge graph embeddings, SNOMED-CT
E-mail of First Presenter: david.chang@yale.edu

Prediction of Drug Co-Prescription Induced Adverse Events Using CANDO

Authors: Zackary Falls, Sarah Mullin, Peter Elkin, Ram Samudrala, University at Buffalo

Abstract: A main challenge in drug discovery is to predict and manage drug-drug interactions (DDIs) and the severity of adverse events (ADEs). Clinically significant DDIs can be hard to find due to the low power of clinical trials to study DDIs with detection of important safety issues rare within the process. In addition, dose dependence and the natural heterogeneity of the population regarding conditions and demographic features can lead to rare or delayed on set identification of ADEs. Therefore, monitoring the safety of drugs post-approval and post-marketing is important for public safety. Herein, we present a hybrid bio-and clinical-informatics approach to the prediction of DDI ADEs. We have added DDI prediction functionality to our drug discovery platform, CANDO (Computational Analysis of Novel Drug Opportunities). The new protocol has been benchmarked against a known set of DDI adverse events extracted from Twosides, for which we get a top10 accuracy of 55.3% (34.7% expected at random) across 1,303 ADEs and 18,525 drug-pairs. Despite this high accuracy, the benchmark is based upon known DDI ADEs and are not novel predictions, therefore they are not clinically useful. We plan to apply this platform in conjunction with EHR data from a local Western New York cohort to validate novel predictions; we will retroactively identify patients in our sample whom have experienced the adverse events while being co-prescribed the corresponding drug pair within a given time window of the event. The results from this study are directly translatable to clinical prescribing practices: by identifying drug combinations which result in high risk of negative outcomes, clinicians can be informed and avoid these previously unknown, potentially life-threatening coprescriptions.

Statement of Significance: We have modified our drug discovery platform, CANDO, to accurately predict drug-drug interaction(s) and corresponding adverse event(s). Subsequently, electronic health record data, from a cohort of patients, will be used to retrospectively validate the predictive power of the modified platform.

Keywords: Drug-drug interactions; Adverse events; Translational research
E-mail of First Presenter: zmfalls@buffalo.edu

A Reward System Polygenic Risk Score for Predicting Obesity and Substance Use Disorders

Authors: Kristen M. Stevens, BA, Daniel L. Marks, MD, PhD, Joyanna Hansen, PhD, RD, Shannon K. McWeeney, PhD, Oregon Health & Science University

Abstract: Genome-wide polygenic risk scores (PRSs) can now predict complex genetic disease risk with nearly the same ability as tests for monogenic diseases. However, there are substantial barriers to implementing such tests in the clinic, most notably how exactly to incorporate PRSs with other known modifiable and non-modifiable risk factors. Some of the most promising diseases for early PRS adoption, such as coronary artery disease and type 2 diabetes, also share a common set of robust modifiable risk factors that includes diet-induced obesity and substance use. However, the cause of both obesity and substance use disorders is at least partially genetic in most individuals. Moreover, intriguing evidence for shared underlying biology between nutrient intake and substance use in humans continues to accumulate. Despite recent advancements, clinical risk prediction models for complex genetic diseases do not leverage any of this shared biology.
Here we develop a reward system aggregate phenotype score from quantitative self-reported "palatable" nutrient intake and substance use (i.e. fat, sugar, alcohol, nicotine, and caffeine) in a subsample of the UK Biobank cohort. We then conduct a genome-wide association study (GWAS) to identify reward system-associated variants in participants of white British ancestry. Using results from this GWAS, we next construct a reward system PRS to predict individuals with obesity or substance addiction. Relevant self-reported environmental risk factors that could be reasonably ascertained during a pediatric clinic visit (i.e. residing in a resource-poor area and history of child trauma) are also included in the prediction model. To evaluate any improvement in the predictive ability of our novel reward system PRS, we compare the performance to both an obesity-only PRS and a substance addiction-only PRS using the same training, validation, and testing sets. Findings from this work will define opportunities for preventative and therapeutic precision medicine in obesity and substance-related addictive disorders.

Statement of Significance: This research is significant, as it demonstrates effective use of genome-wide genotyping for the prediction and potential stratified prevention of complex genetic diseases, specifically obesity, substance use disorders, and their related sequelae (e.g. some cancers, ischemic heart disease and stroke). It leverages cutting-edge biological knowledge implied by recent animal, clinical imaging and lesion studies on reward system function, and directly applies it to a large human population using the latest computational methods in polygenic risk prediction. Curation of an innovative combination of genetic, environmental, dietary and mental health data allows for maximal predictive ability without sacrificing public health, clinical, or personal utility. Furthermore, by investigating a common genetic signal for a reward system deficit in humans irrespective of chemical substance, this work will yield further insight into the biological mechanisms of addiction through the report of specific genome-wide significant variants.

Keywords: Polygenic risk prediction, obesity, substance use disorders

E-mail of First Presenter: stevenkr@ohsu.edu

Multi-region expression profiling of archived breast ductal carcinoma in situ
Authors: Adam Officer, Joseph Steward, Daniela Nachmanson, Alexander Borowsky, Olivier Harismendy

Abstract: Each year in the United States approximately 10000 women are diagnosed with breast ductal carcinoma in situ (DCIS) lesions during screening mammograms. Currently there are no reliable markers to predict risk of progression, likely resulting in overtreatment. The small size and systematic archival (chemical fixation) have rendered the exploration of molecular markers extremely challenging. In particular, the role of the microenvironment, or of the cellular heterogeneity has been overlooked. To address these questions, a large atlas of expression profiles of DCIS lesions with comprehensive clinical and patient follow-up information is needed. We demonstrate the feasibility of the study on archived specimens (3-10 years old) from 21 DCIS patients with accompanying pathology annotation including nuclear grade, DCIS subtype, necrosis and inflammation state. Using laser capture microdissection (median region size - 1.6 mm^2) and RNA-seq library preparation (SMART-3SEQ) specifically developed for small, damaged tissues, we profiled 73 samples, corresponding to 36 regions (29 DCIS, 3 non-DCIS, 1 normal duct, 3 stroma). In each region a median of 7,122 unique transcripts were measured with 15,538 transcripts reliably measured (total read count > 10) across all samples. Correlation between regions within the same patient was significantly higher than regions between different patients (median correlation -0.04 vs 0.21, $p < 0.001$, Mann-Whitney U) suggesting patient-based factors affecting breast tissue expression levels. Furthermore, gene expression analysis revealed pathways significantly associated with nuclear grade and processes differentiating two histological subtypes; solid and cribriform. These results show the reliability of gene expression profiling from small areas of DCIS FFPE samples and how it can be used to characterize DCIS molecular heterogeneity, associating it with histological features to better evaluate factors associated with DCIS progression.

Statement of Significance: There are no effective models for when, or if, DCIS will progress to invasive cancer strongly suggesting overtreatment of this disease. Previous studies have used histological observations such as nuclear grade to predict progression rates but these models are imprecise. Transcriptomic analysis of breast cancer has discovered clinically relevant subtypes that are used in treatment decisions and that have provided insightful into the underlying biology. Analysis of DCIS using RNA-seq techniques is difficult based on the small size of these lesions and the poor quality of archived samples.

Keywords: Ductal carcinoma in situ, transcriptomic, laser capture microdissection, pathology
Email of First Presenter: aofficer@eng.ucsd.edu

Genome-Wide Detection of Epistasis in Antibiotic Resistant *M. Tuberculosis*
Authors: Anna G Green, Roger Vargas, Luca Freschi, Maha R Farhat, Harvard Medical School

Abstract: *Mycobacterium tuberculosis* is a globally prevalent bacterial pathogen with increasing resistance to antibiotics. While prior work has identified individual point mutations leading to antibiotic resistance, the evolution of stable, high-level antibiotic resistance is frequently a multi-step process. A complete understanding of the evolutionary routes to antibiotic resistance would allow us to better understand the biology of this

bacterial pathogen and forecast new resistance evolution. To date, there have been no systematic studies of epistatic mutations – mutations whose effects depend on their genetic context – in *M. tuberculosis* because the bacterium does not recombine, so variants are rarely observed in multiple genetic backgrounds, meaning that their individual effects cannot be disentangled from their combined effects.
Here, we develop a new phylogeny-based method to determine which mutations are more likely to occur following other mutations in the genetic background. We applied our method to a dataset of over 10,000 *M. tuberculosis* genomes spanning the four major global lineages, and found the top pairs of mutations to be within and between genes involved in antibiotic resistance and host adaptation. We detect multiple non-independent mutations in the same codon of *katG*, which confers resistance to isoniazid, that appear to protect against reversion to the antibiotic-susceptible phenotype, as well as an apparent multi-step evolutionary process toward high level isoniazid resistance. In addition to finding variants that both compensate for and amplify preceding resistance-conferring mutations, we also find numerous examples of non-independence between the evolution of resistance to different antibiotics, reflecting consistent selection pressures imposed by the standard multi-antibiotic treatment regimens. We expect that this methodology will illuminate the evolutionary routes to antibiotic resistance in *M. tuberculosis*, and possibly generalize to other non-recombining pathogens and asexual evolutionary processes such as in cancer.

Statement of Significance: Antibiotic-resistant bacteria are a global problem of increasing public health concern. Methods to slow the evolution and spread of antibiotic resistance have the potential to save many lives. Because antibiotics often target essential processes in the bacterial cell, mutations that confer antibiotic resistance may also impair the bacteria's core cellular function. It is speculated that the evolution of antibiotic resistance is accompanied by so-called compensatory mutations to make up for these deleterious effects. But, compensatory mutations have traditionally been difficult to detect. Here, we develop a novel method to detect compensatory mutations in the genome of Mycobacterium tuberculosis, the causative agent of tuberculosis, which has traditionally been difficult to study due to its low mutation rate and lack of recombination. We hope that by discovering compensatory mutations after the evolution of antibiotic resistance, we will better understand the biology of M. tuberculosis and inform possible new treatments

Keywords: Bioinformatics/computational biology, genomics, microbiology, infectious diseases
E-mail of First Presenter: anna_green@hms.harvard.edu

Poster Session 1

Machine Learning in Clinical Environments

Ethnicity-Associated Microbiome and Metabolome Differences Persist During a Short, Plant-Based Diet

Authors: Robert H. Markowitz, Junhui Li, Andrew W. Brooks, Jane F. Ferguson, Timothy Olszewski, Heidi J. Silver, Seth R. Bordenstein, Vanderbilt University

Abstract: Self-reported ethnicity is associated with health disparities, including incidence of obesity and diabetes. In previous work, we determined that taxon differences in the fecal microbiome were

associated with ethnicity based on a population level analysis of amplicon sequence data from the Human Microbiome Project and the American Gut Project. In order to test if this ethnicity signal in the microbiome is influenced by diet, we conducted a small human trial to characterize the oral and fecal microbiomes and urine and blood metabolomes of healthy African American and Caucasian women during a short, plant-based, dietary intervention.
Data analysis of the fecal and oral metagenomes identified significant ethnicity-associated taxon differences, which persisted during and after the transition from a habitual diet to a controlled 4-day plant-based diet. The Shannon index of oral microbiomes, but not fecal microbiomes, was also significantly different between ethnicities. However, no significant differences were observed in taxon richness regardless of oral unique richness, fecal unique richness, or co-existed richness. Analysis of Carbohydrate-Active enzymes (CAZymes) found oral, but not fecal, differences between ethnicities persisted throughout the diet. Additionally, ethnic differences of Antibiotic Resistance Genes were observed in both fecal and oral metagenomes.
On average, the dietary intervention significantly changed 10-12% of ~800 plasma and urine metabolites measured, with increases in sulfurous plant metabolites and decreases in chocolate- and coffee-associated metabolites, as expected. However, broadly these changes were not significantly different between ethnicities. Three metabolites, Epsilon-(gamma-glutamyl)lysine; N(6),N(6)-Dimethyl-L-lysine; Tryptophan Betaine, were identified in plasma which significantly varied between ethnicities following the diet. This study confirms that oral and fecal microbiomes are, in part, structured by factors captured by self-reported ethnicity. Furthermore, these results indicate that short-term dietary contributions to metabolomic profiles neither reflect nor drive ethnicity-associated structuring at a metagenomic level.

Statement of Significance: Ethnicity is a defining factor contributing to health disparity incidence in the United States. In previous work, we determined for the first time that differences in the gut microbiome were associated with ethnicity based on a population level analysis of data from the Human Microbiome Project and The American Gut Project. These differences were subtle but consistently replicated between four self- declared ethnicities in the US. Additionally, 11 of the 12 microbial taxa that recurrently varied in abundance between ethnicities across the two datasets were previously reported to be associated with human genetic variation, supporting a role for genetics in shaping the microbiome. In order to test if this ethnicity signal in the microbiome is influenced by diet, we conducted a small human trial to characterize the oral and gut microbiomes and urine and blood metabolomes of African Americans and Caucasians during a short, plant-based, dietary intervention.

Keywords: Metagenomics, Diet
Email for First Presenter: robert.h.markowitz@vanderbilt.edu

Clostridioides difficile **Infection among privately insured patients in the United States**

Authors: Jessica El Halabi, Nathan Palmer, Kathe Fox, Maha R. Farhat, Harvard University

Abstract: The epidemiology of *Clostridioides difficile* infection in the United States continues to evolve with incidence and severity rising in the past decade. National surveys of *C. difficile* infection are needed but challenged by the lack of strict surveillance data. In a novel approach, we repurposed private insurance claims data between the years 2007-2016 to conduct a retrospective observational cohort study to estimate disease burden and assess for risk factors of *C. difficile* infection and reinfection,

transmission dynamics between individuals of the same household, and outcomes after treatment including fecal microbiota transplantation.
For outpatients, we utilized International Classification Disease (ICD) 9/10 and the filling of anti-C. difficile treatment prescriptions to identify *C. Difficile* cases. For inpatients that lacked prescription data, we built an elastic net model which utilized patient characteristics like age and gender and ICD 9/10 codes to predict patients that were likely treated for a true *C. Difficile* infection. Our analysis will identify possible risk factors that increase the risk of developing *C. Difficile* both in patients in hospital and community settings. We also aim to better understand consequences of treatment and who is more likely to fail a treatment.

Statement of Significance: It is important to understand the risk factors of C. Difficile to help guide prevention strategies and aid in efforts of disease reduction.

Keywords: Clostridioides difficile, Infectious Diseases, Claims Data
E-mail of First Presenter: Jessica_Elhalabi@hms.harvard.edu

Exploring How Temporal Representation Affects Deep Longitudinal Models

Authors: Matthew C Lenert, Jeffrey Blume, Thomas A Lasko, Michael E Matheny, Asli O Weitkamp, Colin G Walsh, Vanderbilt University

Abstract: Many predictive models deployed in healthcare settings abstract away or ignore the temporal nature of disease and the healthcare process. Those that do model time face a myriad of development decisions. One such decision is the representation of time. There are many choices in temporal representation. Some examples include: use of a stochastic process to fit statistical patterns in data, abstracting data into ordered sequences, and dividing data into discrete time-windows.
Researchers found that deep sequence models best utilize temporal relationships compared to other deep neural architectures. Some researchers have found performance advantages to regularly spaced inputs, but the best representation of time is an open question. Two of the most prevalent sequence architectures are the Long Short Term Memory (LSTM) structure and a feed forward model with sequenced multi-head attention mechanisms.
We plan to evaluate how different temporal representations affect prediction performance of the LSTM and Attention architectures. We plan to generate data using a longitudinal mixed-effects statistical model; producing different data sets with varying temporal parameters such as feature collinearity, feature auto-collinearity, sampling scheme, model stationarity, as well as differing relationships between the features and the outcome. We will also vary how time itself is represented as a feature. We then plan to develop and evaluate a software library to help practitioners put these recommendations into practice. We plan to evaluate this suite and its recommendations using well benchmarked learning problems (in-hospital mortality and discharge prediction) from Beth Israel Deaconess' MIMIC III dataset.

Statement of Significance: We hope to contribute an evidence-based guideline for machine learning engineers and data scientists working with longitudinal data. By measuring different temporal properties within their data, we hope these machine learning practitioners will be able to make data-driven decisions when choosing temporal representation beyond traditional trial-and-error approaches.

Furthermore by learning more how these models work with numerical longitudinal data, we hope to inform deep learning architecture to address current weaknesses.

Keywords: Longitudinal Modeling, Time Series, Deep Learning, Temporal Representation
Email of First Presenter: matthew.c.lenert@vanderbilt.edu

Fully Automatic Detection of REM Sleep Without Atonia

Authors: Daniel Yaeger, Phillip Wallis, Soethiha Soe, Alexander Kain, Xubo Song, Miranda Lim, Oregon Health & Science University

Abstract: Typically, skeletal muscles are relaxed during rapid eye movement (REM) sleep, but patients afflicted with REM sleep behavior disorder (RBD) display abnormal muscle activation called REM sleep without atonia (RSWA). RBD is highly predictive of the development of Parkinson's disease and Lewy body Dementia, and early detection of RBD may allow for early intervention to delay or prevent the onset of degenerative neurological diseases. The current practice for diagnosis of RSWA is manual visual inspection of polysomnography signals by a trained clinician, which is time-consuming, expensive, and subjective. In our work we are developing a fully automated pipeline for RSWA detection, in which polysomnography signals are passed sequentially through three separate classifiers to 1) predict which segments of sleep correspond to REM; 2) detect REM segments that are free from apneic/hypopneic events; and 3) identify and classify two types of RSWA events. Each classifier is composed of two stages: a deep-learning based classification model and subsequent determination of the optimal state sequence in accordance with local observation and global transition probabilities. Apneas and hypopneas typically result in an arousal, and our pipeline is the first that is able to distinguish between true RSWA signals and spurious activity caused by apneic/hypopneic events. Rather than predicting whether a patient meets clinical criteria for RSWA using obtuse or hidden features, our pipeline directly identifies RSWA events, which allows clinicians to easily compare the output of the pipeline with their own analysis, and may facilitate clinical adoption of the pipeline.

Statement of Significance: REM sleep behavior disorder (RBD), which involves lack of muscle atonia during REM sleep (REM sleep without atonia, or RSWA) not only presents immediate risks to patient safety, but is also strongly predictive of subsequent development of neurodegenerative disease, particularly synucleinopathies. Accurate and timely diagnosis is important for treatment and interventions that may delay or prevent the onset of disease. Current methods for diagnosing RBD/RSWA are time-intensive, expensive, and subjective. We are developing a pipeline to aid in the diagnosis of RBD which would allow fully automated analysis of polysomnography signals to detect RSWA. Although other pipelines have been developed in the past, our pipeline identifies and classifies two types of RSWA events, allowing clinicians to easily compare their own analysis with pipeline output. Our pipeline is also the first that is capable of distinguishing between RSWA signals and arousals due to confounding apneic/hypopneic events.

Keywords: Machine Learning, Deep Learning, Sleep
E-mail of First Presenter: yaeger@ohsu.edu

Identifying Modifiable Predictors of Patient Outcomes after Intracerebral Hemorrhage with Machine Learning

Authors: Andrew N Hall, MS, [1] Bradley Weaver, BA, [2] Eric Liotta, MD, [2] Matthew B Maas, MD, MS [2] Roland Faigle, MD, PhD [4] Daniel K Mroczek, PhD [1, 3] Andrew M Naidech, MD, MSPH [2]

ABSTRACT: Background/Objective: Demonstrating a benefit of acute treatment to patients with intracerebral hemorrhage (ICH) requires identifying which patients have a potentially modifiable outcome, where treatment could favorably shift a patient's expected outcome. A decision rule for which patients have a modifiable outcome could improve the targeting of treatments. We sought to determine which patients with ICH have a modifiable outcome.
Methods: Patients with ICH were prospectively identified at two institutions. Data on hematoma volumes, medication histories, and other variables of interest were collected. ICH outcomes were evaluated using the modified Rankin Scale (mRS), assessed at 14-days and 3-months after ICH, with "good outcome" defined as 0-3 (independence or better) and "poor outcome" defined as 4-6 (dependence or worse). Supervised machine learning models identified the best predictors of good vs. poor outcomes at Institution 1. Models were validated using repeated five-fold cross-validation as well as testing on the entirely independent sample at Institution 2. Model fit was assessed with area under the ROC curve (AUC).
Results: Model performance at Institution 1 was strong for both 14-day (AUC of 0.79 for decision tree, 0.85 for random forest) and 3-month (AUC of 0.75 for decision tree, 0.82 for random forest) outcomes. Independent predictors of functional outcome selected by the algorithms as important included hematoma volume at hospital admission, hematoma expansion, intraventricular hemorrhage, overall ICH Score, and Glasgow Coma Scale. Hematoma expansion was the only potentially modifiable independent predictor of outcome and was compatible with "good" or "poor" outcome in a subset of patients with low hematoma volumes, good Glasgow Coma scale and premorbid modified Rankin Scale scores. Models trained on harmonized data also predicted patient outcomes well at Institution 2 using decision tree (AUC 0.69) and random forests (AUC 0.78).
Conclusions: Patient outcomes are predictable to a high level in patients with ICH and hematoma expansion is the sole modifiable predictor of these outcomes across two outcome types and modeling approaches. According to decision tree analyses predicting outcome at 3-months, patients with a high Glasgow Coma Scale score, less than 44.5 mL hematoma volume at admission, and relatively low premorbid modified Rankin Score in particular have a modifiable outcome and appear to be candidates for future interventions to improve outcomes after ICH.

Keywords: hematoma expansion; ICH; machine learning
Email of First Presenter: andrewhall@u.northwestern.edu

Improving Prediction of Survival for Extremely Premature Infants Born at 23 to 29 weeks Gestational Age in the Neonatal Intensive Care Unit

Authors: Angie Li, Sarah Mullin, Peter L Elkin, University at Buffalo

Abstract: After an extremely preterm infant is born, the infant is admitted to the neonatal intensive care unit (NICU) after initial resuscitation. However, if it becomes apparent that an infant is decompensating

or unlikely to ultimately survive and be neurologically intact, the family is counseled on withholding further intensive care. Prior to birth, providers often use the National Institute of Child Health and Human Development (NICHD) online risk calculator to predict survival. Because of advances in NICU care, the calculator may now be less accurate; a recent study found that actual survival of a cohort of extremely preterm infants was nearly twice as high as predicted (Andrews et al., 2016). After admission to the NICU, several scoring systems are available to predict mortality. However, these have poorer performance with extremely low birth weight infants, as evidenced by the SNAPPE-II survival model having a predictive performance of .78 ± 0.01 for infants <1500g (Richardson et al., 2001). Despite counseling aided by available risk calculators, providers and families still have challenges with shared decision-making regarding life- support because the individual postnatal course for each extremely preterm infant is highly variable. Therefore, improved models are necessary to better inform providers and assist families. In addition, more clarification regarding maternal and social factors which influence outcome could assist providers with counseling. Our objective is to further elucidate these factors and explore whether novel statistical and machine learning methods can produce models which more accurately predict survival during the initial NICU admission. Records of newborn infants admitted to the NICU in the MIMIC-III database who were born extremely preterm at 23 to 29 completed weeks of gestation were examined for the study. Models have been developed from a variety of information extracted from the records including demographics, vital signs, lab work, medications, procedures, and clinical text.

Statement of Significance:
With more information about the often-difficult hospital course that extremely preterm infants experience after admission to the intensive care unit, medical practitioners can be better prepared to counsel families on expected outcomes. Further study on whether the conclusions from this study improves clinician confidence in counseling patients who may be giving birth or have given birth to an extremely preterm infant can be performed.

Keywords: machine learning, clinical decision support, intensive care, maternal and neonatal care
E-mail of First Presenter: ali83@buffalo.edu

Nonparametric Deep Survival Analysis: Regularization & Missingness
Authors: Shreyas Bhave, Adler Perotte, Biomedical Informatics, Columbia University

Abstracts: Survival analysis methods have long been used to effectively model time-to-event data. In the healthcare setting, the Framingham risk score is a salient use case in which 10-year risk of cardiovascular disease is estimated using a Cox proportional hazards model, regressing over a narrow set of highly predictive clinical features. There are a number of drawbacks to these kind of baseline approaches used in the clinic including: (1) a limited set of features (2) assuming linear relationships (3) the proportional hazards assumption. We propose using the EHR which is a rich source of longitudinal information in conjunction with Deep Survival Analysis (DSA) with nonparametric survival distributions to improve accuracy.
In order to use a more expanded set of clinical features from the EHR for survival analysis, a number of challenges must be addressed: (1) there is a high degree of missingness in EHR data (2) there is no natural event to align all the data (3) many nonlinear relationships likely exist between clinical features.

DSA is an approach for addressing these issues by leveraging a deep conditional model of failure time. We propose a new way to regularize the model using a regularizer directly on log probability.
We ran experiments on simulated data and a social sciences dataset for testing different methods of regularization and exploring the effects of different types of missingness on model robustness. We observe that on simulated data the categorical nonparametric model is able to recover Gamma simulated data. In addition, on the social sciences dataset we outperform cox proportional hazards (0.75 vs 0.63 in-sample concordance). We show that the regularizer is effective in smoothing the distribution and controlling overfitting.
There is promise in using new methods of regularization, missingness modeling and calibration for more complicated survival distributions.

Statement of Significance: Survival analysis methods have been used to model time-to-event data in many clinical scenarios. In particular, the Framingham risk score models coronary heart disease over a 10 year horizon. They find that they are able to achieve an AUROC between 0.72 and 0.74 using the risk score. There is great opportunity to leverage the EHR as a rich longitudinal data resource to improve time-to-event prediction. Deep probabilistic approaches for modeling time-to-event data have the potential to outperform baseline models for prediction of coronary heart disease and other diseases. Understanding how well these models perform out-of-sample prediction in addition to how robust they are to various levels of missingness is understudied and remains very important in clinical scenarios.

Keywords: Survival Analysis, Deep Learning, Nonparametric
E-mail of First Presenter: sab2323@cumc.columbia.edu

Optimizing Machine Learning Models for Clinical Application
Authors: Collin Engstrom, David Page, Varun Sah, Brian Patterson, University of Wisconsin-Madison

Abstract: Recent years have seen an elevated interest in integrating machine learning techniques into patient care. In particular, risk stratification of patients for potential intervention(s) has benefitted from such automated techniques. At the same time, machine learning algorithms have not traditionally been optimized on metrics most relevant to healthcare providers. We show how optimizing on precision@k directly optimizes NNT, a measure often used by physicians in determining the efficacy of treatments and interventions.

Reference
1. Patterson BW, Engstrom CJ, Sah V, et al. Training and Interpreting Machine Learning Algorithms to Evaluate Fall Risk After Emergency Department Visits. Medical Care 2019;**57**(7):560-66 doi: 10.1097/mlr.0000000000001140.

Statement of Significance:
Traditional machine learning methods do not typically account for clinically relevant metrics by default. Such metrics may be useful in risk stratification tasks[1]. We endeavor to show how NNT (one relevant metric) can be included in the optimization of a common machine learning model.

Keywords: risk stratification, optimizing on precision@k, precision medicine, machine learning

E-mail of First Presenter: engstrom@cs.wisc.edu

Predict Late Patients in Pediatric Ophthalmology Clinic Using Machine Learning

Authors: Lin Wei-Chun, Chen Jimmy, Hribar R Michelle, Chiang F Michael, Oregon Health & Science University

Abstract: Purpose: As healthcare shifts towards value-based care, there has been an increased focus on providing efficient and cost-effective clinical services. An important barrier for clinic efficiency is a patient's late arrival. Predicting which patients will be late can allow clinic schedulers to adjust and optimize the schedule to minimize the disruption of patient lateness. However, effectively predicting late patients is a challenging task due to a variety of factors that are associated with late arrival for appointments. The purpose of this study was to develop machine learning models to predict late patients in pediatric ophthalmology clinics at OHSU.
Methods: The data of 6-years outpatient visits at OHSU were extracted. Timestamp and office visit data were used to calculate time-related variables. Patients who checked in more than 5 minutes after their scheduled appointment time were considered late. Models using random forest (RF), XGBoost, support vector machine (SVM), and logistic regression were developed to predict whether the patient would arrive late. We used 5- fold cross-validation to reduce over-fitting. Area under the curve-receiver operating characteristic (AUC-ROC) curve scores, F-measure, and sensitivity were used to evaluate the prediction models. Also, the importance of predictors in the XGBoost model were presented.
Results: The XGBoost model showed the best accuracy (AUC=0.664) and ability to distinguish late patients (Sensitivity = 0.618; F-measure = 0.516). The RF model had the second-best performance followed by the logistic regression and SVM. The top three important predictors identified in the XGBoost model were clinic hour, average time difference between patients check in and scheduled, and types of healthcare insurance.
Conclusions: Machine learning model with secondary use of EHR data can be used to predict late patients with reasonable success. More work is needed to refine the models to improve accuracy. Late arrival prediction has implications for improving clinical scheduling efficiency.

Statement of Significance: This study advances the idea that secondary use of electronic health records data with machine learning method is a feasible way to predict whether the patients will be late arrival. Additionally, it highlights the factors which may be more associated with late patients.

Keywords: Clinical workflow, machine learning, secondary use EHR data
E-mail of First Author: linw@ohsu.edu

Spatio-Temporal Analysis of the Effects of Air Pollution on COPD

Authors: Janette Vazquez, Samir Abdelrahman, Cheryl Pirozzi, Ramkiran Gouripeddi, and Julio C. Facelli, The University of Utah

Abstract: Substantial evidence exists showing a link between exposure to air pollutants and an increase in the risk and severity of pulmonary diseases, including chronic obstructive pulmonary disease (COPD). However, large scale studies were previously not feasible due to lack of effective sensors, computing

resources, data, and analytical methods to effectively analyze exposure health effects at high spatial-temporal resolution. With the increase in computing power and availability of new algorithms for the analyses of large datasets, recent studies have been able to highlight the importance of temporal patterns of exposure in determining health outcomes. A review of lung cancer epidemiology at high spatial-temporal resolution was recently published and listed the analyses methods used in the studies (spatial analysis using various regression modeling techniques and a time series analysis using a classifier and pattern mining techniques [1]); however, none of the papers included combined spatial and temporal analysis. In this study, we demonstrate the use of novel machine learning methods as well as the availability of more computing power to understand the effects of exposure to PM 2.5 on the acute exacerbations of COPD. Due to the broad range of temporal and spatial resolution of our data sources, we plan on exploring various novel temporal advanced machine learning techniques (e.g. temporal clustering, deep learning, etc.) to analyze our dataset and understand the underlying patterns and trends between COPD exacerbation and exposure to PM 2.5. Data will be aggregated using the geographic location from mortality registries from the National Center for Health Statistics (NCHS), smoking regional incidence, weather data from the National Weather Service, and air pollution data from the Environmental Protection Agency and State Environmental Networks. The preliminary focus of this study will be Salt Lake County in Utah.
The analysis of Spatio-temporal data using air pollution data and the NCHS mortality registries will allow us to analyze risk prediction when a vulnerable population is exposed to air pollution. This will be the first study, to the author's knowledge, to combine spatial and temporal data of air pollution with COPD data for the analysis of risk prediction and with the aim of identifying underlying patterns that may exist using novel temporal machine learning methods. While previous studies have used either spatial or temporal data, very few, if any, have actually used a combination of spatial and temporal data for their analysis. A lot of these studies, as well, have used simple regression techniques in their analysis of the datasets. We plan to use more robust machine learning methods appropriate for this type of analysis.

Keywords: COPD, Air Quality, Machine Learning, Deep Learning, Spatiotemporal Analysis
E-mail of First Presenter: janette.vazquez@utah.edu

Using High-Dimensional Pharmacogenomics Data to Predict Effective Antidepressant Treatment in Major Depressive Disorder Patients

Authors: Lauren M Rost, Douglas P Landsittel, Philip E Empey, University of Pittsburgh

Abstract: **Background**: Major depressive disorder (MDD) is the main cause of suicide, and the leading cause of disability from chronic illness in the world (Licinio, 2001; James et al., 2018). One third of individuals treated with antidepressants do not respond sufficiently (Bauer et al., 2013), and namely, patients' initial antidepressant treatment leads to complete remission in only 35-45% of depressed patients (Rush et al., 2006). Thus, there is an imminent need to better inform initial antidepressant treatment.
Objective: We leverage the power of high-dimensional pharmacogenomics data in a cohort of up to 150,000 patients to build a pipeline that 1) robustly supports classification for this heterogenous disease, for both diagnosis and symptom progression, and 2) predicts effective antidepressant treatment through machine learning algorithms.

Methods: The pipeline architecture involves natural language processing of clinical notes to better classify patients and assess symptom remission in MDD patients. With better confidence in diagnosis and symptom progression, we can then stratify patients with unsupervised learning to uncover significant covariates, like sex, race, and disease severity. We then build supervised learning models equipped with high-dimensional pharmacogenomic and electronic health record data to predict effective antidepressants for MDD patients.
Results: Of nearly 5,034 patients enrolled there are 1,097 patients (22%) with at least one ICD-10 code for MDD, 1,868 patients (37%) prescribed at least one antidepressant, and 981 (19%) patients prescribed at least one antidepressant and assigned an MDD ICD-10 code. These data inform algorithms within our pipeline to contribute to standardized data analysis and prediction for this cohort and other MDD patients as well.
Discussion: Our work bolsters the current research surrounding antidepressant prediction through more comprehensive high-dimensional pharmacogenomic and electronic health record data, a systematic approach to data analysis, a high-fidelity classifier for MDD using clinical notes, and greater diversity in our patient cohort.

Statement of Significance: Depression is the main cause of suicide, which is the eighth leading cause of death in the U.S. (Licinio and Wong, 2001; Murphy, 2001). Treatment of depression is generally associated with poor outcomes. Initial antidepressant treatment leads to complete remission in only 35-45% of patients (Rush et al., 2006). Further, in patients with unsuccessful initial outcomes, less than 50% will experience significant improvement from a change in, or the addition of another antidepressant drug (Rush et al., 2006). One approach to improving outcomes is to develop models for predicting which patients respond to specific treatments. Recent studies have used pharmacogenomics to develop predictive models. However, the lack of diversity in patient cohorts and small number of clinical variables included in the predictive models limits their utility and translational potential.

Keywords: pharmacogenomics, major depressive disorder (MDD), prediction, machine learning, depression
E-mail of First Presenter: laurenrost@pitt.edu

Poster Session 1 – Literature Mining, Computational Phenotyping and Exploratory Analyses

A Biomedical Use-Case Library and Voice Enabled Exploratory Search for Developing Intelligent User Interfaces
Authors: Michael Segundo Ortiz, Javed Mostafa MA, PhD, University of North Carolina at Chapel Hill

Abstract: Publication Access Through Tiered Interaction & Exploration (PATTIE) is a voice enabled, visual, and real-time dynamic cluster-based exploratory search approach for PubMed built on the Scatter/Gather information retrieval paradigm. In this presentation, I will discuss the server-side and client-side architecture, provide use-case examples from our biomedical use-case library, and introduce a novel concept in the field of search engines and information retrieval called executable search notebooks.

Statement of Significance: The intramural and extramural research community of NLM has a deep interest in accelerating biomedical discovery within this database and is the motivation for this work. Traditionally there are four themes in the evaluation methodology of information visualization systems – controlled experiments comparing design features within a system; usability of a system; controlled experiment comparing two or more systems; and case-studies. The major challenge and opportunity is the case-study as this method evaluates tools against users with real problems which can be difficult to properly devise. (i.e. use-cases). Although my research has not yet reached the evaluation of real uses (next phase), progress has been made on a model for the creation, organization, and hosting of crowd-sourced and curated use-cases that will serve the NLM extramural research community in their practice of conducting user studies for research and development of intelligent user interfaces such as PATTIE.

Keywords: Search engine, visualization, human-computer interaction, unsupervised machine learning, information retrieval
E-mail of First Presenter: msortiz@unc.edu

Geostatistical Visualization of Ecological Interactions in Tumors
Authors: Hunter B Boyce, Anup Sood, Fiona Ginty, Parag Mallick

Abstract: Tumor heterogeneity makes it difficult to properly assess whether a tumor will progress or remain indolent. Tumor evolution is significantly influenced by the poorly understood ecological processes between cancer cells and their environment; namely the spatial interactions between cell types. Applying ecological methodologies to the tumor microenvironment will help uncover spatially-influenced mechanisms that differentiate between tumors that progress and those that remain indolent. To help elucidate spatial relationships we 1) use agent-based modeling to model tumor growth and progression under four different ecological contexts; predation, commensalism, mutualism, and parasitism; 2) apply spatial statistics to quantify the spatial heterogeneity on a local and global scale; and 3) visualize the spatial patterns of the tumor microenvironment that lead to tumor progression. We show that the small differences in local spatial phenomena are reflected in the emergent global spatial behavior in each of our well-defined ecological models. To demonstrate utility in more complex systems, we validated these spatial patterns and show that they can separate patients in terms of overall survival in a cohort of 252 lung cancer FFPE samples in a multiplexed histopathology data set. The results of this research show that spatial statistics can differentiate between small changes in local spatial interactions in a well-defined tumor ecology context and can be directly applied in computer vision analysis or pathomic analysis of histopathology data to provide better prognostic power.

Statement of Significance: Recent reports have demonstrated that intratumoral heterogeneity can have a significant impact on tumor progression, invasion, and drug resistance. Additionally, differential spatial patterning of cells in the tumor microenvironment (TME) has been shown to classify good and poor survivors. We show that quantifying spatial patterns using geostatistical methods can discriminate between cell-cell interactions in different ecological niches, and can separate a lung cancer cohort in terms of overall survival. This work demonstrates the utility of geostatistical methods in an analysis of the tumor microenvironment.

Keywords: Geostatistics; Pathology; Cancer; Tumor Ecology; Agent-Based modeling, Visualization
E-mail of First Presenter: hboyce@stanford.edu

Identifying Candidate Antibiotic Pairs for Alternating Treatment Strategies
Authors: Andrew Beckley, Erik Wright, University of Pittsburgh

Abstract: **Background**: Antibiotics have enabled many advancements in modern medicine, including organ transplantation and chemotherapy. Widespread use of antibiotics has resulted in increasing antibiotic resistance rates. However, resistance development to some antibiotics may cause pathogens to become sensitive to other antibiotics in a phenomenon known as collateral sensitivity. Alternating between collaterally sensitive antibiotics has been proposed as a means to reverse resistance, yet clinical trials are inconclusive. Effective application of collateral sensitivity requires knowledge of clinically viable antibiotic pairs. In this study, we attempt to identify candidate collaterally sensitive antibiotic pairs using a novel and clinically relevant method.
Methods: Antimicrobial susceptibility test (AST) results represent a useful clinical data source for study because they depict resistance co-occurrence patterns for different species. We used 440,000 AST results from the University of Pittsburgh Medical Center to identify collaterally sensitive antibiotic pairs for six common bacterial pathogens (*E. coli, K. pneumoniae, P. aeruginosa, S. aureus, P. mirabilis*, *E. faecalis*). Our method relies on the observation that collateral sensitivity prevents pathogens from becoming resistant to two antibiotics at the same time (cross-resistance).
Results: Among the six species tested, cross resistance was much more frequent (55% of all antibiotic pairs) than independence between antibiotics (44%) or collateral sensitivity (1%). As expected, cross-resistance was most common between antibiotics sharing a similar mechanism of action. Although collateral sensitivity was rare at the species level, a subspecies analysis revealed 69 new collaterally sensitive antibiotic pairs.
Conclusion: Our results suggest that cross-resistance is more common than collateral sensitivity among clinically relevant resistance mutations, possibly explaining the mixed results of large-scale treatment programs based on alternating between antibiotics. Nevertheless, the identification of many potential antibiotic pairs at the subspecies-level is encouraging and indicates that detailed taxonomic information is required to capitalize on the potential of treatment strategies based on collateral sensitivity.

Statement of Significance: Antibiotic resistance poses a substantial burden on the modern healthcare system. To help mitigate the problem of antibiotic resistance, new treatment strategies have been proposed which attempt to control the evolutionary dynamics of resistance through a phenomenon called collateral sensitivity. Collateral sensitivity is the process by which a pathogen may become more sensitive to particular antibiotics after developing resistance to other antibiotics. Attempts of leveraging collateral sensitivity through alternating antibiotics in the clinic have yielded inconclusive results, which could be due to a lack of clinically focused studies for identifying candidate antibiotic pairs. Our study aims to bridge the gap between promising experimental in vitro studies and inconclusive clinical efficacy by searching for antibiotic pairs which may have clinical viability through the use of antimicrobial susceptibility test results.

Keywords: Antibiotic resistance, collateral sensitivity, cross-resistance, clinical microbiology

E-mail of First Presenter: amb462@pitt.edu

Incorporating Electronic Health Record and Genetic Data to Improve Coronary Artery Disease Risk Prediction

Authors: Harsh Patel, David Crosslin, University of Washington

Abstract: Coronary Artery Disease (CAD) is a global leading cause of death and disability worldwide. Killing over 365,000 individuals in the United States in 2017 alone, CAD is the most common form of heart disease and manifests with a myriad of associated phenotypes. Studies have ascertained a polygenic architecture and substantial heritability, leading to the development of polygenic risk scores (PRS) to categorize individuals who may be at high risk for developing CAD their lifetime. However, existing CAD PRS's have only been successful in distinguishing cases from controls in traditional white European populations and tend to not transfer well between study samples. Furthermore, apart from genetic risk, a patient's environmental and social features are not reflected in these risk scores, thus leaving out a large amount of untapped knowledge. We propose utilizing a more diverse population from the Electronic Medical Records and Genomics Network (eMERGE) and using both traditional statistical genetics and machine learning methodologies to create a more accurate and robust metric for risk prediction. A traditional PRS of CAD individuals in the eMERGE network (n = 78602, cases = 23299) yielded successful case-control separation with a p-value of <0.001. This was achieved using loose phenotyping definitions based on existing literature. We believe better risk prediction can be achieved by tightening the phenotyping schema and by incorporating electronic health record (EHR) data. As combining these features is non-trivial, we plan to utilize machine learning, specifically deep learning, and compare the outcomes with existing techniques.

Statement of Significance: Precision medicine aims to tailor interventions and treatment options to individual patients based on their genetic background and medical history. Much work has been done to incorporate these two components in creating clinically actionable risk prediction models. However, with complex phenotypes, it becomes non-trivial to simply combine features based on traditional statistical models. Furthermore, as researchers get access to larger, more diverse datasets, it will require more robust methods to analyze and make sense of this information. Machine learning, specifically deep learning, is a potential alternative. Combined with domain knowledge and a focus on interpretability, explainable AI may be the next step for precision medicine.

Keywords: Genomics, Machine Learning, Risk profiling
E-mail of First Presenter: hpatel96@uw.edu

Data-driven refinement of gene sets for analysis of disease gene expression

Authors: Alexander T. Wenzel, Pablo Tamayo, Jill P. Mesirov

Abstract: Recent efforts have sought to use data describing the functional state of cancer cells to identify relationships between genetic pathway behavior, disease, and potential therapies. Gene set enrichment analysis (GSEA) is a community standard algorithm for identifying altered processes and pathways, represented by gene sets, in transcriptomic data. These analyses rely on *coherent* gene sets consisting of genes whose coordinate up- or down-regulation is a sensitive and specific indicator of the activity of a specific pathway or process. Many collections of gene sets have been assembled from literature curation to serve as input to GSEA. However, such gene sets are potentially incomplete or contain incorrect genes due to human error. Furthermore, gene sets derived from empirical data from a single experiment are often limited to their specific experimental context. Here we propose to use a large compendia of expression data to refine existing gene sets via a matrix-decomposition-based approach. This method has the advantage of using large collections of expression data to gain a broader quantitative perspective on the relevant biology than was possible in previous gene set curation efforts. Using a mutual-information-based metric comparing enrichment scores with direct measurements of pathway activity, we show that this approach is capable of increasing this association by 20% or more in a broad collection of refined gene sets by removing incoherent genes. We further show that a network propagation-based approach can identify additional co-regulated genes that were not in the original gene set. Our results demonstrate the promise of this data-driven approach to generate a new collection of gene sets that may increase the sensitivity and specificity of gene set enrichment analysis.

Statement of significance: The integration of sequencing technologies in clinical care has largely focused on categorizing patients by genetic variants and prescribing appropriate courses of treatment. However, this approach is not always effective, especially when no significant variant indicating a course of treatment can be detected. Recent work has shown that the detection of aberrant pathway activity in gene expression data from patients can be used to match patients with treatments [**1**]. The analyses such as those described in [**1**] require a pathway annotation approach such as GSEA that relies on strong, coherent gene sets for accurate results. The work described above will be critical in generating the best gene sets representing the pathways often dysregulated in cancer and improve the effectiveness of gene-expression-based drug discovery and treatment selection.

Keywords: genomics; gene expression; precision medicine

Email of First Presenter: atwenzel@ucsd.edu

References

1. Hanaford, A.R., Archer, T.C., Price, A., Kahlert, U.D., Maciaczyk, J., Nikkhah, G., Kim, J.W., Ehrenberger, T., Clemons, P.A., Dancik, V., et al. (2016). DiSCoVERing Innovative Therapies for Rare Tumors: Combining Genetically Accurate Disease Models with In Silico Analysis to Identify Novel Therapeutic Targets. Clinical Cancer Research *22*, 3903–3914.

Leveraging the electronic health record to evaluate the effect of anti-hypertensive medications on mortality in COVID-19 patients

Authors: Zachary Strasser, Hossein Estiri, Shawn N. Murphy, Harvard University

Abstract: COVID-19 has led to unprecedented havoc on economies and health, and has been difficult to manage with existing treatments. The electronic health record data provides critical insight for studying the disease. Here we utilized the Mass General Brigham COVID datamart to evaluate the effect of common anti-hypertensives on mortality. The question is of particular clinical consequence because common anti-hypertensives increase the expression of ACE2 receptors in human cells, the same receptor that COVID-19 uses to gain entry into the cell. It has been hypothesized that patients might be more likely to suffer from a severe COVID-19 infection if they are taking certain anti-hypertensives. A study published in the New England Journal of Medicine initially showed no increase in mortality, but was then retracted when the database validity was brought into question. This study will leverage the Mass General Brigham COVID datamart database to address this critical question.

Statement of Significance: COVID-19 is a worldwide pandemic, yet fundamental questions for treatment are not well understood. The goal of this study is leverage a massive EHR system in order to answer fundamental key questions for guiding therapy.

Keywords: big data, COVID-19, ACE-I
Email of First Presenter: zachary_strasser@hms.harvard.edu

Unsupervised Literature Tagging of Computational Neuroscience Literature, Towards Question Answering

Authors: Evan Cudone, R Andrew Taylor, Robert A McDougal

Abstract: Curation and knowledge dissemination of the computational neuroscience field requires many unique considerations as it utilizes language, methods, and ideas from the likes of biology, chemistry, physics, mathematics, medicine, and computer science. In order to effectively facilitate these informatics tasks for the computational neuroscience community we must first develop a robust representation of its literature. Using unsupervised topic modeling approaches, a metadata tagging schema was developed for computational neuroscience literature from ModelDB (a repository of computational neuroscience models), and compared to that of the larger neuroscience community. This analysis shows key differences in the types of discoveries and knowledge addressed in neuroscience and its computational subdiscipline, and gives some insight into how an automated question answering system might differ between the two.

Statement of Significance: As with all multidisciplinary fields, computational neuroscience requires informatics to facilitate efficient and effective knowledge discovery for its diverse community. To this end, an automated question answering system must be tailored to the content and specifics of this discipline. By exploring computational neuroscience literature using unsupervised topic modeling, we can uncover latent structure in the discipline's semantic space and use that to better inform and design

such question answering systems. These informatics efforts promote dissemination and integration of the research for the larger scientific community.

Keywords: Topic modeling, natural language processing, curation, neuroscience
E-mail of First Presenter: evan.cudone@yale.edu

Poster Session 2

Poster Session 2 – Genomics and Bioinformatics

An Ensemble Approach to Study Tumor Evolution Using Multiple Samples

Authors: Carlos C Vera Recio, Guillermina Lozano, Wenyi Wang, University of Texas MD Anderson Cancer Center

Abstract: Researchers are shifting towards collecting whole-genome/exome DNA sequencing (WGS/WES) data from multiple regions or time-points from a given tumor to understand the tumor's clonal composition and its role on disease and patient outcome. This type of data has ambiguity in its sub-clonal architecture, making clustering of the somatic mutations a critical step, and discrepancies in the results of available clustering methods complicate downstream analysis. Although methods to generate consensus of several clustering algorithms were developed in the Pan-Cancer Analysis of Whole Genomes study, due to differences in mutations and copy number of distinct samples of the same tumor, these do not apply in multiple-sample collection approaches.

We introduce a novel method to ensemble clustering algorithms applicable to DNA sequencing data from multiple samples of the same tumor. In our method, clustering results from each sample are summarized in co-clustering matrices (CCMs); results from all samples from the same tumor are then used to impute values for mutations not observed at a given sample; sparse matrix decomposition is then applied to the imputed CCM to generate final clustering results. To choose the optimal number of clusters, we model the mutation frequencies across samples as a mixture of multivariate Gaussian distributions and use the Bayesian information criterion for decision making.

We will measure the performance of our novel method on simulated data. We will use the method to analyze several clinical datasets generated at MD Anderson Cancer Center, including A) ARTEMIS: A Randomized, TNBC Enrolling trial to confirm Molecular profiling Improves Survival, a clinical trial that has generated WES data of the tumor and matched normal sample before and after treatment for 360 TNBC patients and B) A cohort of 55 patients with prostate cancer followed by liquid biopsies of cell-free tumor DNA, each with samples collected through 3-7 time points.

Statement of Significance: To understand the genetic heterogeneity present in tumors, its impact on disease and patient outcome, researchers have shifted to generating DNA sequencing data from the same tumor over time, or from different sections of the tumor. These multiple-sample collection protocols create a critical need for methods that leverage the increased availability of DNA sequencing data for each tumor to accurately recover its clonal architecture, while addressing the new challenges created by the different mutation profiles between distinct samples for said tumor. Our research addresses this need by developing a novel consensus clonal reconstruction method, which will be made publicly available, that more accurately recovers the clonal composition of a tumor using DNA sequencing data from multiple-samples of that tumor. We will apply our method to two clinical cohorts of prostate and breast cancer, thereby providing new insights into the role of

tumor heterogeneity in these diseases.

Acknowledgement: This work is supported by the NLM Training Program in Biomedical Informatics and Data Science T15LM007093, Program Director Dr. Lydia Kavraki.

Keywords: Tumor Phylogenetics, Tumor Evolution, Somatic Mutations
E-mail of First Presenter: cvera@mdanderson.org

Assessing the Influence of Regulatory Landscape Complexity on Gene Expression

Authors: Mary Lauren Benton, Douglas M Ruderfer, John A Capra, Vanderbilt University

Abstract: Disrupted regulation of gene expression substantially contributes to the risk of many complex diseases. The vast majority of loci associated with disease in genome-wide association studies are in non-coding regions of the human genome, and many contribute to disease risk by altering gene regulatory elements such as enhancers. Although there are numerous examples of enhancers that are mutated in disease, predicting whether mutations in a given enhancer will influence phenotype is still a difficult task. Current strategies for interpreting enhancer variation consider enhancers in isolation, despite evidence from mammals and insects that redundancy in enhancer landscapes buffers the phenotypic effects of enhancer loss on the expression of important genes. This project seeks to leverage genome sequencing data with functional genomic and evolutionary characterization to study the influence of regulatory landscape complexity on gene expression. We hypothesized that the evolutionary history and complexity of the enhancer landscape of a gene influences its robustness to variation. We test this using liver enhancers identified across multiple mammalian species to quantify the relationship between evolutionary conservation of activity and landscape complexity. We found that gains of enhancer activity are more likely to occur in complex enhancer landscapes with higher gene density. This model suggests that the gain of new enhancers is more likely to occur in regions with a larger number of gene targets and a greater amount of existing regulatory activity.
This work further integrates information about the genomic composition of regulatory landscapes across diverse tissues and the 3-dimensional genome structure linked to gene expression to provide insight into the functional impacts of enhancer gain and loss. Further study of the effect of enhancer alteration within the broader regulatory landscape will facilitate better interpretation of non-coding variants and perturbations to the gene regulatory architecture.

Statement of Significance: Non-coding regulatory regions are crucial for the maintenance of proper transcriptional programs in the cell, and genetic variants that disrupt these regions contribute to the architecture of complex disease. Many genome-wide association studies and variant prioritization methods consider the potential effects of single nucleotide variants. However, genetic variants rarely occur in isolation and studies of single variants or regulatory elements can miss important joint effects. New evidence suggests that mammalian genomes maintain stable gene expression levels and robustness to genetic variation through complex regulatory architectures; however, we do not fully understand the contribution of complexity to this robustness. This study integrates a range of genomic information to advance our characterization of regulatory landscape complexity and quantify its functional effects. The resulting quantification will further improve our understanding of how enhancer function may be modulated by its larger genomic context.

Keywords: Genomics, Bioinformatics/Computational Biology, Gene Regulation
E-mail of First Presenter: marylauren.benton@vanderbilt.edu

Correlated mutation analysis related to HIV-1 drug resistance
Authors: Skyler T. Kramer, Kyle Hill, Kamal Singh, Dong Xu

Abstract: Globally, 37 million people are currently infected by the human immunodeficiency virus (HIV), and another 1.7 million people become infected each year. The current recommended first-line antiretroviral treatment consists of two reverse transcriptase (RT) inhibitors and one integrase (IN) strand transfer inhibitor (INSTI). IN is responsible for integrating viral DNA into the host-cell genome, and INSTIs bind the IN-DNA complex. Many polymorphisms in the IN gene cause INSTI resistance. Importantly, recent studies suggest that some off-target mutations in the 3'polypurine tract (3'PPT) can independently cause INSTI resistance. Generation of viral DNA requires concerted activities of RT, RNase H (RH), and IN activities, and our recent experimental results show that mutations in the 3'PPT cause differential RNase H cleavage. Additionally, mutations in one other RT domain – Connection domain (CN) – can also affect the pattern of 3'PPT cleavage. Thus, this study aims to identify correlated mutations between 3'PPT, IN, RH, and CN to develop a better understanding of alternative origins of INSTI resistance. Shannon entropy is used to identify highly polymorphic sites within residue-level multiple sequence alignments of IN, RH, and CN. A corrected mutual information (cMI) score is then used to identify correlated mutations between each set of polymorphic regions and the full 3'PPT sequence. The cMI score incorporates low-count correction, average product correction, and z-score normalization. The results of this analysis suggest that correlated mutations exist between 3'PPT-IN, 3'PPT-CN, CN-RH, and RH-IN. Additionally, these results are being used to direct a series of *in vitro* experiments aimed to assess the kinetic and mechanistic significance of these correlated mutations.

Statement of significance: Despite recent advances in HIV-1 drug discovery and the impact of current antiretrovirals, drug resistance remains a continuous threat to long-time therapy. Thus, it is essential to understand the mechanisms by which INSTI resistance may develop. Recent publications have observed INSTI resistance in patient samples with 3'PPT mutations in the absence of IN mutations, though the mechanism by which the 3'PPT contributed to this resistance is not understood. The present study provides an *in silico* analysis to identify correlated mutations in regions of HIV-1 related to INSTI binding and potential 3'PPT degradation.

Keywords: Correlated mutation analysis; drug resistance; HIV-1
Email of First Presenter: stk7c9@mail.missouri.edu

Inferring Signaling Pathways with Probabilistic Programming
Authors: David Merrell, Anthony Gitter, University of Wisconsin-Madison

Abstract: Cells regulate themselves via dizzyingly complex biochemical processes called signaling pathways. These are usually summarized pictorially as a directed graph where nodes represent proteins, and edges indicate their influence on each other. In order to understand diseases and therapies at the cellular level, it is crucial to have an accurate understanding of the signaling pathways at work. Because signaling pathways can be rewired by disease, the ability to infer signaling pathways from condition- or patient-specific data is highly valuable.

A variety of techniques have been proposed for inferring signaling pathways. We build on past works that formulate signaling pathway inference as a Dynamic Bayesian Network structure estimation problem on phosphoproteomic time course data. Our approach is Bayesian; we use Markov Chain Monte Carlo to estimate a posterior distribution over possible Dynamic Bayesian Network structures. Our contribution consists of a novel proposal distribution that generates samples more efficiently for larger problems. We also relax some of the modeling assumptions made in past works.
We implement our method in Julia, using the Gen probabilistic programming language. Probabilistic programming is a powerful methodology for building statistical models. The resulting code is modular, extensible, and legible. The Gen language, in particular, allows us to customize our inference procedure and ensure efficient sampling.
We evaluate our method on simulated data and the HPN-DREAM Breast Cancer Network Inference Challenge, comparing our performance against a variety of baseline methods.

Keywords: signaling pathways; Bayesian methods; networks
E-mail of First Presenter: dmerrell@cs.wisc.edu

Enhanced Biological Interpretability of Single-Nucleus RNA-Seq Via Computational Correction of Systematic Biases
Authors: John Chamberlin, Aaron Quinlan, Kyoung Jae Won, Younghee Lee, The University of Utah

Abstract: The application of standard genomic assays to single-cells, particularly RNA sequencing (scRNA-seq), is increasing rapidly. Single-nucleus RNA-seq (snRNA-seq) is an attractive extension of the method for translational research in particular because it can be applied to frozen specimens from human donors, which are typically recalcitrant to non-destructive dissociation. However, bioinformatics methods development specifically for snRNA-seq remains almost non-existent: currently, the only difference in standard analysis is that intron regions of genes are included in expression analysis for snRNA-seq, whereas in scRNA-seq, only the exon regions are used. Here, we perform a comprehensive evaluation of the impact of this decision on gene quantification, with respect to both gene detection and to annotation ambiguities such as intron-exon overlaps between distinct genes. We anticipate and demonstrate the presence of a systematic quantification bias towards long genes, which distorts biological interpretation in the form of cell-type assignment and detection of corresponding "marker genes"; we suggest a simple linear model-based normalization approach.
Next, we characterize the impact of overlapping pre-mRNA models. While gene-ambiguous read alignments are a known limitation of standard scRNA-seq quantification software in principle, a systematic evaluation has not yet been performed. Counterintuitively, we find that numerous genes decrease in apparent expression level when intron regions are included in the annotation. This occurs because the supplemented intron regions introduce read-to-feature ambiguity not present in the standard mRNA annotation.
We conclude that distinguishing between mRNA and pre-mRNA in snRNA-seq preprocessing is necessary for effective normalization and substantially improves robustness, consistency, and biological interpretability. While pre-mRNA contributes approximately half of the usable data in an snRNA-seq experiment, we suggest that the unscrupulous addition of an additional one gigabase of intron sequence to the transcriptome could benefit from a more discerning approach.
This work represents the first comprehensive assessment of bioinformatic pre-processing choices in single-nucleus RNA-seq, an emerging technology which will likely enjoy widespread use in translational research for the foreseeable future. Our approach forms a natural foundation for future efforts to optimize normalization approaches for snRNA-seq as well as the application of the technology to new biological questions.

Communication and correction of the described biases is additionally impactful because they remain as-yet unacknowledged in the field. As the role of single-cell genomics transitions from primarily "observational" (e.g. cataloguing cell types across tissues) to "experimental" (e.g. investigating disease-specific gene regulatory networks), unbiased discovery potential becomes paramount.

Keywords: translational bioinformatics; genomic medicine; single-cell sequencing
E-mail of First Presenter: john.chamberlin@utah.edu

Generalized Analysis of Environmental, Host, and Geographic Specificity in Microbiome Data

Authors: John L Darcy, Anthony S Amend, Catherine A. Lozupone

Abstract: Understanding the factors that influence microbes' environmental distributions is important for determining drivers of microbial community composition. Species distributions are governed by parameters that influence their dispersal or survival, which could include continuous environmental variables such as temperature or pH, or more complex distances between sets of variables or geographic distances, or phylogenetic variables such as host range for parasitic species. Specificity is a concept that is often thought about in the context of symbiotic or host parasitic interactions, but here we generalize "specificity" to describe the extent to which species occupy a narrower range of an environmental variable than expected by chance. We show that Rao's (1982, 2010) quadratic entropy is a convenient diversity metric that can be applied to calculate specificity of a focal species to many different environmental variables, including 1-dimensional variables (vectors), 2-dimensional variables (dissimilarity matrices), and phylogenies. With slight modification, we show that this metric is useful for comparing specificity among species in simulated microbiome datasets, even when those species are omnipresent. We also apply this analysis to three empirical microbiome datasets to demonstrate the applicability of our tool: fungi living within the leaves of native Hawaiian plants, bacteria living within Antarctic glacier ice, and bacteria from the human gut microbiome.

Keywords: microbiome, geographic and environmental factors
E-mail of First Presenter: john.darcy@cuanschutz.edu

Inference of Novel Disease-Gene Associations through Network Communities

Authors: Jennifer Asmussen[1], Amanda Koire[1], Panagiotis Katsonis[1], Devika Subramanian[2], Olivier Lichtarge[1], Baylor College of Medicine[1], Rice University[2]

Abstract: Heterogenous biological networks are a promising source of novel disease-gene, disease-drug, and drug-gene associations and the communities derived from these networks contain information that is biologically meaningful and clinically relevant. The work presented here uses communities generated from the STRING protein-protein interaction network to identify novel disease-gene associations in The Cancer Genome Atlas patient cohorts. The underlying hypotheses are, first, that disease genes form functionally relevant groups that can be captured through network-derived communities, and, second, that even when mutational frequencies are too rare over any single gene to link them to a disease, the aggregate accumulation of functionally significant mutations over all the genes in a community can be recognized. To test this, we use a computational method that identifies communities enriched in high-impact functional mutations in a patient cohort. The computational method uses Evolutionary Action to measure the functional impact of a coding

mutation and Kolmogorov-Smirnov (KS) statistics to determine if mutations bias a community toward high-impact functional perturbations relative to the expected background of all mutations in the patient cohort. To control for the oversized effect of known cancer genes, leave-one-out analysis disambiguates which genes drive the bias. The significance threshold is generated for each STRING community through random simulation. This approach extends a prior study of Reactome pathways, curated by experts, and finds STRING communities linked to cancer that were not previously identified in the Reactome analyses. These data show that network communities have the potential to identify novel disease-gene associations. In the future, this approach will be extended to other sets of network-derived communities and other complex disease patient cohorts.

Statement of Significance: It remains difficult to identify genes in which rare mutations contribute significantly to complex diseases. Because these are likely to act in coordination with other genes, we tested whether we could identify genes that were significantly mutated as a group even when each gene is not significantly altered on its own. We used a community-based approach to identify functionally related groups of genes in a protein-protein interaction network and found that some of these communities were significantly differentially mutated such that they revealed low-frequency single-gene mutation events that likely drive disease. This provides a more in-depth understanding of the molecular genetics of complex disease and adds critical new information towards precision medicine. Additionally, this type of approach may prove useful to drug repurposing efforts if performed with network-derived communities generated from drug-gene-disease heterogenous networks. In summary, this work highlights the novel and clinically meaningful information contained in network-derived communities.

Acknowledgement: This work is supported by the NLM Training Program in Biomedical Informatics and Data Science T15LM007093, Program Director Dr. Lydia Kavraki.

Keywords: Precision Medicine, Network Communities, Evolutionary-wide Association Studies
E-mail of First Presenter: asmussen@bcm.edu

Known Variant Aggregation to Aid Exome Variant Interpretation

Authors: David W Sant, Nicole Ruiz-Schultz, Jordan Little, Alexander R Henrie, Krystal Chung, Kim Hart, Andreas Rohrwasser, Karen Eilbeck, The University of Utah

Abstract: The Utah Department of Health (UDOH) Newborn Screening Program (NBS) has recently completed validation of a second-tier exome-based next generation sequencing (NGS) pipeline with *a priori* analytic restriction to candidate genes associated with newborn screening disorders. The aim of this screening methodology is to detect causative genetic variants in individuals with abnormal NBS results. Variant analysis results are reported to the Utah NBS genetic counselor who is tasked with determining the clinical impact of each variant. Gene/disease-specific variant databases, such as ClinVar, contain valuable information regarding the pathogenicity of a variant. However, some variants found by NGS are not present in ClinVar while other variants have multiple interpretations with conflicting pathogenicity information. Investigating variants across multiple databases can aid the genetic counselor in variant interpretation but requires additional effort and time that the counselor may not have. Furthermore, it is not presently known if the other databases contain variants that are not present in ClinVar.
A computational pipeline has been developed to aggregate variant information from multiple variant databases in addition to ClinVar. Due to the differing reference genome versions and normalization styles, the same

variant could be found with different annotations between databases. To allow for accurate comparison, variants from all databases are normalized to HGVS format on the GRCh37 reference genome using the biocommons computational package. Preliminary analysis using 55 genes revealed that 54 of the genes had both variants uniquely found in ClinVar and variants uniquely found in LOVD databases. Only 17% of the variants found were in both ClinVar and the other variant databases, and only 4 of the genes had greater than half of the variants present in both. These findings indicate that there is benefit in obtaining variant information from other variant databases in addition to ClinVar for variant interpretation.
The work presented here is important as it shows that aside from ClinVar there are other repositories of clinical variant information that may be of value to genetic counselors or clinicians that are performing variant interpretation. Additionally, this project has allowed for the creation of a computational pipeline that will gather clinical variant information from multiple databases and normalize the variant annotations to the same HGVS format on the same reference genome (GRCh37). This can be used by the Utah Department of Health to aid the genetic counselors to quickly find the available information without having to search the individual databases. This project is a part of a larger project with the goal of creating one database accessible through a web browser that will allow individuals to easily search information from multiple databases through a single search.

Keywords: Genomics, Sequencing, Bioinformatics/computational biology
E-mail of First Presenter: David.sant@utah.edu

MRP: Rare-Variant Analysis via Bayesian Model Comparison Prioritizes Strong Risk and Protective Effects Across Biomarkers and Diseases

Authors: Guhan Ram Venkataraman, Christopher DeBoever, Yosuke Tanigawa, Matthew Aguirre, Chris CA Spencer, Timothy Poterba, Carlos D Bustamante, Mark J Daly, Matti Pirinen, Manuel A Rivas

Abstract: Whole genome sequencing studies applied to large populations or biobanks with extensive phenotyping raise new analytical challenges. The need to consider many variants at a locus or group of genes simultaneously and the potential to study many correlated phenotypes with shared genetic architecture provide opportunities for discovery and inference that are not addressed by the traditional "one variant - one phenotype" association study. Here, we introduce a model comparison approach we call Multiple Rare-Variants and Phenotypes (MRP) for rare-variant association studies that considers correlation, scale, and location of genetic effects across a group of genetic variants, phenotypes, and studies. We use summary statistic data and apply univariate and multivariate gene-based meta-analysis models to identify those rare-variant associations that have protective or risk effects and can expedite drug discovery. We apply the method to UK Biobank phenotypes across array and exome genotype data for individuals in white British, non-British white, African, South Asian, and East Asian cohorts, and demonstrate that the model comparison approach can aggregate rare-variant association signals across variants, phenotypes, and studies for greater power. We are able to find both previously-documented and novel associations between genes and several biomarkers and anthropometric, sensory, and disease phenotypes. Overall, we show that the MRP model comparison approach is able to retain and improve upon useful features from widely-used meta-analysis approaches in order to prioritize actionable gene targets.

Statement of significance: In the scenario in which rare variant association studies lack power, we can rely on a method like SKAT or the burden test (methods that aggregate signal across gene blocks). MRP is a more generalized, customizable version of these methods and can take into account functional annotation as well. In

incorporating prior information into the Bayesian model, we are able to better prioritize gene targets for therapeutics and better understand disease pathogenesis.

Keywords: Rare variants, genome wide association studies, Bayesian model comparison, therapeutics
E-mail of First Presenter: guhan@stanford.edu

Sequencing Families Reveals Common Deletions that Contribute to Autism Risk

Authors: Kelley Paskov, Dennis Wall

Abstract: Autism is a neurodevelopmental disorder resulting in impaired social interaction and communication skills, now impacting 1 in every 59 kids. While autism is 90% genetic and has been the focus of several large-scale sequencing efforts, known genetic risk factors account for only 6% of cases. Deletions are known to play an important role in autism risk, with large de novo deletions in certain regions causing syndromic autism and smaller deletions showing modest autism associations. However, deletions remain largely understudied since they are difficult to detect from next-generation sequencing data. We show that by sequencing the parents and siblings of affected individuals, we can identify both inherited and de novo deletions with high fidelity using unusual SNV transmission patterns within each family. We validate our deletion detection method using array CGH data from the well-characterized genome NA12878. By applying our method to whole genome sequencing data from more than 1300 multiplex and simplex autism families, we identify a set of common deletions that are predictive of autism status. We then predict the autism status of children from a separate group of families, and are able to identify 20% of autistic children with perfect precision. Many of these deletions fall in regions previously associated with autism. Furthermore these deletions tend to be inherited in families with multiple autistic children and to be de novo in families with only one autistic child. This work suggests that small, common deletions contribute to autism risk and highlights the utility of sequencing the first degree relatives of affected individuals.

Statement of Significance: Understanding the impact of deletions on autism risk helps us to 1) understand why autism occurs 2) develop diagnostic tools for autism 3) identify autism subpopulations that may exhibit similar patterns of behavior and respond to similar behavioral interventions. This work shows that sequencing whole families allows us to more accurately detect deletions which in turn are predictive of autism status. Furthermore, our method for detecting deletions in whole genome sequencing data is not specific to autism, and could easily be applied to other disorders for which we have family data.

Keywords: Autism, Deletions, Copy Number Variants
E-mail of First Presenter: kpaskov@stanford.edu

Integrated Omics Modeling of Transcriptional Regulation in Medulloblastoma

Authors: Owen S. Chapman[1], Tobias Ehrenberger[2], Tenley C. Archer[3], Maxwell P. Gold[2], Filip Mundt[4], Miriam Adam[2], Clarence K. Mah[1], Karsten Krug[4], Sahaana Chandran[5], Jesse R. Dixon[5], Scot L. Pomeroy[3], Ernest Fraenkel[2], Jill P. Mesirov[1], Lukas Chavez[1].
[1]University of California San Diego, La Jolla, CA; [2]Massachusetts Institute of Technology, Cambridge, MA; [3]Boston Childrens Hospital, Boston, MA; [4]Eli and Edythe Broad Institute of MIT and Harvard, Cambridge, MA; [5]Salk Institute for Biological Studies, La Jolla, CA

Abstract: Medulloblastoma (MB) is one of the most common pediatric brain tumors. Current treatment options carry substantial risk for lifelong cognitive and neurological impairment. Consequently, there is a pressing need to elucidate molecular mechanisms with the ultimate goal of developing targeted tumor-specific therapies. In a recent study on proteomes and phospho-proteomes of primary MB tumors, we identified proteomic MB subtypes that were not apparent from transcriptional or methylation data. To identify gene regulatory circuitries that drive these MB subtypes, we have now mapped accessible chromatin and 3D chromosome conformation in 23 and 13 MB tumors from the same cohort, using ATAC-seq and Hi-C respectively. We identify subtype-specific accessible regulatory chromatin regions, including a subset of regions not identifiable by the enhancer mark H3K27ac. We associate regulatory regions to their target gene promoters by correlating transcription with ATAC-seq signal strength, and confirm these relationships using the newly generated Hi-C data. Using an unpublished adaptation of the GSEA algorithm, which we tailored for an enrichment analysis of non-coding regulatory elements instead of genes, we identify transcription factors and master regulators with binding sites specific to an MB subtype. By associating these master regulators with their transcriptional targets, we have derived models of regulatory circuitry in MB which implicate central mechanisms of MB pathogenesis and may inform future therapeutic targeting.

Statement of Significance: This study represents a unique cohort of primary MB tumors for which multiple layers of 'omics information – whole genome, transcriptome, proteome, accessible chromatin (ATAC-seq) and chromatin interactions (Hi-C) – are available for the same samples. We have leveraged these orthogonal data layers to map patterns of transcriptional regulation across our cohort, and to trace the likely downstream consequences of specific oncogenic events in individual patients. Specific innovations include:
ATAC-seq of 23 MB tumors to identify active regulatory chromatin at greater resolution than previously available; Vertical integration of genome, transcriptome, proteome, accessible chromatin and chromatin interaction omics data layers to systematically identify enhancer target genes in MB; A novel network-driven approach to modelling transcriptional regulation in MB, enabling detailed characterization of the downstream targets of MB transcriptional regulators, both known and putative.

Keywords: Cancer genomics, epigenetics, regulatory networks
Email of First Presenter: ochapman@ucsd.edu

Zero-Inflated Random Forests for Genetic Regulatory Network Estimation in Single Cell RNA-Seq Data

Authors: Daniel Conn, Kirby Johnson, Emery Bresnick, and Sündüz Keleş, University of Wisconsin-Madison

Abstract: We present a new random forests (RFs) algorithm for count valued outcomes, and we apply it to single-cell RNA-Sequencing (scRNA-seq) data in the context of genetic regulatory network (GRN) estimation. Specifically, we embed our algorithm into the single-cell GRN network method SCENIC to find direct targets for transcription factors. The algorithm accounts for zero-inflation via an iterative Monte-Carlo expectation-maximization (MC-EM) process. It is crucial that zero-inflation be incorporated into the algorithm because it is believed that the generative process for the excess zeroes is primarily driven by technical artifacts. As a result, we argue that our MC-EM RF algorithm targets biologically relevant parameters, in contrast to standard RFs (or other commonly used machine learning methods). We evaluate our method on 20 scRNA-seq data sets generated by the Tabula Muris Consortium.
Daniel Conn acknowledges support of the NLM training grant: 5T15LM007359.

Statement of Significance: We believe that this work is an important contribution to the field of scRNA-seq analysis. We also believe that our zero-inflated random forests algorithm has application outside of scRNA-seq analysis, as the technique is more broadly applicable to zero-inflated count data. Thus, this technique could have application outside of genomics in other areas of health science.

Keywords: Bioinformatics, Genomics, Single-Cell RNA-sequencing, machine learning, random forests
E-mail of First Presenter: dconn2@wisc.edu

Review of Herbal Natural Products for Pain Management in Biomedical Classification Systems

Authors: Termeh Feinberg, Cynthia Brandt, Department of Veterans Affairs, VA Connecticut Healthcare System, and Yale School of Medicine, Yale University

Abstract: Introduction: Herbal natural products (i.e. dietary supplements) contain many bioactive ingredients and are widely used for pain and related symptoms in the U.S. (e.g., turmeric, ginger, cannabidiol). Despite some clinical research, much remains unknown regarding the benefits of herbal natural product use, and risks associated with herb-drug interactions. Formalized approaches for capturing natural product use in healthcare and research settings are largely lacking. Ontologies are vocabularies of semantic terminology containing information regarding between-term and hierarchical relationships. However, commonly used ontologies in the U.S. healthcare system contain incomplete herbal natural product information, particularly in relation to their use for pain management. The aim of this review was to investigate existing efforts related to herbal natural products for pain management in biomedical classification systems.
Methods: We conducted a narrative review based on publications gathered from various Pubmed search combinations utilizing terms "botanical", "phyto*", "herb", "pain", "ontology", "classification", "lexicon", "term", and "annotation" with no restrictions (9/2019), in addition to Pubmed-suggested articles based on search history. We scanned the references of these publications and engaged in dialogue with colleagues at both local and national presentations (i.e. Yale University Center for Medical Informatics, 2019 Annual AMIA Symposium). Ontobee was also queried to identify relevant ontologies.
Results: We identified ontologies related to drug safety and one recently developed Chinese medicine for rheumatism ontology. In addition, existing reference terminologies (including two focused on creation of a structured vocabulary for indexing natural products) were identified, as was one data source for dietary supplement labels in the U.S. The lack of formalized ontologies for pain and related disorders was also highlighted.
Conclusions: Given the small handful of recent relevant projects from separate research groups, a formalized meeting of ontology researchers and content experts may be necessary to determine the feasibility of developing an herbal natural products for pain ontology. Additional considerations for ontology development include: 1. Re-use of existing ontologies, reference terminologies, and other data sources; 2. Existence of annotation properties required for class representation within the scope of MIREOT (Minimum Information to Reference an External Ontology Term); 3. Potential collaboration with OHDSI (Observational Health Data Sciences and Informatics collaborative, established with the goal of developing an open-source standardized knowledge base); 4. Future usability across international settings; 5. Potential role of Knowledge Engineering to address natural product and related herb-drug safety considerations; 6. Role of pain and pain management in hierarchical structure of potential future ontological development.

Keywords: Biomedical Classification, Pain Management, Natural Herbal Products, Ontologies

Email of First Presenter: termeh.feinberg@yale.edu

Poster Session 2 – Clinical Informatics

Computational Drug Repurposing Validation Strategies: A Review
Authors: Malvika Pillaia, Di Wua, PhD, University of North Carolina at Chapel Hill

Abstract: Drug repurposing (i.e., drug repositioning) is to predict whether the approved drugs can be used for diseases that are not labeled for these drugs. It's cost-effective and has less safety issues in comparison to the traditional process for drug development. Due to the serendipitous nature of past drug discoveries, there has been a push toward data-driven repurposed drug development for more consistent hypothesis generation, inspiring computational drug repurposing efforts. However, the predicted repurposed drugs need future validation (i.e., independent supporting evidence) to build more confidence to follow up. There is a lack of literature reviewing the landscape of validation strategies used by researchers. Therefore, the objective of this review was to examine how researchers provide validation for drug repurposing candidate predictions generated from computational methods. A comprehensive search was conducted on 09/12/2019 across three databases: PubMed, Web of Science, and ACM Digital Library for all relevant articles pertaining to computational methods for drug repurposing, resulting in 3086 query results. Following the methodology presented in the PRISMA Statement for systematic reviews, 416 independent studies using computational methods for prediction with validation were included in the review. Information extracted from the studies included number of citations, whether the paper was condition-specific, computational method used, and validation method used. A citation analysis was conducted for quality assessment. Studies using computational and non-computational validation approaches were described in this review. Due to diversity in how drug repurposing and validation are interpreted, the extent of validation provided varied across studies, regardless of prediction method used. As a summary, concerns about definitions of drug repurposing, strength of validation, and evaluation of validation sources need to be addressed for future computational drug repurposing studies.

Statement of Significance: Over the past few decades, there has been significant increase in spending for drug development, with fewer drugs approved than any other point historically. The term "Eroom's Law" (i.e. the inverse of Moore's Law) is used to describe the inverse correlation of increased monetary input into drug development and the number of drugs approved remaining flat or decreasing. Drug repurposing is currently explored as a strategy to complement this phenomenon. Many computational drug repurposing methods have been developed. To follow up, extra supporting evidence is needed to build confidence in drug repurposing candidates, preparing them for clinical testing.

Keywords: Computation, Drug repurposing, Validation, Review
E-mail of First Presenter: mpillai@live.unc.edu

Early Evaluation of Cancer Prevention and Survivorship Care TeleECHO (Extension for Community Health Care Outcomes) Program
Authors: Zheng Milgrom MD[1,2], Brian E. Dixon MPA, PhD [1,2,3,4], Eneida Mendonca MD, PhD[3,5,6].
NLM T15 Fellowship, Public and Population Health Informatics, Center for Biomedical Informatics, Regenstrief Institute, Indiana University

Department of Epidemiology, Indiana University Richard M. Fairbanks School of Public Health
Center for Biomedical Informatics, Regenstrief Institute, Indiana University
Center for Health Information and Communication, Department of Veterans Affairs, Health Services Research and Development Service, Roudebush VA Medical Center, Indianapolis, Indiana, United States.
Department of Pediatrics, Indiana University School of Medicine
Department of Biostatistics, Indiana University School of Medicine

Abstract: **Objective:** This study evaluates the Cancer Prevention and Survivorship Care TeleECHO (Extension for Community Health Care Outcomes) Program (Cancer ECHO) at Indiana University Fairbanks School of Public Health. Project ECHO is a videoconferencing model for delivering a combination of continuing tele-education and case-based learning to connect specialists (as 'Hub') with health professionals (as 'Spokes'). We aim to assess the effectiveness of Cancer ECHO improving providers' knowledge, self-efficacy, professional practice, and burnout; to evaluate whether its operational and curriculum design is appropriate for the needs of the practitioners; and to assess the professionals' motivations and barriers to participation.
Methods: We will examine using surveys and semi-structured interviews. Our study population includes all the people enrolled in the Cancer ECHO program, either as 'Hub' members or 'Spoke' members, as well as the health care professionals who heard about the Cancer ECHO program independent of their actual attendance or enrollment. We also the unenrolled guests that attended at least one teleECHO clinic session. Primary outcome measures are participants' change in knowledge, self-efficacy, professional practice, and burnout. Secondary outcome measures are participation engagement, satisfaction, motivation and barriers of attendance, and perspectives on the program's operational and curriculum designs. We will capture survey data from participants and transcribe interview recordings to analyze primary and secondary outcomes.
Results: This is an ongoing study, we are expecting to present the results at the conference. From September 16, 2019, to January 21, 2020, 9 Cancer ECHO sessions were held. 57 providers enrolled in the Cancer ECHO program as 'Spoke' members, and 29 people enrolled as 'Hub' members. 19 of the 57 'Spoke' registrants attended at least one session. The average participation among the 19 enrolled participants is 2.37 times.
Discussion: We will present the discussion on the conference after data collection.

Statement of Significance: Other evaluations have been conducted on the ECHO programs focusing on more specific topics of cancer prevention, such as tobacco cessation. (Cofta-Woerpel et al., 2018; Nethan et al., 2019) There is limited evidence supporting the effectiveness of using the ECHO model for general cancer prevention and survivorship care. To our knowledge, this Cancer ECHO is the first ECHO program that is targeting both general cancer prevention and survivorship care without limiting to any population; thus, this is the first study to evaluate such an ECHO program. Our study will inform the professional community of whether the ECHO model is an effective way to improve medical education, workforce capacity, clinical practice and reduce health disparities in general cancer prevention and survivorship care.

Keywords: ECHO, Cancer, Health Disparity, Tele-health, Tele-mentoring
E-mail of First Presenter: omilgrom@iu.edu

Bibliography

Cofta-Woerpel, L., Lam, C., Reitzel, L. R., Wilson, W., Karam-Hage, M., Beneventi, D., … Blalock, J. (2018). A tele-mentoring tobacco cessation case consultation and education model for healthcare providers in community mental health centers. *Cogent Medicine*, *5*(1). doi:10.1080/2331205X.2018.1430652

Nethan, S. T., Hariprasad, R., Babu, R., Kumar, V., Sharma, S., & Mehrotra, R. (2019). Project ECHO: a Potential Best-Practice Tool for Training Healthcare Providers in Oral Cancer Screening and Tobacco Cessation. *Journal of Cancer Education*. doi:10.1007/s13187-019-01549-8

EHR Utilization and Fragmentation: Sequence Analysis of Clinical Workflows

Authors: Amanda J. Moy, MPH,[1] Jessica M. Schwartz, RN, BSN,[2] Jonathan Elias, MD,[1,3] Kenrick D. Cato, RN, PhD,[2] Sarah Collins Rossetti, RN, PhD[1,2]
[1]Columbia University Department of Biomedical Informatics, New York, NY; [2]Columbia University School of Nursing, New York, NY; [3]NewYork Presbyterian/Columbia University Irving Medical Center, New York, NY

Abstracts: Limited measures of electronic health record (EHR) burden among clinicians exist. Considered the gold standard method in examining workflows,[1] prior time-motion studies (TMSs) have largely involved homogenous clinician populations, restricting the ability to comparatively evaluate burden across roles and settings.[2]
We identified interprofessional patterns in EHR use and workflow fragmentation using TMS data.
An analysis was conducted from a TMS of advance practice providers (APPs [residents, physicians assistants, nurse practitioners]) and registered nurses (RNs) from an acute care unit (ACU), intensive care unit (ICU), and emergency department (ED).[3] Workflow fragmentation (switches per minute), task-switch type, and task involvement in transitions were evaluated based on role and practice settings using descriptive and sequence analyses. We compared volume of task-switch types using the Pareto distribution to understand the proportion of tasks accounting for ~80.0% of switches.
Forty-three observations were conducted with 30 APPs and 13 RNs: ACU(n_{APP}=7; n_{RN}=8), ICU(n_{APP}=8; n_{RN}=5), ED(n_{APP}=15). Clinicians averaged 1.43±0.55 switches/minute, with ACU (1.46±0.70) and ED (1.44±0.58) APPs consistent with the mean, ICU APPs below (1.28±0.58), and ACU and ICU RNs exhibiting the highest (1.70±0.49) and lowest (1.13±0.26) mean switch rates, respectively. Twenty-six (26.5%) of the 98 task-switch types presented in the data accounted for 81.0% of all transitions. Of those, data viewing and data entry were involved in the most switches (42.6%); while frequent, median duration was low: data entry (*M*=35secs; $\bar{X}$=65.2secs±77.0), data viewing (*M*=10secs; $\bar{X}$=20.6secs±31.8). Task-switch-volume-to-type ratios were higher among APPs [ACU(80.0% to 21.5%), ED(80.9% to 22.1%), ICU(81.7% to 28.4%)] compared to RNs [ACU(80.9% to 37.8%), ICU(80.7% to 32.1%)].
We identified two EHR tasks (data entry and data viewing) with a median duration of 35 and 10 seconds, respectively, which explained a high proportion of workflow task-switches. Given the recognized extent of documentation burden, these data may shed light on targeted interventions for improving EHR usability.

Statement of Significance: Studying EHR workflow is important for understanding potential impacts on efficiency, documentation burden, cognitive overload, and safety.[4-6] The results presented in this study provide an aperture into the ample possibilities of using interprofessional TMS data to identify and quantify sources of clinician burden, as well as to conduct comparative analysis of workflow fragmentation, task sequence types, and task switches across roles and practice settings. Based on our analysis, clinicians experienced ~1.5 switches/min in their workflow. Data viewing and data entry were involved in 42.6% of task-switches, indicating documentation burden may play a critical role in workflow disruptions. As such, frequency of interruptions evaluated through task switches may serve as proxies for measuring documentation burden. Continued work will further investigate the nature of task-switch types, the fundamental sources of workflow fragmentation,

and the role of multitasking through TMS domains captured in parallel (i.e., communication and physical location).

Keywords: time and motion studies, workflow, electronic health records
E-mail of First Presenter: am3548@cumc.columbia.edu

References:
Lopetegui M, Yen PY, Lai A, et al. Time motion studies in healthcare: what are we talking about? J Biomed Inform. 2014 Jun;49:292-9. doi: 10.1016/j.jbi.2014.02.017. Epub 2014 Mar 7.
Holman GT, Beasley JW, Karsh BT, et al. The myth of standardized workflow in primary care. J Am Med Inform Assoc. 2016 Jan;23(1):29-37. doi: 10.1093/jamia/ocv107. Epub 2015 Sep 2.
Schwartz J, Elias J, Slater C, et al. An Interprofessional Approach to Workflow Evaluation Focused on the Electronic Health Record Using Time Motion Study Methods. AMIA Annu Symp Proc. 2019.
Coiera E. Technology, cognition and error. BMJ Quality & Safety. 2015;24:417-422.
Coiera E, Ash J, Berg M. The Unintended Consequences of Health Information Technology Revisited. Yearb Med Inform. 2016 Nov 10;(1):163-169.
Collins SA, Fred M, Wilcox L, et al. Workarounds used by nurses to overcome design constraints of electronic health records. NI 2012 (2012). 2012 Jun 23;2012:93.

New Clinical Decision Support System for Veterans

Authors: Melissa Resnick, Peter Elkin, University at Buffalo

Abstract: Background: A great deal of text in the Electronic Health Record (EHR) is unstructured.1 One solution to this problem is to use a standard terminology, such as SNOMED. The Department of Veterans Affairs has developed a new terminology called SOLOR, which is a combination of SNOMED CT, LOINC, and RxNorm. The various terms are used to tag the unstructured text: (a) SNOMED CT for diseases and findings and procedures; (b) LOINC for laboratory test results; and (c) RxNorm for medications. Once the text has been assigned terms from SOLOR, the HTP-NLP system developed at the University at Buffalo is used to identify named entities in free text and to create the post-coordinated compositional expressions to be used to drive clinical decision support. The coded clinical text is represented in Analysis Normal Form (ANF), which in turn, drives clinical decision support. The aim is to use this pipeline to provide decision support to Veterans. Methods: The ANF records were obtained which contain: (a) the text that was needed to drive one of many clinical decision support rules, (b) the assigned ANF with SOLOR codes; and (c) the results of the SOLOR and ANF derived from the text needed to drive knowledge artifacts for clinical decision support (HL7 KNARTS). The automated result from our pipeline was compared to the human generated SOLOR and ANF results to determine if all of the terms concepts and relationships were modeled correctly. Then, one of four accuracy measures was assigned: (a) missed, (b) partial map, (c) almost completely mapped, and (d) completely mapped.
Results: At the current time, both SOLOR and the HTP-NLP system have been built and are functioning.

Statement of significance: An ontology containing terms for diseases, diagnoses, laboratory procedures and findings, and medications, is important in NLP, as it provides a method for codifying unstructured text. Once the text is codified with terms from the ontology, the NLP system is used to identify named entities in free text and to create the post-coordinated compositional expressions to be used to drive clinical decision support. Using this pipeline, we will be able to provide clinical decision support to Veterans.

Keywords: Clinical Decision Support System; Ontology/SOLOR; Natural Language Processing
E-mail of First Presenter: mresnick@buffalo.edu

Syphilis Testing Adherence among Women with Livebirth Deliveries: 2014-2016

Authors: Opeyemi C Ojo, MPH[1], Janet NArno, MD[2,3], Guoyu Tao, PhD[4], Chirag G Patel, PhD[4], Andrea Broyles, MPH[5], Brian E Dixon, MPA, PhD[1,5]
[1]Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, IN
[2]Marion County Public Health Department, Indianapolis, IN
[3]Indiana University School of Medicine, Indianapolis, IN
[4]Centers for Disease Control and Prevention, Atlanta, GA
[5]Regenstrief Institute, Indianapolis, IN

Abstract: Background: Congenital syphilis cases are increasing in the US. Effective prevention requires routine serologic testing and treatment of infected pregnant women. The Centers for Disease Control and Prevention (CDC) recommends testing at the first prenatal visit and additional testing at 28 weeks gestation and delivery for women at increased risk. Prior studies on syphilis screening have mixed results. Administrative data from 2009-2010 suggest that women with Medicaid and commercial insurance have nearly universal screening rates, yet rates can vary by geography and population.
Methods: We conducted a retrospective, cross-sectional study of syphilis testing rates among pregnant women with a livebirth delivery using data from a statewide database of electronic health records from 2014-2016. We extracted all syphilis tests linked to birth certificate records provided by a local health department in a large metropolitan area. We excluded all livebirths where gestational age was missing or > 42 weeks. We classified women residing in high syphilis prevalent zip codes as high risk. In addition to descriptive statistics, we calculated syphilis testing rates on the cohort.
Results: Among 20,013 women with livebirths included in the study, syphilis testing in any trimester, including delivery, increased from 75.8% in 2014 to 86.9% in 2016. The number of maternal syphilis tests administered only at delivery decreased from 1,375 in 2014 to 330 in 2016. Additionally, high rates of syphilis screening were seen among women at increased risk of syphilis, testing increased from 84.6% to 94.6%.
Conclusion: Syphilis screening rates increased during the study period. The proportion of women who only received testing at delivery also decreased. Improvement is encouraging, yet approximately 1-in-10 women did not receive any screening during pregnancy. Despite consistent recommendations as well as legal mandates in 45 states, screening for syphilis in pregnancy continues to be suboptimal in certain populations.

Statement of Significance: The findings from this study depicts the gap in prenatal syphilis screening among pregnant women and the need to increase adherence.

Keywords: Congenital syphilis, prenatal screening, sexual transmitted diseases
E-mail of First Presenter: opolorun@iupui.edu

The Spectrum of Recognition and Engagement of Chronic Kidney Disease Care of a Single Healthcare System

Authors: YiFan Wu, Adam Wilcox, University of Washington

Abstract: Chronic kidney disease (CKD) has caused a tremendous burden in the USA and the prevalence of 30 million in 2016[1]. Various measures have implemented to intervene in the progression of the diseases in the population, however, the recognition of CKD in EHR and the engagement of care is rarely reported. We conducted a longitudinal study using the outpatient EHR data from 2015-2020 at the University of Washington health system(UW Medicine) to evaluate the spectrum of engagement of CKD care. We used the CKD National Kidney Foundation (NKF) guidelines to evaluate the population who should have been diagnosed as CKD Stage 3, 4 and 5 by using the abnormal lab values, and we also evaluated the cohort who have a CKD ICD code in their EHRs. Our results suggest that the CKD recognition rate is about 35% in the cohort.
Then we further narrowed down the population of patients who had at least one visit at the nephrology clinic at UW Medicine in the past five and the percentage of patients dropped down even further to ~12%. Once the patients had one visit at the nephrology clinic, the patients are more likely to maintain renal care.
Our analysis is a longitudinal examination of the spectrum of CKD care in a single large healthcare system. The study demonstrates the low engagement of CKD care in a large healthcare system and could be generalized to a larger scale. The previous study at Cleveland Clinic observed a similar pattern and reported an 11% CKD recognition in the EHR system[2]. The descriptive study would be meaningful for primary care physicians and for informaticians to understand the magnitude of challenges of poor engagement in renal care and take measures to bridge the gaps between recognition and care.

REFERENCE:

1. Saran R, Robinson B, Abbott KC, et al. US Renal Data System 2018 Annual Data Report: Epidemiology of Kidney Disease in the United States. Am J Kidney Dis. 2019 Mar;73(3S1): A7-A8. DOI: 10.1053/j.ajkd.2019.01.001. Epub 2019 Feb 21. PMID: 30798791; PMCID: PMC6620109.

2. Jolly SE, Navaneethan SD, Schold JD, Arrigain S, Sharp JW, Jain AK, et al. Chronic kidney disease in an electronic health record problem list: quality of care, ESRD, and mortality [Internet]. American journal of nephrology. U.S. National Library of Medicine; 2014 [cited 2020Feb22]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4056768/

Keywords: Electronic Health Record, Quality Control, Clinical Informatics
E-mail of First Presenter: yifwu@uw.edu

Ventilated Patients, Thromboprophylaxis and Major ICU Stay Related Outcomes: An Analysis of MIMIC-III Clinical Database

Authors: Mrigendra M. Bastola, MD, MS, Craig Locatis, PhD, Paul Fontelo, MD, MPH.

Abstract: Critical care patients are at higher risk for thromboembolic disorders. There are limited studies regarding the effect of Heparin, Warfarin, Enoxaparin on ventilated patients, who are likely to both benefit from prophylaxis and suffer from adverse effects of these blood thinners. This study assessed the MIMIC-III clinical database on 4192 ventilated patients. The authors systematically analyzed relevant data on the thromboprophylaxis agents and their effect on major treatment outcomes. The parameters studied were the length of ventilation, length of Intensive Care Unit (ICU) stays, ICU mortality, inpatient mortality, and occurrence of major thromboembolic events such as pulmonary embolism and deep venous thrombosis. 57% of the 4192

ventilated patients (median=59 years) received thromboprophylaxis, 157(3.7%) developed major thrombocytopenia requiring platelets transfusion, and 181 (4%) developed embolic disorders while on the blood thinners. Heparin, Warfarin and Enoxaparin had confounders adjusted Odds ratios of 0.45(p<0.01, 0.38-0.54), 0.28(p<0.01, 0.15-0.52) and 0.3(P<0.01, 0.16-0.58) for ICU mortality. Only Heparin (OR 1.52, 1.07-2.15, p<0,05) and Enoxaparin (OR2.05, 1.1-3.8, p<0.05) had clinical thrombocytopenia requiring platelets transfusion. None of studied agents had statistically significant effects on the reduction of pulmonary embolism and deep venous thrombosis, or improvement of Sequential Organ Failure Assessment (SOFA) scores at 4, 7 and 10 days of ICU admission, when compared with the control group.

Statement of Significance: The findings of this study support the use of thromboprophylaxis to improve ICU stay outcomes by reducing the duration of hospital and ICU stays, lowering ICU mortality, and reducing ventilator duration. The newer agents (Enoxaparin) might not be significantly better compared to other agents (Heparin, Warfarin). Although the benefits of the thromboprophylaxis agents have been established, their limited effects on reducing thromboembolic events, and effects on lowering platelets found in this study may aid practitioners in selecting and switching the blood thinners, if supported by further studies.

Keywords: Thromboprophylaxis, MIMIC-III, ICU mortality, In-patient mortality, Pulmonary embolism, Deep Venous Thrombosis.
E-mail of First Presenter: mrigendra.bastola@nih.gov

ABSTRACTS
Wednesday, June 24, 2020

Focus Session 4
Machine Deep Learning – Graphical Models

Mechanism Generalization Using Mechanism Centrality

Author: Harrison Pielke-Lombardo

Abstract: I present a method for inferring formerly represented pathway abstractions from sets of concrete pathways using an algorithm called Least Subsuming Subgraph (LeSS). A pathway abstraction (PA) is a schema for a set of pathways where constituent entities or activities are either left out or left unspecified. A PA may depict a submodule or subgraph in common among its subsumed concrete pathways (SCPs). Several PAs can be organized hierarchically, such as in the Gene Ontology Biological Process (GO-BP) ontology, but formal representations of these pathways depicting the structure and organization of entities and activities do not. PAs allow for modularization of known pathways, such as those in Reactome. Abstraction and modularization can assist in the transfer of knowledge from well understood pathways to new and less well understood pathways. Concrete pathway representations are often incomplete, and finding a good PA can facilitate hypothesis generation by making clear which entities or relationships need to be added, removed, or specified. In addition, pathway enrichment methods may not be able to map data to available target pathways because they either do

not take into account causal directionality or require precise mappings to specific entities or activities. PAs can capture the underlying mechanism and allow for abstract matching of constituents. This method is evaluated by inferring PAs from Reactome and reclassifying the SCPs. A good PA will have lost only so much information from its SCPs as to be able to subsume all of them. The success of optimizing the PA information loss will be reflected on the recall and precision of the classification. A good PA will achieve high recall by losing information, thus subsuming more pathways, while it will achieve high precision by retaining information, thus subsuming only those pathways which actually are SCPs.
This method is a novel knowledge transfer approach which may improve interpretability and our understanding of biological mechanisms. The ability to abstract and modularize pathways has many potential uses in improving our understanding of biological mechanisms. It can be used for hypothesis generation by suggesting constituents which need to be specified. PAs can be used as non-specific targets for knowledge based discovery such as pathway enrichment. This method can also be used as a way to compare pathways by making clear which components are shared between them.

Keywords: inference, knowledgebase, pathway enrichment
E-mail of First Presenter: harrison.pielke-lombardo@cuanschutz.edu

A Bayesian Approach to Local Causal Discovery in the Presence of Latent Confounding
Authors: Bryan Andrews and Gregory F Cooper, University of Pittsburgh

Abstract: We describe a new Bayesian approach for deriving posterior probabilities of causal relationships from observational data. Using the framework of ancestral graph Markov models, the method can account for patterns of conditional independence in the empirical distribution consistent with latent confounding. However, the number of possible causal models grows exponentially in the number of variables. Therefore, we perform a model search that only considers relationships local to a user-specified target variable. We derive the posterior probability of each of these local causal models and use them to compute the probability of each local variable's relationship (cause, effect, or confounded) with the target variable. We report the evaluation of this algorithm using synthetic data and real clinical data.

Statement of Significance: In this work, we perform local causal discovery with ancestral graph Markov models. Previously, methods have been developed that use conditional independence tests to derive constraints on what are admissible local causal models. In contrast, we use a Bayesian approach to derive the probability of such models. In particular, we compute the probability of causal relationships (cause, effect, or confounded) among a local set of variables. The main technical contribution of this work is the development of a factorization for ancestral graph Markov models which facilitates the use of a Bayesian score within local causal discovery.

Keywords: Local Causal Discovery, Latent Confounding, Bayesian Probability
E-mail of First Presenter: bja43@pitt.edu

Lopsided Bootstrap Loss Enables Segmentation Networks to Learn from Annotations with Multiple False Negatives
Authors: Darvin Yi 1, Endre Grovik 2, Michael Iv 3, Elizabeth Tong 3, Greg Zaharchuk 3, Daniel Rubin 1,3

Affiliations:
Dept. of Biomedical Data Science, Stanford University
Dept. of Medical Physics, Oslo University Hospital
Dept. of Radiology, Stanford University

Abstract: Deep learning has proven to be an essential tool for medical image analysis. However, the need for accurately labeled input data, often requiring time-and labor-intensive annotation by experts, is a major limitation to the use of deep learning. One solution to this challenge is to allow for use of coarse or noisy labels, which could permit more efficient and scalable labeling of images. In this work, we develop a lopsided loss function based on entropy regularization that assumes the existence of a nontrivial false negative rate in the target annotations. Starting with a carefully annotated brain metastasis lesion dataset, we simulate data with false negatives by (1) randomly censoring the annotated lesions and (2) systematically censoring the smallest lesions. The latter better models true physician error because smaller lesions are harder to notice than the larger ones. Even with a simulated false negative rate as high as 50%, applying our loss function to randomly censored data preserves maximum sensitivity at 97% of the baseline with uncensored training data, compared to just 10% for a standard loss function. For the size-based censorship, performance is restored from 17% with the current standard to 88% with our lopsided bootstrap loss. Our work will enable more efficient scaling of the image labeling process, in parallel with other approaches on creating more efficient user interfaces and tools for annotation.

Statement of Significance: Deep learning methods, especially dense prediction tasks like segmentation, have a heavy dependence on labeled data. For medical imaging tasks, these labels come at a heavy societal cost, often requiring hours of expert annotations for a single scan. Learning to use noisy label data, such as lesion labeling with heavy type II (false negative) errors, will enable machine learning teams to scale up data collection processes by lowering the quality burden given to the experts labeling the data. We hope that our work is a step forward in creating deep learning algorithms that are both more robust and more sustainable.

Keywords: Deep Learning, Segmentation, Noisy Labels
E-mail of First Presenter: darvinyi@stanford.edu

Decoding Cancer Evolution via Clinical Phylogenetics and Machine Learning

Authors: J Nick Fisk[1,2], Stephen Gaffney[3], Katerina Politi[4], Vincent Cannataro[5], Jeffrey Townsend[1,3,6]
[1] Interdepartmental Program in Computational Biology and Bioinformatics, Yale University
[2] Cancer Biology Training Program, Yale Cancer Center
[3] Department of Biostatistics, Yale University
[4] Department of Pathology, Yale Cancer Center
[5] Department of Biology, Emmanuel College
[6] Department of Ecology and Evolutionary Biology, Yale University

Abstract: Cancer progression is an evolutionary process influenced by restrictive microenviornmental factors and, thus, successful strategies to prevent or delay disease progression require understanding of the underlying molecular evolutionary mechanisms. While the evolutionary trajectories of neoplasms of the same type share many similarities in their etiology, patient life history and the clinical approaches utilized are shaped by unique and personal factors calling for precision medicine approaches in treatment. Due to the additional selective

pressure levied during patient therapy, molecular evolutionary tools of phylogenetic analysis and ancestral state reconstruction are well-suited to decoding cancer evolution. Enriching these phylogenetic analyses with patient clinical information and enhancing them with machine learning techniques enable nuanced probing of the evolution of cancer within individuals and the etiology of the disease as a whole, including the acquisition of resistance to targeted and general chemotherapy.
Here, we 1) infer ancestral states in tumor phylogenies including tumor stage to better identify milestones of disease progression in individual patients, 2) use machine learning to infer the stages and other tumor morphologies of unobservable ancestral states, 3) trace the evolution of mutational signatures across a tumor phylogeny with clinical information superimposed, enabling examination of the shifting contributors to cancer mutational load—including the mutagenic effects of therapies commonly utilized in the standard of care, and 4) quantify the selective advantage conferred by somatic variants in a response to treatment, revealing how cancer evades elimination and, ultimately, recurs. Taken together, we provide phylogenetically-informed descriptions of cancer progression using a novel approach which can be applied to better inform clinical decision-making and the identification of targets for drug development.

Statement of Significance: Cancer is on pace to supplant heart disease as the number one cause of death in the United States within the next 5-10 years, having killed over six-hundred thousand people the year in 2019 alone. Further, it remains the number two cause of death worldwide. While the myriad improvements in cancer therapy are notable, the war is not won and tumors persist, progress, and spread despite our best therapeutic efforts. The complexities of the cancer family of diseases and their molecular underpinnings call for more powerful and nuanced techniques derived from synthesis of methods from interdisciplinary fields. Understanding the enigmatic evolution of tumors is essential to subverting it. The methods and results found here have clear, actionable, and translational utility in clinical decision making in addition to enhancing our understanding of the foundational biology of cancer.

Keywords: Cancer, Phylogenetics, Machine Learning, Precision Medicine, Translational Research, Therapeutic Resistance, Etiology
Email of First Presenter: jeffrey.fisk@yale.edu

Identifying Novel, Precancer-associated Stem Cell Subtypes Through Regulatory Network Inference and Machine Learning

Authors: Bob Chen, Alan J Simmons, Austin Southard-Smith, Marisol Ramirez, Cody Heiser, Qi Liu, Ken Lau, Vanderbilt University

Abstract: Prevention and enhanced surveillance methods can be improved through the utilization of precancer characterizations and their predicted developmental profiles. Over the past decade, much has been learned about the cellular transitions from early colonic adenomas to carcinomas; however, the pathway remains underspecified. To refine these pathways, we utilize single-cell transcriptomics to characterize colorectal polyps and tumors biopsied from Vanderbilt University Medical Center patients. This project, encompassed by the Human Tumor Atlas Network, further investigates the genomic and histological properties of biopsied polyps and tumors. The high dimensionality of these transcriptomes, alone, necessitates the use of data-driven methodologies.
In this work, we outline novel, precancer-associated stem cell subtypes and their transcriptomic characterizations. These are identified through semi-supervised learning of gene regulatory network activity by

stochastic gradient boosting and protein-DNA interaction mining. The developmental trajectories described by these regulatory networks can then be transformed into biologically interpretable graphical models. Cellular potency, or a cell's potential to differentiate, is inferred through transcriptional diversity models. Within the studied precancer stem cell subtypes, we observe the enrichment of biological programs associated with chromatin remodeling, epithelial-to-mesenchymal inducing transcription factors, and cell-cell interactions canonically associated with progression to carcinomas.

Statement of Significance: By establishing colorectal pre-cancer characterizations at single-cell resolutions, we can augment the specificity of future translational approaches in the detection and treatment of neoplasia before malignant progression. In addition, these characterizations augment prospective hypothesis generation. Our framework enables the data-driven discovery and deep characterization of previously elusive precancer cell subtypes, of which may have been undetectable by candidate-based genomic or proteomic paradigms.

Keywords: Colorectal Cancer, Cancer Genomics, Precancer, Stem Cells, Machine Learning, Gene Regulatory Networks, Developmental Trajectories
E-mail of First Presenter: bob.chen@vanderbilt.edu

Predictive Modeling of Post-Operative Discharge to Skilled Nursing Facilities (SNF)
Authors: NE Lopez, Mike Hogarth, RA Gabriel, SL Ramamoorthy

Abstract: Background: Post-surgical discharge to skilled nursing facilities (SNF) are associated with poor prognosis and increased length of stay. Accurately predicting discharge to SNF in the pre-operative setting may allow for improved pre-operative preparation for patients and hospitals.
Objectives: To develop a predictive model to determine, pre-operatively, which patients will require discharge to SNF.
Design: This is a retrospective study using the American College of Surgeons National Surgical Quality Improvement Project (ACS-NSQIP) database. Patients undergoing elective, inpatient colorectal surgery in 2017 were included in the study. The primary outcome of interest was discharge to SNF. Using a training set, we performed a multivariable logistic regression to generate a model for discharge to SNF. We then used a validation set to assess the model. We used the area under the receiver operating curve and the Hosmer-Lemeshow test to evaluate discriminatory ability and goodness-of-fit.
Results: 5, 557 patients were included. The training set included 4,167 patients (75%) and the validation set included 1,390 (25%) patients. Demographics in the two groups were similar. Multiple logistic regression revealed the following factors had significant effects on post-surgical discharge to SNF: sex, transfer status, age, dyspnea, functional status, ascites, hypertension, dialysis, disseminated cancer, albumin, hematocrit, platelets, ASA classification, operative time, mechanical bowel prep, oral bowel prep and surgical approach. The area under the receiver operating characteristic curve of the model on the validation set was 0.86 (95% CI: 0.83-0.89) and demonstrated adequate goodness-of fit (Hosmer-Lemeshow test, p=0.5632).
Conclusion: Using pre-operative factors, we have developed a model to accurately predict post-operative discharge to SNF. Further studies will be aimed at validating this model on an external data set. If the model proves accurate, next steps will be to incorporate the model into electronic health records for use in pre-operative counseling.

Keywords: Predictive modeling, skilled nursing facility, colorectal surgery

Email of First Presenter: nelopez@ucsd.edu

Applying Open-Set Learning Methods to Cancer Treatment Predictions

AUTHORS: ALEXANDER CAO[†*], YUAN LUO[‡], AND DIEGO KLABJAN[†]

Department of Industrial Engineering and Management Sciences Northwestern University Evanston, IL Department of Preventive Medicine Northwestern University Chicago, IL

Abstract: Cancer treatment regimens often consist of a combination, or "cocktail", of drugs. The landscape of cancer drug cocktails evolves with discoveries of novel cocktails with improved treatment and lessening side effects. Predicting cancer treatments can, therefore, be natu- rally formulated in terms of an open-set learning problem. Our goal is to classify patients by cocktail based on features including their lab results and demographics. In addition, how- ever, sufficiently unique patients unlike those historically associated with "known" cocktails, or cocktails in the training set, should be classified as "unknown". This "unknown" class is an indication that different cocktails may be more suitable for those patients' treatments.

To this end, we study a novel extension of variational autoencoders utilizing a Gaussian mixture embedding in order to perform supervised clustering through deep generative mod- els. Reciprocally, we also investigate improvements to open-set recognition through several means. First, our model includes multiple subclusters within each class, which addresses patient heterogeneity within cocktails. And second, we introduce a new open-set recognition algorithm, incorporating both distance measures and orientation, leading to better accuracy and robustness with regards to increases in the number of "unknown" classes. We apply our methods to several standard academic datasets as well as electronic health records from Northwestern's hospitals.

Statement of Significance: Personalized medicine through quantitative, phenotypic profiling shows promise in medical care by guiding drug combination strategies. In cancer treatments, combination therapies are becoming the standard of care and many drug combination therapies have been approved or are under clinical trials.

Although some guidelines exist for certain cancer types, individual patients' responses to various drug combinations are still not well understood. For instance, which drug combina- tions would likely benefit a specific patient the most is still a critical, open question. With the massive adoption of Electronic Health Records (EHRs) and the ensuing big data, there is an opportunity to apply machine learning methods to investigate this challenge. Extract- ing clinically important phenotypes from EHRs in an automated way has gained increasing attention as a promising approach to speedup findings.

KEYWORDS: OPEN-SET LEARNING, DEEP GENERATIVE MODELS, CANCER TREATMENT PREDICTIONS

E-mail of First Presenter: a-cao@u.northwestern.edu

Focus Session 5
Data Science – Population Health

Automatic discovery of complex interactions between multiple risk factors

Authors: William Baskett, Murugesan Raju, Chi-Ren Shyu

Abstract: Specific medical outcomes are often the result of many different factors acting together. In many cases, these factors do not act independently but rather the effect of one factor may depend on the presence/non-presence of another factor. To understand how different risk factors interact, a computational method which can not only discover these interactions but also present them in a way which is relatively simple for humans to understand is needed. We have developed a contrast pattern mining method capable of identifying complex logically structured interactions which are associated with specific outcomes using a large number of factors. These patterns of interactions can provide significant insight into how specific risk factors interact to influence the probability of developing a disease or how two similar cohorts of patients differ from each-other in key ways. We demonstrate that our method is capable of automatically discovering key interactions which affect medical outcomes using several different medical datasets, each illustrating a different kind of problem and containing different kinds of data. This data includes EHR data for several conditions and genetic data from individuals with Autism and their unaffected family members.

Statement of Significance: Risk factors for diseases often do not behave independently of each-other and it is necessary to understand the structure of these interactions to better understand the underlying causes of diseases and how to prevent and treat them. This method can identify complex logically structured patterns of risk factors which are disproportionately common or uncommon in specific cohorts of patients. This can be used to contrast patients with and without a specific disease to discover patterns of risk factors for that disease. It can also be used illustrate key differences between similar patients with a disease phenotype such as by contrasting individuals who developed symptoms for a specific chronic disease as a child vs those who developed symptoms later in life. The results produced by this method are useful for hypothesis generation and provide direction for more detailed future research focused on a smaller number of factors.

Keywords: Explainable AI, Contrast mining, Data mining
Email of First Presenter: wibpp9@mail.missouri.edu

Interactive Visualization for Medical Images in Jupyter Notebooks

Authors: Brian Pollack, Kayhan Batmanghelich, University of Pittsburgh

Abstract: Computer vision research using radiological imaging is a popular and important subject within the machine learning community. Data and image augmentation is commonplace for all types of computer vision, but these changes must be verified during the analysis process. The most straightforward manner to verify these changes is via visual inspection. However, dealing with unwieldy data formats, multidimensional images, and additional layered annotation or metadata can be frustrating for researchers and impede progress for a given analysis. Useful tools exist to view medical images, but typically require a researcher to leave their work environment in order to use stand-alone software packages.
Leveraging new and powerful Python packages such as xArray, Holoviews, and Dask, we have developed a simple package that can be used for real-time display and manipulation of multidimensional medical images, segmentation masks, and more, for an entire patient population, without ever leaving an active Jupyter Notebook. Not only does this tool improve medical image workflows, it also facilities communication with medical experts due to the inclusion of features that are common among radiology workstations.

Statement of Significance: There is a lack of tools in the python ecosystem specifically for medical image analysis via machine learning. This software uses new and powerful data visualization packages to greatly improve the ability to perform image analysis in an interactive, real-time, and self-contained manner that is not possible using basic matplotlib functionality.

Keywords: Computer Vision, Bioinformatics, Radiology, Data Visualization
E-mail of First Presenter: brp98@pitt.edu

Exploring Administration Time-of-Day Efficacy for Seasonal Influenza Vaccinations

Authors: Darwin Y Fu, Jacob J Hughey, Vanderbilt University

Abstract: Despite evidence that many medication targets have biological rhythms, a patient's circadian rhythm is rarely considered in determining administration times for medical interventions. In particular, there are significant circadian oscillations of immune system components. One important, regularly administered immune intervention is seasonal influenza vaccinations. Previous work have observed higher antibody titers in older adults with morning vaccinations compared to those with afternoon vaccinations, but efforts to expand findings to multiple strains and age groups have found mixed results. We hypothesized there is a time of day effect in influenza vaccine efficacy, and conducted a retrospective study using administration time data extracted from electronic health records.
We collected seasonal influenza vaccination data from Vanderbilt University Medical Center with follow up on whether or not patients were diagnosed with influenza. We fit a mixed effects Cox regression model with predictors sex, age, a cosinor component to represent time, and a random effect representing flu season. Preliminary results indicate a slight but statistically significant increase in relative risk for patients vaccinated in the afternoon. Work is ongoing to validate influenza phenotyping and account for possible confounders such as previous patient health.

Statement of Significance: Since 2010, seasonal influenza results in upwards of 45,000,000 illnesses and 800,000 hospitalizations each year. The flu creates a significant public health burden, particularly among vulnerable populations. Seasonal flu vaccines are a major preventative intervention, but vaccines can vary significantly in efficacy from year to year. One opportunity to increase efficacy is in optimizing time of administration. Previous studies have found higher immune responses to morning vaccinations for older adults, but limited epidemiological evidence exists for a broader population. We utilized electronic health records to conduct a retrospective study of vaccine efficacy accounting for time of administration, year, age, sex, and health of patients.

Keywords: Circadian Rhythms, Immunity, Health Informatics
E-mail of First Presenter: darwinyfu@gmail.com

EVALUATING ELECTRONIC HEALTH RECORD USE AMONG ACADEMIC OPHTHALMOLOGISTS

Authors: Sally L. Baxter, Mitul Mehta, Scott E. Rudkin, John Bartlett, James D. Brandt, Cathy Q. Sun, Marlene Millen, Christopher A. Longhurst

Abstract: Ophthalmologists have high-volume practices that demand efficiency working with electronic health records (EHRs). Several EHR systems aggregate audit log data to generate metrics describing time spent on various tasks. The aim of this study was to leverage these metrics to better understand ophthalmologists' EHR use, and particularly after-hours EHR use as that is a well-documented source of burnout. We aimed to determine which personalization features or efficiency tools within the EHR were associated with decreased after-hours EHR use in order to inform future strategies for optimizing efficiency. EHR use data were obtained over 12 months (11/2018-10/2019) for 139 academic attending ophthalmologists at 5 institutions using the same vendor. Metrics included time spent on activities such as notes, orders, clinical review, and in-basket. After-hours use was approximated using metrics describing time in the system outside periods with scheduled appointments. Ophthalmologists' use of EHR personalization features and efficiency tools were also analyzed. Linear mixed effects models were used to evaluate which factors were associated with decreased after-hours EHR use. Ophthalmologists (n=139) spent a mean (SD) of 76.3 (36.0) minutes (min) in the EHR per day. On average, ophthalmologists spent the most time on notes per day (31.9 [21.0] min), followed by orders (11.6 [7.9] min) and clinical review (7.4 [5.5] min). Mean time outside scheduled hours was 28.1 (21.1) minutes. Entering medication orders on the vendor's mobile client was associated with significantly less time spent on after-hours EHR use (coefficient -13.40, $p<0.001$). Most other EHR efficiency tools were not associated with decreased after-hours EHR use. By quantifying time spent on EHR activities both during appointments and after-hours, EHR use metrics can enable analyses providing insights into variations in EHR use and inform future efforts to promote efficiency of EHR use and physician satisfaction.

Statement of Significance: EHR audit log data can be used to better understand workflows and to capture after-hours EHR use, which has been associated with physician burnout and is not easily attainable by traditional time-motion studies. These data can be used to improve ongoing EHR implementation and optimization efforts.

Keywords: ophthalmology, electronic health records, audit logs, burnout, clinical informatics
Email of First Presenter: s1baxter@health.ucsd.edu

Investigating the Link between Obstructive Sleep Apnea and Air Quality: What can we learn from CPAP device data?

Authors: Jay P. Kitt*, Krishna Sundar, Ramkiran Gouripeddi, Cheryl Pirozzi, Julio C. Facelli. The University of Utah

Abstract: Obstructive sleep apnea (OSA) is a common sleep disorder which affects an estimated 22-million Americans. In patients with OSA, periods of airway closure occur throughout the night resulting in a hypoxic sleep state which contributes increased risk of many conditions including heart disease, depression, stroke, and dementia. Inflammatory events, which occur following acute exposures to irritants lead to an increase in apnea. Exposure to air pollution has been linked to airway inflammation suggesting that air-pollution is likely to negatively impact OSA. First-line treatment for OSA is use of a CPAP device which keeps the external air supply at higher pressure to help maintain an open airway. These log of apnea events and report daily AHI, a measure of apnea events per hour, for every night of active use. In this study, apnea data collected from CPAP devices of 4000 patients was compared with EPA atmospheric pollutant concentration data. Simple or time-lagged regression of AHI severity and pollutant data showed little correlation, however classification of "good days," where no notable apnea changes occurred, and "bad days," days where more than 10% of patients had an AHI greater than two standard deviations above their mean, were categorized using a Random Forest classifier with 96.7% accuracy. Examining the important factors utilized by the algorithm allowed identification of pollutants

which drove selection of bad days. Those include PM 2.5, carbon monoxide, and ozone. Through further analysis, it was discovered that pollutant-sensitive sub-populations drive categorization. Overall, some level of pollutant sensitivity impacts ~50% of CPAP patients. Those with sensitivity average a 26% higher AHI and have 46% more clear-airway apneas. These results demonstrate the impact of air pollution on OSA and provide patients and physicians information to make care adjustments for the future.
In this work CPAP device data is utilized for the first time to study the impact of air pollution exposures on obstructive sleep apnea. Machine learning techniques are utilized to predict days where a large fraction of the CPAP treated population will suffer worse apnea episodes based entirely on pollutant concentration profiles with 96.7% accuracy. Specific pollutant-sensitive populations are identified through machine learning and statistical analyses. The results of this work demonstrate the impact of air pollution on OSA and provide patients and physicians information necessary to make care adjustments for the future.

Keywords: Translational Clinical Informatics, Exposure Science, Air Pollution, Apnea
E-mail of First Presenter: jay.p.kitt@gmail.com

Patient Perspectives on Using Conversational Agents for Hypertension Medication Self-Management
Authors: Ashley C Griffin, Zhaopeng Xing, Sean P Mikles, Arlene E Chung, University of North Carolina at Chapel Hill

Abstract: Conversational agents or chatbots, which are computer systems that converse with people, have demonstrated the potential to deliver self-management interventions in an engaging manner. However, little is known about patient perspectives towards using conversational agents to support medication self-management in hypertension. The objective of this study was to assess information needs and preferences for using conversational agents to support hypertension medication self-management.
Semi-structured interviews were conducted to assess hypertension self-management needs and preferences for using a conversational agent. Purposeful sampling was used to select adults with hypertension on at least one medication based on age, race, gender, education, and number of prescription medications. Participants were shown an example of a commercial chatbot that focused on health and then perceptions and barriers and facilitators were elicited using an interview guide. Interviews were audio-recorded, transcribed verbatim, and imported into NVivo software. Transcripts were analyzed using applied thematic analysis by two reviewers (AG & ZX), and discrepancies were adjudicated by a third reviewer (SM).
15 participants were enrolled for interviews out of 145 eligible participants. Mean age was 59 (SD = 11), and 53% of participants were female, 66% were Caucasian, and 60% had at least a college degree. The majority of participants (60%) had been diagnosed with hypertension for ≥5 years, took 6 prescription medications (SD = 4), but only 20% had used a chatbot before. Preliminary results reveal that chatbots were viewed positively, and most participants reported they would like to use a chatbot to help track their medications, refills, and blood pressures. Some felt that chatbots may provide too much information or might be intrusive if the recommendations were focused on lifestyle modification. These initial findings contribute towards a more nuanced understanding of patient perceptions towards conversational agents. Additional analysis in progress is exploring information needs and desired system functionalities.

Statement of Significance: Hypertension is the most common chronic disease in the U.S. and leading risk factor for heart disease. While many factors are associated with uncontrolled hypertension, poor medication self-management is an important contributor. Approximately 75% of adults with hypertension are taking

medications, but only half have blood pressures that are under adequate control. Conversational agents confer a number of advantages to support medication adherence beyond other digital health approaches (e.g., mobile apps, text messaging) as these agents can mirror therapeutic processes such as motivational interviewing or cognitive behavior therapy and deliver interactive education. Conversational agents for health have demonstrated improvements in self-management behaviors, including knowledge, skills, and health outcomes, but have not been utilized in hypertension self-management. Thus, this research seeks to develop a better understanding of users' needs and perceptions to inform the design of a chatbot intervention to support blood pressure management and medication adherence.

Keywords: Consumer health informatics, conversational agents, self-management, hypertension
E-mail of First Presenter: acgriffi@live.unc.edu

Characterizing Patterns of Disease Diagnosis Across Men and Women

Authors: Tony Y. Sun, Harry Reyes Nieva, Noémie Elhadad, Columbia University

Abstract: With the push from the National Institutes of Health's to consider sex as a biological variable, we sought to characterize systematically the patterns of disease diagnosis across men and women at a tertiary academic medical center.

We selected patients with at least one year of observations between 2009-2019 at Columbia University's Irving Medical Center (335,242 men and 470,898 women). To characterize disease diagnoses, a given condition's diagnosis was approximated as the first occurrence of its SNOMED-CT code in a patient's timeline. These resulted in 5,816 codes in our cohort. For each diagnosis, its prevalence difference and mean age at onset difference between men and women were computed. To characterize differences in diagnosis according to their symptoms, we selected 15 disease-specific cohorts. We collected all "presenting codes" — SNOMED-CT codes that occurred within three years of each disease diagnosis (11,000 codes)— and computed differences in their prevalence and time to diagnosis between men and women.

The prevalence of conditions was higher for women (mean prevalence difference across conditions: 0.045%, p=3.097e-16). In 68% of the 5,816 conditions, women were diagnosed when older (mean age-at-disease-onset difference across conditions: 2.4 years). In 62% of all presenting codes, women were diagnosed later than men with the same symptom (mean difference of time to diagnosis: 18.53 days).

There are systematic differences in patterns of diagnosis between men and women across all conditions. While it is expected that sex-specific conditions would be diagnosed differently, this study suggests that different symptoms of disease are weighed differently across sexes at the time of diagnosis. Women are diagnosed later than men (mean difference in age: 2.4 years), all the while presenting with symptoms more prevalently than men (mean difference in prevalence: 0.045%) for a longer time (mean difference in time to diagnosis: 18.53 days).

Statement of Significance:

We are in the process of replicating our study on a large national claims dataset. Our analysis suggests that for a wide range of conditions, women are diagnosed at an older age and with a longer gap to diagnosis than men with the same symptoms. There are multiple implications to this work. From the standpoint of analysis of observational health data, stratification by sex seems critical to ensure reliable analysis. Machine learning algorithms trained on electronic health record data must also account for differences in prevalence, time to diagnosis, and time between symptom presentation and diagnosis.

Keywords: Observational health data, characterization study, gender bias
E-mail of First Presenter: tys2108@cumc.columbia.edu

Focus Session 6
Pharmacoinformatics

TOWARD *IN VIVO* MOLECULE EFFECT PREDICTION WITH DEEP LEARNING AND FAST PROTEOMICS

Authors: Jesse G Meyer, Natalie M Niemi, David J Pagliarini, Anthony Gitter, Joshua J Coon, University of Wisconsin-Madison

Abstract: Measuring chemical effects in biological systems is slow and costs billions of dollars annually. Prediction of chemical effects in silico would increase the speed and decrease the costs of toxicity testing and drug development. Machine learning has been successfully applied to predict specific types of toxicity, and more recently, for targeted drug development. However, target-agnostic prediction of chemical effects in biological systems is needed to discover cures for complex diseases and discover unexpected types of toxicity. We describe advancements that bring general target-agnostic chemical effect prediction closer to reality. First, a deep learning strategy was developed for general molecular effect prediction (MoEPred) in cultured cells; changes in the biological "omic" states induced by chemicals are predicted using those chemical's structures as model input. Here the biological omic states are represented by the collection of molecules that are present, for example the transcriptome, proteome, or metabolome. Proof of concept MoEPred results were generated using public transcriptomic changes induced by thousands of chemical perturbations (LINCS project L1000 assay data). The second contribution herein is a method to quickly generate proteomic data on the scale of minutes per sample using mass spectrometry (MS), which is needed for training MoEPred models. This method eliminates time consuming liquid chromatography analyte separation and instead achieves separation quickly in the gas phase. To demonstrate the utility of this new fast proteomic strategy, in only 4.4 hours of MS data collection, 132 proteomes were collected from various combinations of genotype, nutrients, and mitochondrial toxins. A total of over 39,000 non-unique proteins were quantified in total at a rate of >3 proteins/second. These experiments revealed protein changes specific to each of the factors, and provide several testable mechanistic hypotheses. Together, this computational framework for MoEPred with the fast omic data collection will enable faster drug discovery and toxicity prediction.

Statement of Significance: The ability to predict changes in cellular states that result from chemical treatment with MoEPred will revolutionize the way we develop drugs and test chemical toxicity. Minute-scale proteomic data collection is not only useful for building a MoEPred training set, but also for standard phenotypic drug screening and biomarker discovery from large clinical cohorts.

Keywords: Cheminformatics, machine learning, toxicity, proteomics, drug discovery
E-mail of First Presenter: jgmeyer2@wisc.edu

Leveraging Partitioning Statistics for Scalable Microbial Identification

Authors: Pavan K Kota, Daniel E LeJeune, Rebekah A Drezek, Richard G Baraniuk, Rice University

Abstract: The rapid identification of pathogens for infection diagnostics must be both scalable and robust to sudden genomic mutations and emergent species. While nucleic acid sequencing provides universal information from a sample, it remains costly and time-intensive for routine use. We report our ongoing research towards the use of arbitrary, nonspecific DNA sensors for microbial genomic fingerprinting. While most probe-based microbial identification strategies depend on highly specific sensing mechanisms, our sensors each interact with many locations along each copy of microbial genomes. We show that the use of a handful of sensors can reliably discriminate between pathogens and that the signal processing technique known as *compressed sensing* can be used to unmix and quantify fingerprints in samples with a few pathogens. Moreover, recent technologies in sample partitioning enable the isolation of at most a few cells per partition (e.g. a droplet or nanowell). By splitting a sample over thousands of such partitions, much more information can be gleamed, and inevitable background contamination can be isolated and excluded from analysis. Most partitioning techniques have depended on highly dilute samples for single-cell characterization, meaning that concentrated samples may saturate the system with multi-cell partitions. We report *in silico* results of a novel compressed sensing algorithm that, in combination with our genomic fingerprinting sensors, can resolve multi-cell capture for an improvement in dynamic range of 1-2 orders of magnitude. Lastly, we discuss how conventional signal amplification strategies can be adjusted for nonspecific sensing to enable single-cell sensitivity in practice. We present a preliminary *in vitro* demonstration of these techniques and the upcoming challenges.

Statement of Significance: There is an urgent need for a rapid diagnostics platform that offers an alternative to time-intensive sample culture. Clinicians are often forced to preemptively prescribe broad-spectrum antibiotics while waiting on results which inevitably engenders drug resistance. By merging ideas from signal processing and biosensing, our research takes an unconventional approach towards a scalable and robust diagnostic platform capable of routine use across many types of infection. As sequencing becomes increasingly prevalent, large public databases can augment the capability of our platform because our DNA-based sensors' signals can be thermodynamically predicted from sequence. Fingerprints can be updated in real-time based on pathogen evolution, obviating the need for the re-design and distribution of a new diagnostic tool for every new pathogen. Our research will therefore synergize with the shifting technology landscape while serving a critical function for clinicians.

Acknowledgement:
This work is supported by the NLM Training Program in Biomedical Informatics and Data Science T15LM007093, Program Director Dr. Lydia Kavraki, and by the Rice University Institute of Biosciences and Bioengineering.

Keywords: Nucleic acids, data-driven diagnostics, sequence analysis, signal processing
E-mail of First Presenter: Pavan K Kota, pkk1@rice.edu

cando.py: Software for Drug Candidate Prediction with Application to SARS-CoV-2

Authors: William Mangione, Zackary Falls, Ram Samudrala, University at Buffalo

Abstract: Elucidating drug-protein interactions is essential for understanding the beneficial effects of small molecule therapeutics in human disease states. Traditional drug discovery methods focus on optimizing the efficacy of a drug against a single biological target of interest. However, evidence supports the multitarget theory, i.e., drugs work by exerting their therapeutic effects via interaction with multiple biological targets.

Analyzing drug interactions with a library of proteins provides further insight into disease systems while also allowing for prediction of putative
therapeutics against specific indications. We present a Python package for analysis of drug-proteome and drug disease relationships implementing the Computational Analysis of Novel Drug Opportunities (CANDO) platform. The CANDO package allows for rapid drug similarity assessment, most notably via the bioinformatic docking protocol where billions of drug-protein interactions are rapidly scored and the similarity of drug-proteome interaction signatures is calculated. The package also implements a variety of bench-marking protocols to determine
how well drugs are related to each other in terms of the indications/diseases for which they are approved. Drug predictions are generated through consensus scoring of the most similar compounds to drugs known to treat a particular indication. Support for comparing and ranking novel chemical entities, as well as machine learning modules for both benchmarking and putative drug candidate prediction is also available. We applied cando.py to the recent pandemic of SARS-CoV-2, showcasing its ability to quickly and efficiently suggest novel drug candidates
that may be effective against the virus. Two pipelines are utilized: the above consensus scoring protocol using drugs identified via two high-throughput screens as actives, and a method that uses the aforementioned docking protocol to rank drugs based on their predicted binding affinities to SARS-CoV-2 proteins. The CANDO Python package is
available on GitHub at https://github.com/ram-compbio/CANDO, through the Conda Python package installer, and at http://compbio.org/software/.

Statement of Significance: The cando.py package shows promise in helping to alleviate the time and monetary burdens associated with modern drug discovery. By considering drugs in multitarget context, we can better describe how they behave in biological systems and therefore progress towards a deeper understanding of the science of drug discovery. The application to SARS-CoV-2 provides a pertinent example of how our platform may be utilized to combat complex and urgent issues in medicine.

Keywords: drug discovery, drug repurposing, bioinformatics, computational biology, proteomics, SARS-CoV-2, COVID-19
E-mail of First Presenter: wmangion@buffalo.edu

Identifying Neoantigens From Noncanonical Transcription and Translation Events at Complex Structural Variants in Human Cancers

Authors: Vinayak Viswanadham, Simon Chu, Victor Mao, Isidro Cortes-Ciriano+, Peter J Park+ (+ = co-corresponding)

Abstract: Recent whole-genome analyses have revealed pervasive complex structural variants (SVs) across human cancers, and how extensively do SVs contribute to tumor neoantigen loads remains a critical question. To investigate whether SVs generate functional templates for neoantigens, we characterized transcription at 86,612 SVs detected from whole-genome and RNA sequencing data of 582 tumors across 16 tumor types from the PCAWG/TCGA project. In ~11% of SVs (n=10,080), we observed intergenic transcription at SV-generated fusion sequences within 0.5 kb of the ends of breakpoints. Using LC-MS/MS data from the CPTAC/TCGA project matched to breast, colon, ovarian,

and rectal cancer samples; we identified on average 3-4 high-confidence peptides per sample uniquely originating from fusion sequences, with considerable variation among tumor types. To examine whether SVs trigger transcription and translation outside of annotated reading frames, we identified 36,719 strongly transcribed intergenic regions within 100 kb of SVs across our tumor samples, with ~3% of intergenic regions per CPTAC tumor type yielding proteomically-backed peptides, with 100-300 high-confidence intergenically-derived peptides identified per tumor sample. We causally linked ~1% of transcribed intergenic regions to local SVs and confirmed proteomic support for a subset of these regions. We also identified recurrent intergenic transcription and translation across cell lines from the CCLE and ENCODE collections using analyses of WGS, RNA-seq, and MS/MS datasets from matched cell line types and demonstrate signals of transcriptional initiation and ribosomal loading within intergenic space in a subset of samples. Our analysis demonstrates that complex SVs and intergenic transcripts contribute to novel peptides that could be exploited as putative neoantigens for immunotherapies and studied further for effects upon tumor biology.

Statement of Significance: We have mapped out rates of transcription and translation arising from structural variants and at intergenic regions across several cancer types, and we have linked a subset of intergenic transcription and translation events to the presence of nearby structural variants. Our findings suggest that dramatic genomic rearrangements can generate *de novo* transcriptional events in intergenic space, which we found to contribute significantly to the tumor transcriptome and peptidome. Intergenically-derived and SV-derived peptides represent an untapped source for neoantigens in future immunotherapeutic approaches and a potentially important component of the tumor proteome.

Keywords: Cancer, Proteogenomics, Bioinformatics/Computational Biology
E-mail of First Presenter: vvv776@g.harvard.edu

Investigating the Antibiotic Biosynthetic Potential of the ADT Microbiome

Authors: Reed M Stubbendieck, Cameron R Currie, Department of Bacteriology, University of Wisconsin-Madison

Abstract: The aerodigestive tract (ADT) is comprised of the respiratory and upper digestive tracts and is the primary portal through which pathogens and other microbes enter the body. To successfully colonize the ADT, bacteria must compete with other microbes for scarce resources, including nutrients and physical space. Competition may occur indirectly by preventing competitors from accessing resources or directly through the production of antibiotics or other secondary metabolites that inhibit the growth of competitors. Within microbial genomes, the genes responsible for production of secondary metabolites are organized into biosynthetic gene clusters (BGCs), which encode the proteins necessary for biosynthesis, gene regulation, export, and self-resistance. As the ADT directly faces the external environment, we hypothesize that bacteria colonizing the ADT may produce antibiotics or other metabolites involved in competition with both co-occurring and invading microbes. As a first step to address our hypothesis, we identified secondary metabolite biosynthetic gene clusters (BGCs) from the expanded human oral microbiome database (eHOMD) using the antibiotics & Secondary Metabolite Analysis Shell (antiSMASH). From the eHOMD, we have identified 3996 BGCs

within 1570 bacterial genomes, representing 475 different bacterial taxa comprising 62% and 85% of all known and cultured bacteria, respectively, colonizing the ADT. To determine if different BGCs are characteristic of bacteria that colonize distinct sites (e.g., nasal cavity, mouth, and pharynx) within the ADT or are associated with different disease states, we are currently aligning metagenomic reads from the human microbiome project and other studies onto the ADT BGCs. We hypothesize that specific BGCs may be enriched or depleted, when comparing the microbiomes of healthy and sick individuals. The bacteria that harbor BGCs of interest will isolated and directly assessed for their ability to inhibit respiratory pathogens of interest.

Statement of Significance: The human microbiota provides essential services that include aiding digestion, producing vitamins, and pathogen defense. Administration of traditional broad-spectrum antibiotics may disrupt the healthy human microbiota and predispose individuals to life-threatening diseases, such as Clostridium difficile infection. To address this problem, it is necessary to develop narrow-spectrum antibiotics that target specific pathogens and are harmless against beneficial bacteria. As antibiotics are thought to function as a form of chemical warfare between competing microbes, we hypothesized that human symbiotic bacteria occupying the same niches as pathogens have been evolutionarily selected to produce molecules that both inhibit pathogens and are not harmful toward the human host or its beneficial microbiota. Our analysis will provide a first indication of whether specific BGCs are underrepresented in the metagenomes from sick children relative to healthy controls. The bacteria harboring these BGCs will become candidates for producing antibiotics and will be isolated and investigated further

Keywords: antibiotics, microbiome, microbial interactions
E-mail of First Presenter: stubbendieck@wisc.edu

The Unreasonable Effectiveness of Naïve Bayes for Predicting Combinatorial Drug Effects
Authors: Hannah A. Burkhardt, Devika Subramanian, Justin Mower, Trevor Cohen, University of Washington

Abstract: Drug combinations, rather than individual agents, best treat many cancers by achieving a treatment effect that is greater than the sum of its parts; this is called drug synergy. Because of the combinatorial explosion of possibilities, clinical evaluation of all 'drug cocktails' is not feasible, and there is an urgent need for methods prioritizing promising combinations for evaluation in animal models and clinical trials. Consequently, there has been a burgeoning interest in the development of computational models to predict combinatorial effects, with some authors reporting high predictive performance. Facilitating this computational exploration, the NCI-ALMANAC data set comprises in-vitro tests of the effectiveness of combinations of 104 different drugs at inhibiting the growth of 60 cancer cell lines. Previous work has explored modeling this drug synergy data to predict novel treatment options, with potential applicability to previously untested drugs or cell lines (including patient-derived cellular models). However, it is unclear to what extent this performance emerges from capturing combinatorial effects compared to information implicit in the datasets used for training and evaluation. In this work, we demonstrate the strikingly high performance of the Naïve Bayes (NB) probabilistic model, which assumes conditional independence of features, showing that high predictive accuracy can be achieved based on class priors - without modeling combinatorial effects at all. We then demonstrate the utility of class priors as calculated by NB in improving the performance of an explicitly combinatorial model, using a method named Embedding of Semantic Predications (ESP), in a simple but effective ensemble approach. This work indicates that the ALMANAC dataset, which has been used to study combinatorial effects, can be modeled under an independence assumption with high performance. However, combining this class prior-based

predictive power with explicit modeling of combinatorial effects further improves performance on the task of recovering held-out synergistic combinations.

Keywords: Pharmacoinformatics; cancer; pharmaceuticals; machine learning
E-mail of First Presenter: haalbu@uw.edu

Open-Mic Session 1
Data Science and Bioinformatics

Dissecting the Tumor Ecosystem Mediating Metastatic Breast Cancer

Authors: Emily K. Stroup[1,2], Hongbin Wang[2], Zhe Ji[2,3]
[1]Driskill Graduate Program in Life Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL
[2]Department of Pharmacology, Feinberg School of Medicine, Northwestern University, Chicago, IL
[3]Department of Biomedical Engineering, McCormick School of Engineering, Northwestern University, Evanston, IL

Abstract: In this project, we aim to decode the gene transcriptional network underlying the tumor microenvironment driving breast cancer metastasis, by cutting-edge high-throughput genomic experiments and computational modeling. We will integrate high-dimensional genomics data from patient samples and synergistic mouse models using machine learning and deep learning algorithms to develop new biological insights on inflammation within the microenvironment and the progression of tumor metastasis. We will generate and utilize unbiased information collected from DNA-seq, single-cell and bulk RNA-seq, ATAC-seq, ChIP-seq, and flow cytometry for these samples. We hope to reveal the transcriptomic and epigenetic changes within cancer cells and immune cells (such as macrophages, neutrophils, and myeloid-derived suppressor cells (MDSCs)), that create the immunosuppressive microenvironment and mediate breast cancer metastasis. Once we identify the key regulators through this integrative analysis, we will develop new methods to push the tumor into an immunologically "hot" state where the cancer cells will be more susceptible to cytotoxic immune responses and checkpoint inhibition.

Statement of Significance: While the overall prevalence of breast cancer is decreasing in the United States, the incidence of metastatic disease is rising and survival rates remain below 40%. Although the genetic drivers and expression changes in metastatic tumors have been heavily studied, the regulatory roles of the tumor microenvironment remain poorly understood. The tumor microenvironment is a dynamic ecosystem composed of tumor cells, immune cells (such as CD8+ T cells, CD4+ T cells, B cells, NK cells, dendritic cells, MDSCs, and macrophages) and stromal cells (such as fibroblast, adipocytes, and endothelial cells). These cells interact with each other through secreted cytokines, chemokines, and paracrine signaling. It is increasingly recognized that the tumor ecosystem plays a critical role in mediating cancer progression. For example, tumor cells can activate NF-kB, STAT3, and AP-1 signaling thereby suppressing anti-tumor immunity. This inflammatory response can reshape the microenvironment to create a pro-tumorigenic niche that directs cancer progression and metastasis.

Keywords: bioinformatics, computational genomics, multi-omics, metastasis, breast cancer
E-mail of First Presenter: EmilyStroup2023@u.northwestern.edu

Deep Generative Graph Neural Networks for de Novo Drug Design and Optimization
Authors: Benjamin Kaufman, Sebastian Raschka, Anthony Gitter, University of Wisconsin-Madison

Abstract: The current pipeline to bring a drug to market is incredibly expensive and time-consuming, with much of the development money being spent on screening large chemical compound libraries on the drug target of interest. Predictive virtual screening methods attempt to reduce the cost of this process, but can only be applied to compounds available in screening libraries. Unfortunately, synthesizable compounds with the desired properties may not be present in these libraries. We attempt to address this issue by training various deep generative network architectures directly on molecular graphs. With an appropriate choice of loss function and techniques in optimization and self-supervised learning, we can produce novel compounds with desired molecular properties. Preliminary results for these methods are promising and further work is being done to evaluate their performance on different properties of interest. If these methods prove effective on general chemical properties, they serve to improve the efficiency of the drug discovery process.

Statement of Significance: These methods contribute to the field of de novo molecule design for drug discovery. What sets these methods apart from other methods in the field are the ways in which we incorporate constraints and self-supervised learning so the model learns to generate compounds from the chemical space that can actually be synthesized in practice.

Keywords: Machine Learning, Drug Discovery, Generative Models
E-mail of First Presenter: bkaufman3@wisc.edu

Imputation of Sparse Metabolomic Data
Authors: Elin Shaddox, Debashis Ghosh, Katerina Kechris, University of Colorado Anschutz Medical Campus

Abstract: We propose a multivariate biomarker discovery model, which accounts for varying levels of sparsity, or missing values, encountered in metabolomics data. Our multiple imputation approach incorporates a left censored model to account for non-random missingness attributed to limits of detection of instruments for mass spectrometry with gas or liquid chromatography (GC-MS or LC-MS). Our model is applied to an untargeted metabolomics study in plasma for chronic obstructive pulmonary disease (COPD), with the goal of discovering notable biomarkers. We also demonstrate the performance of our approach to improve biomarker detection and impute missing values compared to alternative approaches, through simulated datasets of varying levels and patterns of missingness.

Keywords: Left Censored Imputation, Metabolomics, Biomarker Discovery
E-mail of First Presenter: elin.shaddox@cuanschutz.edu

Modeling Histone Modifications at a Single-Cell Level Using EpiTOF
Authors: Laurynas Kalesinskas[1], Michele Donato[1], Ananthakrishnan Ganesan[1], Purvesh Khatri[1]
[1]Center for Biomedical Informatics Research, Stanford University

Abstract: Post-translational histone modifications of chromatin are a key regulator of many biological processes. However, their role in disease and in different immune cell types has been largely unexplored, due to lack of data at a single-cell resolution. Recently, with the aid of the high-throughput Epigenetics profiling by Time-Of-Flight method (EpiTOF), we have been able to characterize levels of histone modifications in hundreds of samples from subjects at the single cell level. In this presentation, I will focus on preliminary analyses of this data, exploring histone modifications associated with age, sex and different immune cell types.

Statement of Significance: Elucidating the patterns of histone modifications at a single-cell level can help us understand regulation of genes and immune cell differentiation, which can lead to increased mechanistic understanding of disease and potential therapeutic targets.

Keywords: Epigenetics, Single-Cell, Histone Modification
Email of First Presenter: lkalesin@stanford.edu

Diversity in Pharmacogenomics: A Review Article
Authors: Roderick Gladney, PharmD, University of North Carolina at Chapel Hill

Abstract: Pharmacogenomics (PGx) is a branch of precision medicine that studies the impact of genetic variants on drug efficacy and toxicity in order to optimize and individualize treatment plans. By combining advanced biomedical tools such pharmacogenetic testing and big-data analytics, clinicians are now beginning to devise personalized treatment plans based on an individuals' genetic, biological or environmental profile with the goal of improving diagnosis and risk assessment, and ultimately clinical outcomes. Genome-wide association studies (GWAS) have been salient in identifying genetic variants and mitigating disease risk. However, participation of individuals of non-European ancestry is underrepresented leading to a growing health disparity in precision medicine. Inter-ethnic and racial differences in drug pharmacokinetics and/or drug receptor sensitivities can lead to different drug responses between populations. With now 132 pharmacogenomic dosing guidelines for 99 drugs and pharmacogenomic information included in 309 medication labels, it is imperative that efforts are made to not widen the existing gaps in health outcomes for minority populations. In this review, we will synthesize existing literature to outline the ethnic differences in genetic variations and the advantages of diverse populations in pharmacogenomic studies in providing broader clinical applications.

Statement of Significance: Precision medicine is an emerging domain within the health care field that aims to revolutionize conventional approaches to patient care. Recent pharmacogenomic studies have demonstrated that an individual's genetic profile can affect disease outcomes and mitigate drug efficacy and toxicity risks. However, most pharmacogenomic studies have been conducted almost exclusively on individuals of European descent, thereby limiting the broader application to minority populations. The clinical implications of the lack of diversity in pharmacogenomics may lead to misdiagnosis and poor treatment plan for minority populations. Consequently, this further widens existing gaps in health disparities and outcomes for individuals of non-European ancestry.

Keywords: Pharmacogenomics, Precision Medicine, Diversity
E-mail of First Presenter: rodeglad@email.unc.edu

Multidrug Resistant Organism Carriage in Wisconsin Children

Authors: Ashley Kates, Nathan Putman-Buehler, Lauren Watson, Tamara LeCaire, Kristen Malecki, Paul Peppard, Ajay Sethi, Ellen Wald, Julie Mares, Dan Shirley, Garret Suen, Nasia Safdar, University of Wisconsin School of Medicine and Public Health

Abstract: Children attending daycare are at increased risk of carrying multidrug resistant organisms (MDROs) compared to children not attending daycare. Carriage of MDROs greatly increases the risk of infection, not only in the child but also for others living in the household. Here we present the findings from a cross-sectional study assessing MDRO carriage in daycare-attending and -non-attending children in Wisconsin.
Children between 6 months and <6 years of age were enrolled. Diet and risk factor history was collected. Samples were collected from the nares, axilla/groin, and stool and were cultured for common pathogens in children and MDROs. 16s rRNA sequencing of the v4 region for the nasal and skin samples and shotgun sequencing for the stool samples is ongoing. Skin/nasal samples will be assigned taxonomy using amplicon sequence variants to the genus level when possible. For stool samples, predicted ORF translations from our assembled metagenomes will be subjected to functional annotation using the Kyoto Encyclopedia of Genes and Genomes (KEGG) Automated Annotation Server to annotate each predicted protein according to the KEGG orthology. These will be mapped back to their respective metabolic pathways and normalized based on metagenome size.
Forty-four children were enrolled. The average age was 2.6 years and 50% were female. Twenty-three children (52.3%) were enrolled in daycare. Eighteen children were positive for at least one organism, nine of which had daycare exposure. In the stool, 6 children (13.6%, 2 in daycare) were *C. difficile* carriers, 3 were VRE carriers (6.8%, 1 in daycare), 8 carried an ESBL GNR (18.2%, 4 in daycare), and 3 carried MRSA (6.8%, 1 in daycare). There were no significant differences between children with and without daycare exposure for any organism. Children in this population had higher than expected rates of ESBL GNRs and MRSA for a community population. Sequencing is ongoing.

Statement of Significance: There is a paucity of data MDRO carriage and the microbiomes of children. However, we do know the microbiome of children is greatly impacted by behaviors – such as introduction of hands and objects into the mouth, close contact with floors, and contact with the skin of adults – especially during the early years of life. It has also been shown environmental factors have a greater influence on the microbiome of children than adults. Researchers have hypothesized there is a critical period during the early years of life that shape the microbiota and disruptions to the "healthy" microbiota may favor the development of disease and infection later on in life. Children are most frequently enrolled in daycare in the early years of life when their microbiota are most vulnerable. Therefore, there is a need for better understanding of the microbiome and MDRO colonization at various body sites in young children.

Keywords: Multidrug resistant organisms, children, daycare
E-mail of First Presenter: akates@medicine.wisc.edu

Synteny Maps Lead to High Quality Ortholog Predictions

Authors: Nicholas P Cooley & Erik S Wright, University of Pittsburgh

Abstract: Typically, shared order, or synteny, between two genomes is identified from predicted orthologous gene pairs. Groups of orthologous gene pairs that provide contiguous and sequential blocks in their respective

genomes indicate where blocks of genomic content and context are conserved. This conservation is useful for studying genome-scale events like rearrangements or inversions, or topics like ancestor genome reconstruction. Genomic context is also a useful tool for other purposes as well; prokaryotes tend to physically cluster genes involved in single or related functions, meaning that the genes involved in secondary metabolite construction are often physically co-localized in sequence space. This co-localization tendency is leveraged by tools like ANTISMASH, a pipeline for prediction of secondary metabolite gene clusters.
The foundational step of processes that use synteny as a tool for studying various research topics is precise prediction of orthologous gene pairs. This is problematic, as most orthology prediction strategies tend to be greedy by nature. Fortunately, this script can be flipped and genomic synteny itself can be predicted from significant and unique k-mers shared between genomes, instead of from orthologous gene pairs. Generating synteny maps from shared k-mers avoids costly all-vs-all alignments and the shared k-mers themselves serves as markers for which pairs of genes to evaluate as potential orthologs. Herein we present a method for predicting orthologous gene pairs using high quality synteny maps constructed using the R packages DECIPHER and SynExtend. This method is unique in that it is completely alignment free. It is also entirely contained within the R-Bioconductor environment, with comprehensive help files and vignettes designed for improved user accessibility.

Statement of Significance: Orthology prediction is a fundamental task in computational biology. This task is formally an assignment of the phylogenetic relationship between two or more sequence features in different genomes, but is more often equated with the simple question of 'do these two sequences share a common function?'. The nature of orthology prediction, and the historic nature of the data involved have made this a task where prioritization of sensitivity over specificity has been a logical tradeoff. However, with decreases in sequencing costs and the increase in available sequence data, that kind of permissive prediction becomes less necessary, and potentially problematic. Predicting orthology from synteny maps selects the opposite trade-off, prioritizing specificity over sensitivity. This approach, though conservative in pairwise prediction scenarios, allows for more a more complete analysis in n-way prediction scenarios.

Keywords: synteny, orthology, comparative genomics
E-mail of First Presenter: npc19@pitt.edu

A Grammar of Patient-Centered Communication Documentation

Authors: David Chartash; Mona Sharifi; Beth Emerson; Robert Frank; Cynthia Brandt; R Andrew Taylor

Abstract: Within the clinical encounter, *finding common ground* and *enhancing the patient-doctor relationship* are tasks facilitated by the overall concept of patient-centered communication. In the documentation process, the encoding process of language and culture of medical practice interferes with the translation of the spoken content of the clinical encounter into the medical record. This translation and the cognitive processes of writing by the provider are included in the concept of the medical sublanguage. Querying providers (32 physicians, 11 advanced practice providers) to articulate the components of the medical sublanguage which were representative of patient-centered communication in their own documentation, we have obtained the following preliminary results: The permutation of a noun and verb phrase, such that the discourse agent (represented as a noun phrase) is either preceded by a discourse action (verb phrase), or followed by transitive agreement or disagreement (verb phrase)

The combination of a noun and verb phrase, such that a discourse action (verb phrase) is preceded or followed by a set of nouns indicative of the components of uncertainty within shared decision making
Next, in order to build a model of patient-centered communication, we propose to develop a parallel grammar of patient-centered communication from spoken language within the clinical encounter. Reconciling these grammars, we intend to develop a more robust model of patient-centered communication. A computational language model will be developed to perform part-of-speech tagging on two sets of clinical notes, followed by an information extraction task facilitated by a context-free grammar built upon both provider and patient communication perspectives. Validating this model will take place using notes which have been shared with the patient (subject to OpenNotes policies) compared with notes which are not available to the patient; this comparison will allow for the validation of the specificity of the patient-centered grammar derived from patients.

Statement of Significance: This is a step towards understanding the language of both patient-provider communication and clinical documentation. Studying these language use contexts, we intend to facilitate the use of computational linguistic tools to understand clinical discourse; the medical sublanguage as expressed in complex logical form for clinical reasoning and patient-provider interaction. An understanding of language supports more robust semantic morphology within the medical sublanguage which takes into account the complex propositional and structural semantic nature of information used for clinical reasoning. In seeking to use the medical record as a retrospective source of semantic information for clinical reasoning research, therefore, we can model the clinical note as containing information beyond the lexical item, and provide character to the information within documentation as it contributes to the diagnostic process. In particular, this supports the evaluation of Entrustable Professional Activity 5 (Document a Clinical Encounter in the Patient Record) within undergraduate medical education.

Keywords: computational linguistics; patient-centered care; clinical documentation
E-mail of First Presenter: david.chartash@yale.edu

MREC: A Fast Framework for Aligning Single Cell Molecular Data

Authors: Andrew Blumberg, Mathieu Carrière, Michael Mandell, Soledad Villar, Raul Rabadan, Columbia University

Abstract: Comparing and aligning large datasets is a pervasive problem occurring across many different knowledge domains, and which is particularly relevant in single-cell genomics. We introduce and study MREC, a recursive decomposition algorithm for computing matchings between single-cell datasets. The basic idea is to partition the data, match the partitions, and then recursively match the points within each pair of identified partitions. The matching itself is done using black box matching procedures that are too expensive to run on the entire dataset. Using an absolute measure of the quality of a matching, the framework supports optimization over parameters including partitioning procedures and matching algorithms. By design, MREC can be applied to extremely large datasets. We analyze the procedure to describe when we can expect it to work well and demonstrate its flexibility and power by applying it to a number of alignment problems arising in the analysis of single cell molecular data.

Statement of Significance: Matching and alignment techniques based on optimal transport have recently been

applied to various problems and datasets in genomics. However, they are currently limited by the dataset sizes and cannot handle more than a few hundreds of cells. We are the first to propose a large-scale matching technique and explore a number of applications to datasets coming from single-cell characterization of different biological systems, showcasing the usefulness of our method for various problems in this field. The results we obtain outperform or are comparable to alternative algorithms on small datasets, and produce interesting results on datasets too large to be amenable to any other techniques.

Keywords: Genomics, Computational biology, Optimal transport
E-mail of First Presenter: mc4660@cumc.columbia.edu

Determining optimal patient selection for adjuvant chemotherapy after surgical resection in patients with Stage II and III colorectal cancer using computational modeling.

Authors: Justin J. Hummel[1], Yuanyuan Shen[1], Wesley C. Warren [1,2], Chi-Ren Shyu[1], Jonathan B. Mitchem[1,2,3]; [1]Institute for Data Science & Informatics, University of Missouri-Columbia, Columbia, MO; [2]Department of Surgery, University of Missouri College of Medicine, Columbia, MO; [3]Harry S. Truman Memorial Veterans' Hospital, Columbia, MO.

Abstract: Colorectal cancer (CRC) is the 2nd leading cause of cancer death in the US today. Patients with Stage II and III CRC are often considered for adjuvant chemotherapy after surgical resection. However, our current ability to predict the benefit of adjuvant chemotherapy is limited to a few simple clinicopathologic features. This is important both to select patients where recurrence can be prevented, but also to avoid needless toxic therapy in patients that do not receive benefit. In this study, we propose to utilize novel computational modeling in addition to advanced molecular features to address this critical issue in patient care. Using publicly available data that include clinicopathologic and molecular features we will conduct a comprehensive comparative analysis of applied statistical models as well as novel data mining techniques to identify potentially critical patient subgroups. We will utilize patient datasets that include patients that did and did not receive adjuvant chemotherapy. All models will be compared to existing models for prognosis and prediction using standard cross-validation methods while confirming statistical significance for both variables and results. At the completion of this study we will present a novel computational model to improve the prediction of patient benefit from adjuvant therapy after surgical resection in Stage II and III colon cancer patients. These results will support clinical providers in choosing the optimal therapy for their patients, moving closer toward the goals of precision medicine. The use of both currently available gene panels (FoundationOne CDx®) as well as unsupervised patient subgroup data mining allow us the potential to build a rapidly translatable model. Successful completion of this work will assist both patients and physicians and determining the best course of treatment, while ultimately clarifying the underlying mechanisms of treatment that have eluded us.

Statement of Significance: This research aims to assess patient populations using novel exploratory mining methods for subgrouping patients and sort responders from non-responders for clinical support assistance. Many of the molecular and clinicopathological attributes can be leveraged in a more conservative fashion that is deliberate. Our approach seeks a more effective, and efficient, statistical framework for measuring the probability of recurrence for a patient that receives adjuvant chemotherapy than existing methods. Avoiding unnecessary treatments that are harmful and painful improves the standard-of-care for mCRC patients and decreases the likelihood of recurrence.

Keywords: Adjuvant chemotherapy; statistical modeling; colorectal cancer
Email of First Presenter: jjh42v@mail.missouri.edu

Automated Node-Positive Classification and Feature Extraction Using Tumor Micro-Environment Imaging
Authors: Gautam Machiraju, Alex Derry, Sehj Kashyap, Parag Mallick, Sylvia Plevritis, Stanford University

Abstract: Detecting lymph node infiltration in cancer is critical for identifying the progression of cancer and determining patient prognosis. Novel multiplexed imaging platforms like CODEX have recently been developed which allow the detection and examination of subcellular structures in the tumor microenvironment using an array of cellular biomarkers. However, due to the novelty and richness of information in these images, there exist no algorithms for computer-aided detection of node infiltration in tumors. The amount of information provided in the cell marker channels provides a perfect opportunity for developing novel computer vision systems in order to improve our understanding of the factors involved in cancer progression and increase the accuracy of cancer screening. For the classification of lymph node positivity, we apply variations on the VGG19 deep convolutional architecture and introduce a novel transfer learning approach designed for operating on high-dimensional data, demonstrating good performance on a subimage patch level. By consensus voting among patches for each patient, we were able to correctly classify every patient in our held-out test set. Additionally, by computing saliency maps across the image channels, we identify the biomarkers that are most predictive of nodal involvement in cancer. Many of these markers are known to be associated with the progression of cancer to an aggressive state, providing a high degree of confidence in our model and the potential for enabling the prioritization of prognostic biomarkers.

Statement of significance: We demonstrate that multiplexed transfer learning (MTL) can learn latent features of correlated labels without explicit training and channel-wise saliency mapping (CSM) can enable feature selection agnostic of model architecture. The proposed model and preliminary methods for interpretability have the potential to advance informatics in the domain of multiplexed imaging. This approach could be deployed on *any* multiplexed image dataset or learning task, especially as this data type becomes more widely available with increasingly-adopted platforms such as CODEX, MIBI, image-based transcriptomics, CellDive, etc. Finally, these contributions also provide the groundwork for unsupervised methods for anomaly detection and representation learning for each patient sample.

Keywords: deep learning, image processing, cancer informatics, multiplexed imaging, geo-statistics
E-mail of First Presenter: gmachi@stanford.edu

Open-Mic Session 2
Clinical and Public Health Informatics

Improved physician follow-up using Post-Handoff Report of Outcomes (PHaROs)
Authors: Cindy C Reynolds, Robert El-Kareh, Cynthia L Kuelbs, University of California, San Diego

Abstract: Patient follow-up after transitions of care is known to improve physician learning and diagnostic calibration, but is difficult in settings such as the emergency department (ED) where physicians work in shifts and there is high patient turnover. To facilitate ED physician follow-up of patients, this study developed an algorithm to identify patients who had significant outcomes of interest after discharge from the ED and present the information to physicians in a curated list in the electronic medical record (EMR).
A daily automated algorithm was used to generate a list of patients that a provider had seen in the ED over the last 14 days. Patients on the list were flagged if they died, had an escalation of care (ICU admission, code, or rapid response), had a significant change of service (e.g. medical to surgical), re-presented to the ED in 72 hours, or had a child protective services (CPS) case filed. EMR tracking data on physician chart review was analyzed to detect whether there was increased access of charts after the intervention when compared to prior to the intervention.
Results are pending, but an interrupted time series analysis of the data will hopefully show that there was an increase in patient chart access after the intervention when compared to prior to the intervention.
If there is a significant increase in physicians accessing patient charts after the intervention, this would be the proof-of-concept for a scalable, easily-maintained solution to improve patient follow-up. Furthermore, this algorithm could be adapted for use in other settings such as night floats or surgery floors.

Statement of Significance: Numerous studies note the importance of patient follow-up for safety and physician learning, however, there are numerous barriers to follow-up including increased "shift" work by physicians. The electronic medical record (EMR) contains a wealth of patient follow-up data, but is poorly-structured for easy patient follow-up. This study provides proof-of-concept for a scalable, easily maintainable solution to facilitate information access and improve patient follow-up.

Keywords: clinical informatics, patient follow-up, medical education
Email of First Presenter: ccreynolds@health.ucsd.edu

A Convex Framework for Optimal Resource Allocation and Influence Maximization

Authors: Michael Liou, Po-Ling Loh. University of Wisconsin-Madison, Department of Statistics

Abstract: Dynamics over networks is becoming an increasingly important applied data problem to study with the wide prevalence of social networks and interconnected socioeconomic institutions. We study two interrelated optimization problems of these dynamics, influence Maximization (IM) and Optimal Resource Allocation (ORA). IM focuses on selecting some set of nodes to boost in order to maximize spread of information to the highest expected number of influenced nodes. ORA, on the other hand, focuses on selecting a set of nodes to distribute resources in order to prevent or contain information (or a virus) from spreading as efficiently as possible, thereby minimizing the expected number of influenced nodes. Here we consider a unified convex optimization framework that relate the two problems with approximate solutions and summarize our findings with rigorous simulations, characterizing the critical nodes that are identified by both types of problems. We illustrate the applicability of our approach in context of resource allocation to nodes to boost for targeted advertising or to support in the prevention of epidemic outbreaks.

Statement of Significance: This research problem has wide applicability in domains such as epidemiology, information security, marketing and many others. For example, due to a quickly globalizing world, the risk of epidemic outbreak is higher than ever and it's increasingly important for us to be well prepared for risk containment. An optimization framework for understanding these dynamics would help guide planning of how to optimize allocation of a limited number of resources in a dynamic network environment. A simulation that can be quickly solved with approximate solutions is also important to help provide insight into contingency plans, especially since real dynamics on networks are inherently random. Furthermore, having a unified framework for studying the inverse problem of increasing spread as much as possible would help better characterize general viral dynamics on network structures.

Keywords: Network, Optimization, Dynamical Systems, Epidemiology
E-mail of First Presenter: myliou@wisc.edu

EHR-based CDS Implemented to Deliver Public Health Information to Prevent and Manage Disease Outbreak Following the Inception of Meaningful Use – *A Scoping Review*

Authors: Jean Frédo Louis, MsIM, Janette Vazquez, MsCS, Catherine Staes, RN, PhD, Charlene Weir, RN, PhD, Guilherme Del Fiol, MD, PhD., The University of Utah

Abstract: The increasing adoption of electronic health record (EHR) systems has created opportunities for computerized clinical decision support (CDS). Meaningful Use regulations and EHR certification programs have incentivized the use of CDS to improve medical decision-making and patient outcomes. The impact of CDS in clinical care had been amply studied, and CDS in the clinical settings had been recognized to impact population health. However, less effort has been dedicated to enabling EHR-based CDS systems to deliver public health information to clinical settings, especially during disease outbreaks. The objective of this scoping review was to identify and characterize EHR-based CDS interventions that have been implemented to deliver public health guidance to prevent and manage disease outbreaks. We followed the PRISMA-ScR guideline for scoping review to search MEDLINE and four other databases. Each step was done in duplicate and independently. We classified the articles based on the type of public health interventions, diseases of interests, and CDS characteristics.
In this study, we identified the focus of the work done in the field and highlighted areas where more research is needed. The review can be used as a starting point to map and discuss the characteristics of CDS implemented to deliver public health guidance to healthcare practices in the context of an outbreak.

Keywords: clinical decision support, public health clinical guidance, electronic health record, diseases outbreaks
E-mail of First Presenter: jeanfredo.louis@utah.edu

Learning Tasks of Pediatric Providers from Electronic Health Record Audit Logs

Authors: Barrett Jones, You Chen, Vanderbilt University

Abstract: The amount of time spent working in the Electronic Health Record (EHR) has become a burden for many providers. Understanding provider activity patterns in the EHR can help healthcare organizations or EHR vendors to identify where EHR workflows can be improved to reduce burnout. To understand provider activities in the EHR, we propose an unsupervised learning framework to infer EHR tasks of Pediatric providers, including general pediatrics and pediatrics house-staff, in the treatment of normal newborns.
We investigated the EHR audit logs corresponding to 234 newborns who received care from 30 general pediatrics, and 63 pediatric house-staff providers from 3/1/2019 to 3/30/2019. The providers performed 37,644 EHR events. We leverage word2vec, k-means, and ProM process mining software on audit log data to learn EHR tasks for each of the two provider roles and visualize them.
Our analysis results show that pediatrics house-staff more commonly perform note preparation and results viewing relative to general pediatrics. General pediatrics perform more communication and chart review relative to pediatrics house-staff. Task learning analysis discovered 2 tasks for general pediatrics and 3 tasks for pediatric house-staff. The general pediatrics task focus on note creation and communication, respectively. Primary themes for pediatrics house-staff tasks are note creation, admit/discharge and order creation, and communication. Each task workflow was visualized using ProM process mining software.
We are successful in finding variation in EHR activity between the selected physician roles. Visualizations of task processes show that there are common tasks and task workflows within the EHR for the studied providers.

Statement of Significance: This work is a step towards understanding EHR tasks and task workflows using purely automated computational methods. Past studies are reliant on physician interviews or observation to understand tasks, in contrast our work can be automated and save time. Also, our approach can be used to learn tasks for a wide range of clinicians. Additionally, studies have focused on the learning of EHR tasks for physicians in the outpatient setting, by quantifying the efficiency of each task (e.g., EPIC signal – physician efficiency profile (PEP)). Our work provides the ability to apply these methods to the inpatient setting.

Keywords: Audit Log, EHR Task, Inpatient Setting, Newborn, General Pediatrics, Pediatric House-staff, Task Workflow, Process Mining, Electronic Health Records, Clinical Informatics
E-mail of First Presenter: barrett.jones@vanderbilt.edu

Examining Organizational Characteristics and Perceptions of Clinical Event Notification Services in Health Care Settings

Authors: Kevin Wiley, Jr.[1], MPH; Joshua Vest, PhD, MPH; Jessica Ancker, PhD. [1]Indiana University, Indianapolis, IN; [2]Regenstrief Institute, Inc.; Weill Cornell Medicine

Abstract: Introduction: Event notification systems are an approach to health information exchange (HIE) that provides real-time automated alerts to notify end users of patient interactions with the health care system.[1,2] Subscribers rely on data feeds that provide detailed patient health information derived from multiple sources during an encounter.[1] Existing research has demonstrated that event notification

services are associated with user satisfaction, organizational efficiencies, improve care coordination, and care quality.[1,3–5]

However, organizational factors may influence the usage and utility of event notification systems.[3] Organizational and clinical workflows, internal policies and procedures, and staff attitudes determine how, when, and for which patients event notification systems are most useful.[3,5] Previous studies have focused on the utility and user acceptance of event notification services provided by a single HIE organization.[1,6,7]

The purpose of this study was to examine user perceptions of event notifications systems provided by three HIE organizations. Using a questionnaire, we examined associations between organizational capabilities, event notification system use, and clinical and non-clinical staff perceptions on clinical efficiency, care quality, care coordination, and patient satisfaction.[8]

Methods: We surveyed healthcare professionals from 160 healthcare organizations that subscribed to event notifications through one of three HIE organizations. Respondent and organizational characteristics were quantified using frequencies and percentages. User perception was measured using Likert-scale questionnaire items. We conducted Exploratory Factor Analysis[9] to determine independent and dependent variables. Two dependent variables include 1) information and care quality and 2) care coordination. Independent variables include 1) policies and procedures, 2) knowledge management and quality, 3) patient consent, 4) staff support and attitudes, and 5) staff capacity. We conducted Pearson chi-squared and Student's t-tests to examine associations between independent and dependent variables. Mixed-effects regression models were estimated to account for clustered survey responses at respondents' organizations and qualified entities.

Statement of Significance: Measuring and examining adoption and advance use of health information technologies is still relatively new in the health services and informatics research disciplines. We sought to examine associations between event notification subscription services among healthcare organizations and user perceptions, care coordination and quality, and organizational characteristics. Such evidence is important in determining organizational characteristics that may facilitate or inhibit the use of health information technologies in the future.

Keywords: health information exchange, healthcare organizations
E-mail of First Presenter: kkwiley@iu.edu

References

1. Moore T, Shapiro JS, Doles L, et al. Event detection: a clinical notification service on a health information exchange platform. AMIA Annu Symp Proc 2012;2012:635–42.
2. Dixon BE, Schwartzkopf AL, Guerrero VM, et al. Regional data exchange to improve care for veterans after non-VA hospitalization: a randomized controlled trial. BMC Med Inform Decis Mak 2019;19(1):125.
3. Vest JR, Ancker JS. Health information exchange in the wild: the association between organizational capability and perceived utility of clinical event notifications in ambulatory and community care. J Am Med Inform Assoc 2017;24(1):39–46.
4. Anand V, Sheley ME, Xu S, Downs SM. Real time alert system: a disease management system leveraging health information exchange. Online J Public Health Inform [Internet] 2012;4(3). Available from: http://dx.doi.org/10.5210/ojphi.v4i3.4303
5. Unruh MA, Jung H-Y, Kaushal R, Vest JR. Hospitalization event notifications and reductions in readmissions of Medicare fee-for-service beneficiaries in the Bronx, New York. J Am Med Inform Assoc

2017;24(e1):e150–6.
6. Campion TR Jr, Ancker JS, Edwards AM, Patel VN, Kaushal R, HITEC Investigators. Push and pull: physician usage of and satisfaction with health information exchange. AMIA Annu Symp Proc 2012;2012:77–84.
7. Frisse ME, Tang L, Belsito A, Overhage JM. Development and use of a medication history service associated with a health information exchange: architecture and preliminary findings. AMIA Annu Symp Proc 2010;2010:242–5.
8. Berg M. Patient care information systems and health care work: a sociotechnical approach. Int J Med Inform 1999;55(2):87–101.
9. Spearman C. "General Intelligence," Objectively Determined and Measured. Am J Psychol 1904;15(2):201–92.

Delays in Care and Differential Presentation of Acute Myocardial Infarction

Authors: Harry Reyes Nieva, Tony Y. Sun, Sharon R. Gorman, Grace Mao, Noémie Elhadad, Columbia University

Abstract: The objective of this retrospective observational cohort study is to elucidate differences in presentation of acute myocardial infarction (AMI) across sex and race/ethnicity, and identify potential disparities in timeliness of treatment. Delays in treatment may contribute to prolonged atrophy of cardiac muscle and increased lifelong disease burden.

We examined emergency department (ED) visits at the Columbia University Irving Medical Center in New York between January 2010 and June 2019 for patients presenting with AMI. We extracted electronic health record (EHR) data and excluded patients with prior history of myocardial infarction. Using named-entity recognition, we identified sign and symptom mentions from ED admission notes. We assessed differences in presentation of signs/symptoms among patients using Chi-squared tests with false discovery rate correction for multiple comparisons. Using EHR timestamps, we fit multivariate Cox proportional-hazards models for in-hospital survival and time to first measurement, medication, and procedure.

We identified 2,299 patients with their first diagnosis of AMI during an ED visit (54% male; 30% Hispanic/Latinx, 16% Caucasian/White, 11% African American/Black). Compared to male patients, female patients were more likely to present with generalized pain (OR:2.83; 95%CI:1.29-6.21; p=0.02), head pain (OR:2.45; 95%CI:1.73-3.46; p<0.001), dyspepsia (OR:1.94; 95%CI:1.21-3.11; p=0.02), epigastric pain (OR:1.72; 95%CI:1.34-2.20; p<0.001), leg pain (OR:1.55; 95%CI:1.24-1.94; p<0.001), esophageal reflux (OR:1.46, 95%CI:1.13-1.90, p=0.01), musculoskeletal pain (OR:1.36, 95%CI:1.14-1.62, p=0.003), malaise/fatigue (OR:1.32; 95%CI:1.1.11-1.56; p=0.006), and neck pain (OR:1.32, 95%CI:1.07-1.63, p=0.02). Differences in in-hospital survival were not statistically significant for sex (p=0.84) or race/ethnicity (p=0.18). Compared to all other patients, however, Latinx and African American patients experienced longer time-to-first-measurement (aHR:0.88; 95%CI:0.78-0.99; p=0.04) and time-to-first-procedure (aHR:0.83; 95%CI:0.73-0.94; p=0.003).

Interim analyses support consideration of sex-based differences in clinical presentation when diagnosing AMI and suggest disparities in time to treatment based on race/ethnicity. Future work will further examine symptom clusters and their association with care delivery.

Statement of Significance: Prior studies on medical intervention following acute myocardial infarction (AMI) have focused on in-hospital mortality and long-term outcomes based on patient sex and race/ethnicity. Here, we propose to assess differences in time to events related to care during AMI-related ED visits. Furthermore, our analysis incorporates all documentation from ED visits, including the patient narrative at admission. Because prior research has shown the importance of early intervention at time of AMI, assessing these differences in care practices during ED visits has the potential to produce actionable insights for mitigating disparities in care.

Keywords: health disparities, quality of care, cardiovascular health
E-mail of First Presenter: hr2479@cumc.columbia.edu

Thinking Fast and Slow: Visualization of Financial Conflict of Interest Disclosure in Clinical Trial Publications

Authors: Alex Rich, University of North Carolina at Chapel Hill

Abstract: Much of the damage caused in the US opioid crisis can be traced to corporate influence on clinical research, medical education, and health policy. The 1990s and early 2000s saw a surge in industry-funded research publications suggesting that there was an epidemic of untreated pain in America and that prescription opioids offered a safe and cost-effective means of addressing that problem. The resulting explosion of opioid prescribing has led to hundreds of thousands of overdose deaths. Opioids are not the only drug category to experience this kind of manufacturer-friendly influence on clinical trials. Nonsteroidal anti-inflammatory drugs, selective serotonin reuptake inhibitors, and other drug categories have been the subject of unusually favorable drug trial publications whose results have later been disproven by significant patient harm after reaching the market.
This raises a key question: how do instances of bias in drug trial publications go undetected by the highly trained professionals who read them? Part of the answer lies in the way that human beings approach informational problems. Nobel laureate economist Daniel Kahneman frames this process as dividing between two different modes: System 1 thinking and System 2 thinking. System 1 is characterized by fast, automatic consumption of information and instinctive reaction. System 2 is characterized by slower, more contemplative and critical evaluation of the information at hand. Kahneman's most recent book sums this approach up in its title: "Thinking fast and slow." All readers, regardless of their level of sophistication, are capable of sliding into rapid, passive consumption of the information they are being presented. Nudging readers into slower, more critical "System 2" thinking is necessary to facilitate recognition of instances of bias in drug trial publications.

Statement of Significance: The core question of the research I am conducting is: can a visualization of authors' financial conflict of interest in a drug trial publication provide a cognitive speed bump to help highlight instances in which readers might benefit from slower, more critical evaluation of a drug trial with significant backing from manufacturers who have a large financial stake in the published results?

Keywords: Public Health, Opioid Crisis, Visualization, Data Evaluation
E-mail of First Presenter: alex.rich@unc.edu

Private Insurance Market Participation by Psychiatrists in Massachusetts

Authors: Nicole M Benson MD, Catherine Myong BA, Joseph P Newhouse PhD, Vicki Fung PhD, John Hsu MD MBA, Harvard University

Abstract: Countless patients rely on health insurance to help pay for care; however, many psychiatrists do not accept health insurance reimbursement which may limit access to specialty mental health care. Using 2013 Massachusetts licensing data and the All-Payer Claims Database, we conducted a population-based, cross-sectional analysis examining psychiatrists licensed in Massachusetts (MA) treating patients through insurance in 2013. Of the 2,348 licensed psychiatrists in Ma, 78.5% had at least one paid claim for an outpatient visit in the APCD but only 6.4% had claims for at least 300 patients/year (a full caseload). Psychiatrists had a median number of 18 patients with claims (mean = 73). Compared to those 30-39 years from medical school graduation, psychiatrists within 19 years of graduation were less likely to bill for an outpatient (Odds Ratio [OR] 0.67 for 7-19 years, 95% Confidence Interval [CI]: 0.47-0.94) and less likely to have claims for 300+ patients/year (OR 0.49 for 7-19 years, CI: 0.29-0.83). Participation varied across insurance types (93.3% for group commercial plans vs 32.7% for Medicaid Managed Care plans). Among Massachusetts psychiatrists, participation in the private insurance market appears to be limited. Currently, older psychiatrists are more likely to participate, suggesting that access to psychiatrists for patients seeking care using insurance could worsen as these psychiatrists retire.

Statement of Significance: Psychiatrists in MA appear to participate in the health insurance market to a limited degree, with many physicians seeing only a small number of patients through insurance, possibly because they are seeing patients outside of the insurance market. These challenges are likely to worsen substantially in the next decade and complicate policy efforts to improve access because half of licensed psychiatrists have thirty or more years of practice and may be approaching retirement.

Keywords: Psychiatry, Health Insurance Participation
E-mail of First Presenter: nbensonharva@mgh.harvard.edu

Impact of a Clinical Decision Support Tool in the Evaluation of Acute Pulmonary Embolism

Authors: Keaton Morgan, MD; Devin Horton, MD; Stacy Johnson, MD; Troy Madsen, MD; Joe Habboushe, MD; Yi Lu, MS; Kensaku Kawamoto, MD, PhD, The University of Utah

Abstract: Several validated clinical decision rules exist for the evaluation of patients with suspected acute pulmonary embolism. Recent studies, however, indicate that these decision rules are not appropriately applied as often as 7-13% of the time, leading to unnecessary advanced imaging studies. Clinical decision support (CDS) systems have been demonstrated to increase compliance with clinical practice guidelines and may be valuable tools in the evaluation of suspected pulmonary embolism. OBJECTIVE: We propose a CDS tool for the pulmonary embolism rule-out criteria (PERC) rule with the aim of decreasing the number of unnecessary advanced imaging studies performed in the evaluation of acute pulmonary embolism. METHODS: The study will be a before-and-after prospective study at two emergency departments (one academic and one community) in one healthcare system. The patient population will include all adult patients with a d-dimer or advanced imaging study ordered for suspected pulmonary embolism. INTERVENTION: The CDS tool will be an MDCalc application embedded

into a commercial electronic health record (EHR) system. All data that can be extracted from the EHR will automatically populate into the tool. The tool will fire as an interruptive pop-up if a d-dimer is ordered and the patient is PERC negative by criteria automatically extracted from the EHR. OUTCOMES: The primary outcome will be the number of computed tomography pulmonary angiography and ventilation-perfusion scans performed for evaluation of acute pulmonary embolism per total number of patient visits during the study period. The secondary outcome will be the missed diagnosis rate. HYPOTHESIS: We hypothesize that implementation of a PERC rule CDS tool for patients undergoing evaluation for suspected pulmonary embolism will result in a decrease in the number of advanced imaging studies performed in PERC negative patients without an increase in the missed diagnosis rate. Non-adherence to clinical practice guidelines in the evaluation of acute pulmonary embolism leads to overuse of advanced imaging studies, such as CT pulmonary angiography. Effective clinical decision support (CDS) may improve compliance with these guidelines and decrease the number of advanced imaging studies performed on low-risk patients. We plan to implement a CDS tool embedded into the electronic health record system for the use of the pulmonary embolism rule-out criteria (PERC) rule in the evaluation of acute pulmonary embolism. We estimate that successful implementation may result in a reduction of more than 100 advanced imaging studies per year at a single health care center. This study would represent the first successful implementation of a computerized CDS alert for the PERC rule and could provide important guidance for other institutions interested in implementing a similar CDS tool.

Keywords: Clinical decision support (CDS), Pulmonary embolism rule-out criteria (PERC), Acute pulmonary embolism, Emergency department
E-mail of First Presenter: keaton.morgan@utah.edu

Health Information Exchange in Primary Care: Use Among Teams

Authors: Nate C Apathy, BS; Joshua R. Vest, PhD, MPH; Julia Adler-Milstein, PhD; Brian Dixon, PhD; MPA; Justin Blackburn, PhD; Christopher A Harle, PhD, MS

Abstract: Health care delivery reform efforts in the United States (US) have aimed at improving care coordination and reducing the negative impact of care fragmentation on care quality. To facilitate these quality improvements, efforts such as the Patient-Centered Medical Home (PCMH) model emphasize team-based care that relies upon robust health information technology, including health information exchange (HIE). To understand the extent to which HIE in PCMHs is used in a way consistent with team-based care models, we linked HIE log data and clinical EHR data to quantify the proportion of HIE use in three PCMHs that was undertaken by a delegate user or multiple users (team-based). Second, we analyzed the relationship between team-based HIE use and 1) diversity of information accessed and 2) a measure of HIE use intensity. We stratified our analysis by timing of use, separating use in the two weeks prior to a visit from same day use and use in the two weeks following the visit. Of the 12,556 visits with HIE use, 10,702 (85.2%) met our criteria for team-based HIE use. Rates of team-based HIE use varied by use timing ($p<0.001$); less occurred on the day of the visit. In regression analyses, team-based HIE use was not associated with use diversity during any time period. However, team-based HIE use was negatively related to use intensity in the two weeks prior to the visit as well as the two weeks following the visit.

Statement of Significance: Delivery reform efforts rely on robust health information technology (IT) infrastructure to be successful, however little is known about the nature of health IT use and the degree to which current use supports innovative care delivery models like the PCMH. Without effective use of health IT that supports delivery redesign efforts such as those that emphasize team-based care, the health care system is unlikely to realize the full benefits of health IT. This study explores the extent to which health information exchange (HIE) use in PCMHs reflects team-based use, and analyzes whether or not team-based HIE use has informational benefit over non-team-based HIE use in terms of viewing additional information (diversity) or viewing more detailed information (intensity). Findings from this study can inform the implementation and design of health IT tools like HIE to support innovative care delivery models.

Keywords: Health information exchange; primary care; patient centered medical home
E-mail of First Presenter: natea@iu.edu

Yes, We Closed the Tap! Now What?

Authors: Meenakshi Mishra, Nicole Weiskopf, Oregon Health & Science University

Abstract: In response to the current opioid epidemic, national and state-level policies have been enacted to limit opioid prescriptions. The resulting decrease in opioid prescriptions is concurrent with a decline in deaths from prescription opioid overdose, but the impacts of prescription opioid policies on patient-specific pain and function outcomes are not well understood. Moreover, many patients are being provided medication assisted therapy and adjuvant psychosocial therapies to treat opioid use disorder. There is a need for real-world evidence regarding the effectiveness of these therapeutic strategies in this tremendously complex space.
Many health organizations have developed electronic health record (EHR) registries to track patients on chronic opioid therapy. There is enormous potential for combining these structured registry data with relevant patient-reported outcomes regarding pain, function, and quality of life, which are stored in a combination of structured and narrative fields within the EHR. There is an opportunity to use these data sources to evaluate the effectiveness of these pain management and opioid use disorder therapies, as well as to understand and ideally predict which patients may benefit from which therapeutic.

Statement of Significance: Very little is known about the effectiveness of chronic pain management strategies in the face numerous national level policies to reduce opioid prescription for chronic pain. This research will contribute to the better understanding of alternate strategies for managing chronic pain.

Keywords: opioid, machine-learning, chronic pain management
E-mail o First Presenter: mishram@ohsu.edu

A Proposal for the Relationship Between Physical Therapy and Opioid Use for US Veterans

Authors: Lindsey Brown-Taylor, Shardool Patel, Michael J Buys, MD, Benjamin S Brooke, George E. Whalen Veterans Affairs Medical Center Salt Lake City Salt Lake City Utah

Abstract: Background: Chronic opioid prescription is one catalyst for opioid addiction that may lead to abuse and overdose. Prescriptions for chronic opioid use are particularly prevalent and problematic among orthopaedic surgical patients for postsurgical pain. Non-pharmacologic alternatives are an urgent priority of the Veterans Health Administration (VHA) to combat the opioid epidemic. Physical therapy was proposed as a non-pharmacologic alternative to opioids by the American Physical Therapy Association, the Centers for Disease Control and Prevention, and the United States Surgeon General. Despite these recommendations, the relationship between physical therapy and opioid use is not well understood. The purpose of this project is to investigate the relationship between perioperative physical therapy and chronic postoperative opioid use after orthopaedic surgery. We hypothesize that increasing physical therapy utilization will be related to reduced risk of chronic postoperative opioid use.
Methods: The Corporate Data Warehouse of the VHA will be utilized to query a sample of patients who underwent elective orthopaedic surgery between Jan 1 2005 and Dec 31 2018. Physical therapy clinical encounters recorded in the electronic health record will be utilized to quantify the number of pre- and post-operative visits within 90 days of surgery. Chronic opioid use will be defined as receiving ≥1 opioid prescription 91-365 days postoperatively. Binomial multiple regression will be used to identify significant predictors for chronic postoperative opioid use. Independent variables will include number of pre- and post-operative physical therapy visits, the interaction of pre- and post-operative physical therapy visits, demographics, previous opioid use, co-morbidities, multimodal pharmacotherapy, intraoperative pain management strategy, and surgical procedure.
Anticipated Impact: Understanding the relationship between physical therapy and opioid use provides the opportunity to (1) identify patients who may benefit most from perioperative physical therapy, (2) inform perioperative physical therapy prescription frequencies, and (3) educate patients of non-pharmacologic alternatives for postoperative pain management.
Statement of Significance: The field of informatics has extensive application to medical, nursing, and pharmacological practice; however, informatics has not been heavily integrated to rehabilitation sciences. This abstract will detail a proposal to apply informatics to understand the relationship between physical therapy and opioid use after elective orthopaedic surgery. Healthcare utilization metrics related to physical therapy attendance before and after surgery will be evaluated in relation to postoperative opioid use to identify if, when, and for whom physical therapy may be most beneficial to reduce need for opioid medication.

Keywords: Rehabilitation, Opioids, Orthopaedics
Email of First Presenter: Lindsey.brown.taylor@utah.edu

Diagnostic Test Follow-up Support: Incomplete Test Notifications, Purge of Non-essential Notifications, and Flexible Secondary Pool Assignment

Authors: Peter Hong, MD; Jowell Sabino, RN, MSN, CPNP; Chase Parsons, DO, MBI; Jonathan Hron, MD; Marvin B. Harper, MD — Boston Children's Hospital, Boston, Massachusetts

Abstract: Timely and effective follow-up of laboratory tests and medical imaging results is crucial to patient care. We describe our multidisciplinary approach toward delivering health information technology solutions which integrate with the electronic health record system to expand functionality to include (1) notifications for incomplete test results, (2) purging of clinically-irrelevant results, (3) adaptive assignment of results to a secondary "pool" of clinicians for additional oversight of test results,

and (4) e-mail reminders of unendorsed notifications. Better quantification of the various factors of this process, in conjunction with quality improvement methodology, will help to demonstrate & increase the value of these systems.

Statement of Significance: Widespread adoption of government-certified electronic health record (EHR) systems has brought great promise to the health information technology (HIT) landscape; nonetheless, there remains significant variability with regard to follow-up practices of results, as well as notable challenge in locally tailoring solutions.(1-3) The capacity to address certain aspects of test follow-up, such as notification for incomplete studies or e-mail reminders, may not be comprehensively addressed by all EHR implementations.(4,5) Out-of-the-box EHR functionality, such as individualization of test-result routing at the provider level provides incredible customizability—but is often difficult to maintain, oversee, and audit. Clinical informaticists are positioned to thoughtfully interface with clinicians and IT professionals in the design, maintenance, and review of diagnostic test follow-up procedures to enhance patient safety and reduce cognitive burden on clinical teams.
Our procedures offer a unique set of approaches to the dilemma of diagnostic test follow-up across a heterogenous clinical departments and sites. Challenges with corner cases of complex routing logic and with process transparency & feedback present opportunities for next steps. Measuring the impact of these solutions upon patient safety and provider satisfaction will help to guide future direction.

Keywords: Test follow-up, EHR customization, High reliability
Email of First Presenter: peter.hong@childrens.harvard.edu

References:

1. Georgiou A, Li J, Thomas J, Dahm MR, Westbrook JI. The impact of health information technology on the management and follow-up of test results – a systematic review. Journal of the American Medical Informatics Association. 2019 Jul 1;26(7):678–88.
2. Callen J, Georgiou A, Li J, Westbrook JI. The impact for patient outcomes of failure to follow up on test results. How can we do better? The Journal of the International Federation of Clinical Chemistry and Laboratory Medicine. 2015 Jan;26(1):38–46.
3. Colicchio TK, Cimino JJ, Del Fiol G. Unintended Consequences of Nationwide Electronic Health Record Adoption: Challenges and Opportunities in the Post-Meaningful Use Era. J Med Internet Res. 2019 Jun 3;21(6):e13313.
4. Singh H, Thomas EJ, Mani S, Sittig D, Arora H, Espadas D, et al. Timely Follow-up of Abnormal Diagnostic Imaging Test Results in an Outpatient Setting: Are Electronic Medical Records Achieving Their Potential? Arch Intern Med. 2009 Sep 28;169(17).
5. Dalal AK, Poon EG, Karson AS, Gandhi TK, Roy CL. Lessons learned from implementation of a computerized application for pending tests at hospital discharge. J Hosp Med. 2011 Jan;6(1):16–21.

Open-Mic Session 3

Machine Learning

Stabilized Image Segmentation in the Presence of Noise

Authors: Jonas A. Actor, Rice University; Beatrice Rivière, Rice University; David T Fuentes, University of Texas MD Anderson Cancer Center

Abstract: Medical image segmentation is increasingly performed using deep convolutional neural networks. While these deep learning methods produce state-of-the-art results, such methods are unstable: small noise in input can cause drastically different segmentation results. As a result of this instability, deep learning segmentation methods are viewed as "brittle black boxes", raising questions of reliability and clinical liability. We address these concerns by improving the stability of deep learning methods. To do so, we propose a novel method that incorporates into the training process a notion of how much each convolution layer can be distorted due to noise. In our approach, we use Holder's Inequality to bound the spectral norm of the convolution linear operators, as this norm measures the possible distortion at each layer. Our bound applies to a broader class of convolutions than previous methods, is faster to compute, and requires no additional implementation beyond the tools available in current deep learning toolkits. We incorporate this bound as a regularization term in training as part of this method. We evaluate the effectiveness of our method on models for liver segmentation from CT image data, and we demonstrate that our method helps to stabilize segmentation output in the presence of noise.

Statement of Significance: Over the last decade, deep learning for medical imaging has rapidly become a valuable tool to gather quantitative data in devising better treatments and in improving clinical care. Despite many high profile successes, there are concerns about the viability of deep learning methods,

since many promising results on specific datasets have not been replicable, due to subtle differences in image properties such as noise or image acquisition parameters. This work addresses these concerns by proposing a novel technique to improve the stability of deep learning segmentation methods. Our mathematical analysis applies to a broader class of convolutions than previous methods, and the resulting stability bound is both faster to compute and requires no additional implementation beyond the tools available in current deep learning toolkits.

Acknowledgement:
This work is supported by the NLM Training Program in Biomedical Informatics and Data Science T15LM007093, Program Director Dr. Lydia Kavraki.

Keywords: Medical imaging, image segmentation, hepatocellular carcinoma, deep learning
E-mail of First Presenter: jonasactor@rice.edu

Decomposition of Clinical Disparities with Machine Learning
Author: Noah Hammarlund, University of Washington

Abstract: Differences in average treatment rates for conditions such as emergency cardiac surgery point to racial disparities in healthcare. The optimal approach to alleviate a given disparity depends on whether the main driver is differential health risk or differential treatment within the healthcare system. I propose an extension to the Oaxaca Blinder decomposition framework to capitalize on advances in clinical data and machine learning prediction to quantify the portions of a given disparity due to differential clinical health and differential healthcare treatment. The proposed method applied to the surgery decision for heart attacks using electronic medical records data from a major academic hospital system in Indiana suggests a smaller potential healthcare treatment disparity compared to the conclusion from the standard approach. However, the method reveals that the entirety of the cardiac surgery rate difference can be explained by differential healthcare treatment for Black patients even after machine learningbased clinical health adjustment.
The machine learning decomposition approach can highlight potential increased average health risks or decision problems to build upon standard inequality decompositions and fairness-aware machine learning approaches. A large healthcare treatment portion indicates a differential relationship between clinical variables and treatment for Black patients. Differential clinical treatment in cardiac care suggests the importance of system- and physician-level policies to address differential treatment. The proposed decomposition approach can scale to apply to many clinical decisions to better differentiate health and healthcare disparities to better target policy solutions.

Keywords: Machine Learning, Health Disparities, Decision Support, Electronic Medical Records
E-mail of First Presenter: noahh@uw.edu

Electronic Health Record Phenotyping of Patients with Drug-Induced Renal Injury

Authors: Zaid K. Yousif, PharmD,[1] Linda Awdishu, PharmD MAS,[1] Michael Hogarth, MD,[1] Jejo Koola, MD MS[1]. [1]University of California San Diego, La Jolla, California

Abstract: Adverse Drug Reactions (ADRs) occur in up to 20% of hospitalized patients, result in 100,000 deaths annually and are major contributors to cost, morbidity, and mortality in the United States. Up to 20% of hospitalized patients will have at least one ADR. DIRI (Drug-Induced Renal Injury) is a significant ADR in hospitalized patients. DIRI is characterized by a significant acute reduction in renal function. Prospective cohort studies of Acute Kidney Injury (AKI) have documented the frequency of DIRI to be up to 26%. Detection of patients with DIRI is a practical challenge due to the ubiquity and complexity of AKI risk factors at inpatient care. In addition, confirming DIRI diagnosis requires advanced clinical training. Our goal is to develop an algorithm to identify DIRI cases retrospectively using machine learning techniques.

This is a retrospective analysis of the Rationale and Design of the Genetic Contribution to Drug-Induced Renal Injury Study (DIRECT). A total of 211 clinical predictors were evaluated including demographics, laboratory values, vital signs, medications, diagnoses, and procedures. Clinical predictors were evaluated at different time points during hospital stay. Statistical analysis was performed using logistic regression with L_1 penalty (LASSO regularization) and ten-folds cross-validation to construct a probabilistic model with DIRI vs. Not-DIRI as the outcome.

A total of 314 subjects met the inclusion and exclusion criteria and completed the study. Significant predictors in the multivariable analysis were days of exposure to nephrotoxic drugs prior to AKI, receiving prescription medications during hospital stay, past medical history of cardiac surgery prior to hospital admission, heavy proteinuria during hospital stay, vancomycin trough ≥ 20 mcg/dL during hospital stay. The model achieved a performance of 0.79 ± 0.03 assessed using Area Under the Receiver Operator Characteristic Curve (AUC).

Statement of Significance: This is one of the first endeavors to create a probabilistic model to identify cases of DIRI. Previous research has been mainly focused on predicting AKI incidence regardless of the etiology of renal injury. This work can be utilized to retrospectively identify DIRI patients for cohort and genotyping studies. Additionally, this work can be utilized to guide medical intervention when treating patients with AKI. Future work includes validating the model performance using an external dataset.

Keywords: Adverse Drug Reactions, Acute Kidney Injury, Machine Learning
Email of First Presenter: zyousif@health.ucsd.edu

Developing and Testing Clinical Decision Support for Neonatal Ventilator Management

Authors: Lindsey A Knake, Wael Alrifai, Christoph U Lehmann, L. Dupree Hatch, Vanderbilt University

Abstract: Mechanical ventilation (MV) in the Neonatal Intensive Care Unit (NICU) is a life-saving therapy, but is associated with neurodevelopmental impairment, increased mortality, and long-term respiratory complications. In neonates, volume-targeted ventilation (VTV) modes are associated with shorter MV courses and improved clinical outcomes. However, only 42% of NICUs in the US and Canada report use of VTV as the primary mode. The slow clinical adoption is due, in part, to a lack of providers' knowledge of appropriate tidal volumes (TV) for the patient's weight and disease state, which may lead to apparent "failure" of VTV. Clinical decision support tools (CDS) could facilitate appropriate use of VTV

in clinical care. We hypothesize that implementation of CDS tools for choosing an evidence-based initial TV will improve the use of VTV in the NICU.
Using data from the electronic health record (EHR) and a novel ventilator data capture program, we will conduct a multi-center retrospective cohort study to determine the prevalence of evidence-based ventilator use, including choice of mode and TV, and to identify demographic and practice factors related to their use. Subsequently, we will conduct an online questionnaire in three Vanderbilt Health Affiliated Network NICUs to develop local consensus recommendations for TV. Finally, using the data gained during this process, we will create CDS tools to recommend evidence-based TV values. Using an interrupted time series design, we will deploy a CDS tool in multiple locations within the clinical workflow in two study NICUs with the third serving as a control. Our primary outcome will be the percentage of mechanically ventilated infants who receive an initial TV that is congruent with evidenced-based recommendations. This study design will allow us to evaluate where in the clinical workflow CDS is most effective.

Statement of Significance: Our study design will determine if clinical decision support for neonatal mechanical ventilation is beneficial and where in the clinical workflow the tool is most effective. Our proposed work will lay the groundwork for future studies in clinical informatics, neonatal mechanical ventilation, and implementation science.

Keywords: Clinical Decision Support, Ventilator, Pediatrics.
E-mail of First Presenter: lindsey.knake@vumc.org

Acceleration Signals in Determining Gait-Related Difficulties and the Motor Skill of Walking in Older Adults

Authors: Pritika Dasgupta[1], Ervin Sejdić[4]
Associated Authors/Collaborators: Jessie VanSwearingen[2], Alan Godfrey[3], Mark Redfern[4], Manuel Montero-Odasso[5], Brian Suffoletto[7], Adam Frisch[7], James Huber[7], James Alexander Hughes[8], Mark Daley[9]

Affiliations:
[1]Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15261, USA. [2]Department of Physical Therapy, School of Health and Rehabilitation Sciences, University of Pittsburgh, Pittsburgh, PA, 15261, USA. [3]Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, Tyne and Wear, England, UK. [4]Department of Electrical and Computer Engineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, PA, 15261, USA. [5]Gait and Brain Lab, Parkwood Institute and University of Western Ontario, Canada. [7]Department of Emergency Medicine, University of Pittsburgh School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15261, USA. [8]Department of Computer Science, St. Francis Xavier University, Antigonish, Nova Scotia, B2G 2W5, Canada [9]Department of Computer Science, University of Western Ontario, London, Ontario, N6A 3K7, Canada

Abstract: In adults 65 years or older, falls or other neuromotor dysfunctions are often framed as walking-related declines in motor skill; the frequent occurrence of such decline in walking-related motor

skill motivates the need for an improved understanding of motor skill in walking. Simple gait measurements, such as speed, do not provide adequate information about the quality of the translation of the body motion during walking. Furthermore, there is a great need in the clinical literature and clinical practice for more accurate measures of the loss of the motor skill of walking, so that clinical practice can provide better therapeutic interventions to improve the motor skill of walking.
This open mic suggests a consensus on what the motor skill of walking is and dissects it into seven interrelated characteristics and traits. Subsequently, we purport that these characteristics of the motor skill of walking cannot be represented by simple gait measurements or raw sensor measurements alone. Gait measures from accelerometers placed on the lower trunk, or trunk-acceleration gait measures, can enrich measurements of walking and motor performance.
To support our claim, we will map these acceleration gait measures (AGMs) to the various aspects of the motor skill of walking. Additionally, influential AGMs will be elected through feature selection methods. Various machine learning algorithms ranging from logistic regression, non-linear regression, evolutionary algorithms, and ensemble methods will be used to make predictions on age-related gait-related difficulty outcomes (such as fall risk). Various examples of AGM mapping to the motor skill of walking will be demonstrated through three datasets: a dataset of 10 healthy volunteers in a controlled lab setting, data from the PRIMA (Program to Improve Mobility in Aging) dataset of 248 community-dwelling older adults of 65-91 years of age, and data from older adults taken during emergency room visits at University of Pittsburgh.

Statement of Significance: Overall, we hope to find that the combination of high-fidelity artificial intelligence algorithms and acceleration gait measures derived from low-cost sensors can fulfill the severe and crucial need for the clinical measurement of the motor skill of walking in older adults.

Keywords: Gait, Signal Processing, Older Adult Research, Machine Learning
E-mail of First Presenter: prd17@pitt.edu

Predicting Total Daily Dose of Insulin for Optimal Blood Glucose Control in Hospitalized Patients
Authors: Minh Nguyen; Larry Kalesinskas; Ivana Jankovic, MD; Jonathan Chen, MD PhD, Stanford University

Abstract: Diabetes mellitus is a chronic disease characterized by hyperglycemia. Uncontrolled hyperglycemia in hospitalized patients can lead to serious complications and increased mortality. Inpatient treatment for hyperglycemia is insulin, but therapeutic range is narrow. Predicting total daily insulin dose (TDD) is a first step in prescribing insulin to improve patient health outcomes. It is often undertaken by clinicians using simple weight-based formulas and clinical judgement. We aim to determine if machine learning prediction models of TDD in hospitalized patients can be useful in practice and develop good models for such tasks in the future.
We used the inpatient electronic health record (EHR), available from the Stanford Research Repository (STARR) database to identify a "non-complex" cohort from patients who achieved good glucose control. We first included the most relevant features that an endocrinologist considers in prescribing insulin. These are hemoglobin A1c, height, weight, creatinine level, sex, age, steroid dose, diet, whether or not patients received basal insulin, and counts of all blood glucose measurements within 48 hours in specified value ranges before prediction time. We performed a baseline univariate regression with

weight as the input. We also performed a multivariate regression, a random forest, and a fully connected neural network model using the all the above features.
Initial results showed that using weight only had the worst performance, with a mean absolute error (MAE) of 7 units. We would argue that the current practice of weight-based dosing insulin can be improved. The multivariate regression, random forest and neural network models showed significant improvement, yielding a MAE of 5.4, 4.9 and 4.2 units respectively. Important features for prediction were basal insulin, weight, age, creatinine, and height.
We are working on expanding the feature space, getting better data, developing better deep learning models for prediction, understanding variable importance, and the interpretability of our models.

Statement of Significance: The current practice for determining total daily insulin dose is based on a combination of weight-based calculations and clinical judgement, which can result in significant practice heterogeneity and sub-optimal blood glucose control. Most work has done in predicting blood glucose within an hour based on previous glucose measurements from continuous glucose monitoring devices. To the best of our knowledge, there are no publications on predicting insulin daily dose for hospitalized patients, using glucose measurements from glucometers.

Keywords: clinical informatics, healthcare, machine learning
E-mail of First Presenter: minh084@stanford.edu

Predicting Suicidal Ideation and Attempt in Children Aged 9-10

Authors:
Gareth Harman[1,2], Dakota Kliamovich[3], Michael Mooney[2], Bonnie J. Nagel[1,3]
1) Department of Psychiatry, Oregon Health & Science University, Portland, OR, USA
2) Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA
3) Department of Behavioral Neuroscience, Oregon Health & Science University, Portland, OR, USA.

Abstract: Suicide remains the second leading cause of death for individuals aged 10-24, yet risk factors for suicidal ideation and attempt are poorly characterized in those under the age of 10. The ability to identify individuals that may have an increased risk for either suicidal ideation or attempt is a critical precursor for early intervention.
The Adolescent Brain Cognitive Development (ABCD) study is a landmark project that aims to characterize normal brain development from childhood through adolescence and into young adulthood. The ABCD study has enrolled 11874 healthy children between the ages of 9-10 from 21 sites across the US representing the national diversity of ethnic, racial, and socioeconomic backgrounds.
This project examines risk factors implicated in existing suicide literature by selecting these elements as features for model training to predict suicidal ideation and attempt in the ABCD sample. These features span several domains, including patient demographics, measures of impulsivity, neurocognition, mental and physical health, and cultural and environmental factors.
We employed random forest, logistic regression and penalized regression models with and without feature selection to classify both suicidal ideation and attempt. Logistic regression with feature selection and elastic-net without feature selection were the best performing models able to classify individuals with suicidal ideation from controls with an area under the curve (AUC) of .70 (CI 95% .70-71). The

important features in this classification task included measures of loneliness, impulsivity, feeling unloved, and aggression. In addition to predicting suicidal ideation, we were able to use this same model trained on those with suicidal ideation to classify those reporting suicide attempt from controls (AUC of .77; CI 95% .76-.77). To our knowledge, this study is the first of its kind to utilize predictive modeling for suicidal ideation and attempt in a large and diverse sample of children aged 9 and 10.

Statement of Significance: It is known that suicidal ideation and attempt are major concerns in the adolescent population, however, little is known about this phenomenon in children aged 9 and 10. There is a large body of literature that has identified several domains of risk factors that may contribute to an individual's risk of suicidal ideation or attempt. This study examines the use of machine learning using risk factors implicated in adolescence to identify those in a younger population that may be at greatest risk for suicidal ideation or attempt. An important element of this study is the use of "transdiagnostic" to further elucidate the contributing features by examining the more granular feature i.e. "difficulty concentrating", rather than the diagnosis i.e. attention deficit disorder. This provides both a more detailed and generalizable risk feature space. The ability to detect those at greatest risk provides a critical opportunity for early intervention.

Keywords: Suicide, Mental Health, Children, Machine Learning
E-mail of First Presenter: harmang@ohsu.edu

Clinical Decision Support for Predicting Postpartum Depression Using Machine Learning

Authors: Houda Benlhabib, Ian Bennett, Peter Tarczy-Hornoch, Sean Mooney, University of Washington

Abstract: Postpartum depression (PPD) is a depression that occurs after childbirth. Importantly, if left untreated PPD can have a severe outcome on the mother and offspring. Current statistics show that most of the maternal deaths in the US were in postpartum period. Toxicology testing revealed that women in postpartum period are at risk of suicide, accidental drug overdose and homicide. PPD can also lead to infanticide, decreased maternal sensitivity and attachment to infants which leads to poor child development. Importantly, the clinical diagnosis of PPD remains challenging due in part to high percentage of women with the disorder that fail to report it and do not seek the appropriate interventions. One of the main contributors of the latter is the lack of a routine protocol to screen women for depression during and after pregnancy. This indicates the need of better tools to screen women for PPD. The objective of the current work is to develop computational tools using machine learning and natural language processing to predict women who develop PPD using Electronic Medical Health Records (EHR). Here we report the use of deidentified aggregated reporting query tool, Leaf, to identify a population of pregnant women who delivered at the UW Medicine that suffer from PPD and to assess and characterize attributes associated with it using EHR. These attributes can serve develop data science approaches for decision support. In this presentation, we will describe the study population and our approach toward developing new methodology using structured EHR data and clinical text.

Keywords: Bioinformatics
E-mail of First Presenter: hbenlh@uw.edu

Developing a Predictive Model for Endometriosis Diagnosis

Authors: Amber Kiser, Karen Schliep, Karen Eilbeck, The University of Utah

Abstract: Endometriosis is a common yet debilitating disease that is estimated to affect 11% of reproductive-age women. A significant delay of 6 – 12 years exists between the start of symptoms and diagnosis. This delay contributes to a significantly lessened quality of life for these women. The reason for this delay is complex; however, it is partly due to the difficulty of diagnosis. The current gold standard requires laparoscopic surgery that carries with it inherent risks and costs. The search for a non-invasive diagnostic method has been ongoing for decades, which includes looking for biomarkers as well as developing predictive models. Our goal is to join in this endeavor and build a predictive model, employing novel methods of machine learning that have not been used before. The data comes from a cohort of 473 women who underwent laparoscopic surgery for a number of indications, including tubal ligation, fibroids, pelvic pain, and infertility. In this cohort, 173 women were diagnosed with endometriosis. A number of descriptive methods, including principle component analysis (PCA) and linear discriminant analysis (LDA) were used to see any patterns in the data. While PCA was not effective at uncovering patterns, LDA proved to be better suited to the data and was able to discriminate between patients with and without endometriosis. Next, we used several different machine learning methods, including logistic regression, Naïve Bayes, decision tree, random forest, and support vector machine, to develop a classifier that would distinguish patients with endometriosis from patients without endometriosis. We will present our results from these different classifiers. The ultimate goal of this project is to develop a model that can serve as a diagnostic tool and decrease the delay of diagnosis for patients with endometriosis.
Endometriosis is a common yet debilitating disease that is estimated to affect 11% of reproductive-age women. A significant delay of 6 – 12 years exists between the start of symptoms and diagnosis. This delay contributes to a significantly lessened quality of life for these women. The reason for this delay is complex; however, it is partly due to the difficulty of diagnosis. The current gold standard requires laparoscopic surgery that carries with it inherent risks and costs. The goal of this project is to develop a model that can serve as a diagnostic tool and decrease the delay of diagnosis for patients with endometriosis.

Keywords: machine learning, endometriosis, clinical informatics
E-mail of First Presenter: amber.kiser@utah.edu

A Bayesian Hidden Markov Model for Assessing Seizure Risk

Authors: Emily T Wang[1], Sharon Chiang[2], Zulfi Haneef[3], Robert Moss[4], Marina Vannucci[1]
[1]Department of Statistics, Rice University, Houston, Texas, U.S.A.
[2]Department of Neurology, University of California, San Francisco, San Francisco, California, U.S.A.
[3]Department of Neurology, Baylor College of Medicine, Houston, Texas, U.S.A
[4]Seizure Tracker LLC, Springfield, Virginia, U.S.A.

Abstract: Epilepsy, a chronic neurological disorder, is characterized by frequent, unpredictable seizures arising from abnormal electrical disturbances in the brain. The unpredictable nature of seizures not only

makes treatment of epilepsy difficult but can also have a high impact on patients' quality of life. Existing methods for assessing seizure risk heavily rely on historical raw seizure counts, which by themselves do not accurately capture the underlying fluctuations and predict current seizure risk for the patient. As such, clinical management of epilepsy is impaired by difficulties in estimating changes in a patient's propensity to seizures over time.
To address these issues, we propose an unsupervised Bayesian nonhomogeneous hidden Markov model for discretizing and assessing daily seizure risk in patients with epilepsy using raw seizure count data as well as external clinical covariates such as age, gender, medication adherence and seizure triggers. To handle the high proportion of zeros and complex temporal structure characteristic of daily seizure count data, we model the seizure counts using a statistical distribution with a flexible dispersion parameter and a zero-inflation parameter. Accuracy of the model was evaluated through simulation studies.
We apply the model to daily seizure count data from SeizureTracker.com, a patient-reported electronic seizure diary containing over 1.7 million recorded seizure events since 2007. Results show improvements in accuracy over existing seizure risk assessment methods as well as traditional methods that rely solely on raw seizure counts. Clinical application of this method will directly improve patient care through a better understanding of patients' epilepsy disease trajectory through model results, allowing clinicians to act proactively with antiepileptic drug (AED) interventions.

Statement of Significance: Epilepsy affects approximately 1% of the population, with about one third of affected patients refractory to antiseizure medication. Clinical management of epilepsy is further limited by the unpredictable nature of seizures. Existing methods for seizure risk assessment depend heavily on raw seizure counts, which alone cannot adequately distinguish true changes in seizure risk from natural variations in the number of seizures. Our approach for assessing seizure risk utilizes external clinical covariates as well as raw seizure counts to estimate changes in a patient's sequence of seizure risk states over time. This quantitative approach allows us to directly measure seizure propensity rather than relying on the proxy measure of raw seizure counts. Results on simulated daily seizure counts indicate that our model achieves higher accuracy than existing seizure risk assessment tools. The model may be used by clinicians to understand changes in patients' epilepsy disease progression and to plan treatment interventions.

Acknowledgement:
This work is supported by the NLM Training Program in Biomedical Informatics and Data Science T15LM007093, Program Director Dr. Lydia Kavraki.

Keywords: Seizure risk, epilepsy, hidden Markov model, Bayesian inference
Email of First Presenter: emilywang@rice.edu

How to Capture, Characterize, and Classify What We Don't Know
Authors: Mayla R Boguslav, Lawrence E Hunter, Sonia M Leach

Abstract: There is an extensive natural language processing literature focused on extracting information, i.e. characterizing what we know. We propose to flip this emphasis and instead focus on extracting what we don't know--known unknowns--in order to specifically characterize the current state of scientific inquiry. These known unknowns are stated in the scientific literature in the form of

hypotheses, claims, future opportunities, anomalies, and evidence statements that identify goals for knowledge and future research. We present an efficient annotation schema that can formally represent such statements and be used to train automatic classifiers. The schema includes a taxonomy of types of statements about unknowns, a list of over 830 lexical cues (words that can indicate such statements) with many examples, and a preprocessing step to ease the work of annotators and accelerate the annotation process. Lexical cues include "not known," "calls into question," "complicated," "remarkably," "possible," "might," etc. We report on our progress including the strengths, weaknesses, and difficulties of annotating what we don't know, as well as classification results using the data collected.

Statement of Significance: Characterizing statements of known unknowns in the scientific literature is a step towards scientific transparency – disseminating the state of scientific ignorance as well as scientific knowledge to the public, allowing an exploration of the dynamic landscape of research.

Keywords: NLP, Ontology, Unknowns
E-mail of First Presenter: Mayla.Boguslav@CUAnschutz.edu

An interpretable Electronic Health Record (EHR) Phenotype for Treatment Resistance in Depression (TRD)

Authors: James T Brown, Michael A Ripperger, Colin G Walsh, Vanderbilt University

Abstract: Treatment-resistant depression (TRD) patients suffer increased healthcare costs, depressive symptoms, and notably increased risk for suicide attempt (30% lifetime risk). Pharmacogenomics-informed interventions might improve treatment precision, but insufficient sample size has prevented genome-wide association studies from identifying specific TRD variants. We aimed to increase sample size in future studies by developing a quantitative TRD phenotype model from routinely gathered EHR data. The data included demographic, diagnostic, pharmaceutical, and comorbidity predictors of more than 146,000 depression and dysthymia patients that visited Vanderbilt University Medical Center in the past 20 years. Using electroconvulsive therapy as a surrogate outcome for TRD, we developed multiple models through a combination of generalized linear regression and logistic regression with L1-norm regularization. These methods emphasize model interpretability to encourage trust and uptake in potential clinical use. Ultimately, we show a scalable, interpretable quantitative model can predict a patient's likelihood of having TRD, a complex disease whose clinical definition lacks general consensus, with relative calibration and discrimination. Predicting each patient's likelihood of treatment resistance will allow researchers to include all patients in genome wide association studies, circumventing the limitations of binary TRD definitions. These findings additionally lay the groundwork for potentially useful clinical prediction.

Statement of Significance: The present study presents the development of an interpretable quantitative phenotypic model capable of predicting a depression patient's likelihood of being treatment resistant. This model has potential to resolve sample size issues in genome wide association studies targeting genetic variants corresponding with treatment resistance in depression (TRD). The model also lays the groundwork for a clinically useful TRD prediction model, reducing morbidity and suicide risk for TRD patients.

Keywords: TRD, Predictive Modeling, Machine Learning, Psychiatry, Precision Medicine
E-mail of First Presenter: james.t.brown@vanderbilt.edu