National Library of Medicine Informatics Training Conference

June 27-28, 2016

The Ohio State University
The Ohio Union
Columbus, Ohio

TABLE OF CONTENTS

Posters (Days 1 and 2)

NLM Informatics Training Conference 2016
The Ohio State University

Agenda
Monday, June 27, 2016

6:30 – 8:00 AM	**Transportation Time | Conference Hotel → Ohio Union** There will be a complementary shuttle from the Columbus Hilton Downtown to The Ohio Union.
7:00 – 7:55 AM	**Registration and Breakfast** *Location: Outside of the Great Hall Meeting Room* **[Posters to also be set-up during this time]**
8:00 – 11:30 AM	**US Bank Theater, Ohio Union**
8:00 – 8:10 AM	**Welcome to Ohio State** | Dr. Bruce McPheron Provost and Executive Vice President, The Ohio State University
8:10 – 8:20 AM	**Opening Remarks from Hosting Training Site** | Dr. Philip R.O. Payne Chair, Department of Biomedical Informatics, The Ohio State University
8:20 – 8:30 AM	**Introduction to Training Directors and Trainees** | Dr. Valerie Florance Director, NLM Extramural Programs
8:30 – 9:45 AM	**Plenary Session #1** | Moderator: Dr. Alexa McCray, Harvard University (1 hour 15 min, 5 papers) (12 minutes per presentation, 3 minutes for Q&A) *Location: US Bank Theater* 1. Evaluating Publically Available Personal Health Records for Home Health – Laura Kneale/University of Washington 2. Data in Emergency Department Provider Notes at Time of Image Order Entry – Justin Rousseau/Harvard Medical School 3. Pediatric ECG Feature Identification – Emily Hendryx/Rice University 4. Learning to Diagnose with LSTM Recurrent Neural Networks – Zachary Lipton/University of California, San Diego 5. Automatic Detection of Drug-Drug Interactions Between Clinical Practice Guidelines – Geoffrey Tso/Veterans Administration
9:45 – 10:30 AM	**Posters and Coffee Break** *Location: Near registration table, outside of The Great Hall Meeting Room* **Topic 1 – Healthcare Informatics:** #101 Movement Disorders Journal: Testing an App to Track Parkinson's

Symptoms – Jeff Day/National Library of Medicine

#102 Design of a Subscription-Based Laboratory Result Notification System – Benjamin Slovis/Columbia University

#103 Medication Use Among Veterans Across Health Care Systems – Khoa Nguyen/Veterans Administration

#104 Designing a Telehealth Training Curriculum using a Telemental Health Model – Pamela Hoffman/Veterans Administration

#105 A Multi-Axial Based Knowledge Management System for Alerts – Rajdeep Brar/Yale University

#106 Improving and Applying Medical High-Throughput Machine Learning – Paul Bennett/University of Wisconsin-Madison

#107 Assessing the Delay in Communication Regarding Digital Inpatient Documentation – Ross Lordon/University of Washington

#108 Quantifying Burden of Treatment in Patients with Breast Cancer – Alex Cheng/Vanderbilt University

#109 Evaluating the Use of an Automated Section Identifier for Focused Information Extraction Tasks on a VA Big Data Corpus – Le-Thuy Tran/University of Utah

#110 Automatic Identification of High Impact Articles in PubMed to Support Clinical Decision-Making – Jiantao Bian/University of Utah

#111 Design Thinking in Radiation Oncology – Adam Rule/University of California, San Diego

#112 Prospective Study of a Kawasaki Disease Natural Language Processing Tool – Juan Chaparro/University of California, San Diego

#113 Modeling of Hypoplastic Left Heart Syndrome for Improved Decision Support – Charles Puelz/Rice University

#114 Taxonomic Classification of HIT Hazards Associated with EHR Implementation: Initial and Stabilization Phases – Paul Varghese/Harvard Medical School

#115 Teamwork Behaviors of Emergency Medical Service Teams in Pediatric Simulations – Nathan Bahr/Oregon Health & Science University

#116 Large-Scale Family Cohorts Linked to Electronic Health Records – Scott Hebbring/University of Wisconsin-Madison

Topic 2 – Bioinformatics/Computational Biology:

#201 Predicting Accidental Falls in People Aged 65 Years and Older – Mark Homer/Harvard Medical School

#202 Content-Based fMRI Activation Maps Retrieval – Alba G Seco de Herrera/National Library of Medicine

#203 The Epigenomic Landscape of Aberrant Splicing in Cancer – Donghoon Lee/Yale University

#204 Identifying and Resolving Inconsistencies in Biological Pathway Resources – Lucy Wang/University of Washington

#205 Conserved Transcriptional Regulators Control Divergent Toxin Production in Fungi – Abigail Lind/Vanderbilt University

#206 Determining Gene Expression Trends using Single-Cell RNA-seq with CREoLE – Geoffrey Schau/Oregon Health & Science University

#207 Analysis of Orphan Disease Gene Networks to Enable Drug Repurposing – Kelly Regan/The Ohio State University

#208 Signal-Oriented Pathway Analyses Reveal a Signaling Complex as a Synthetic Lethal Target for p53 Mutations – Songjian Lu/University of Pittsburgh

#209 Towards a Knowledge-Base for Biochemical Reasoning – Daniel McShan/University of Colorado

Topic 3 – Clinical Research Translational Informatics:

#301 Informatics Approaches for Evidence Appraisal and Synthesis – Andrew Goldstein/Columbia University

#302 Using Wearable Technology to Aid in the Classification of Different Cardiac Arrhythmias – Jessica Torres/Stanford University

#303 Predicting Heterogenous Causal Treatment Effects for First-Line Antihypertensives – Alejandro Schuler/Stanford University

#304 Acquiring and Representing Drug-Drug Interaction Knowledge and Evidence – Jodi Schneider/University of Pittsburgh

#305 Impact of Missing Data on Automatic Learning of Clinical Guidelines – Yuzhe Liu/University of Pittsburgh

#306 Understanding Clinical Trial Patient Screening from the Coordinator's Perspective – En-Ju Lin/The Ohio State University

#307 Standardizing Sample-Specific Metadata in the Sequence Read Archive – Matthew Bernstein/University of Wisconsin-Madison

#308 Causal Inference During Multisensory Speech Perception – John Magnotti/Baylor College of Medicine

#309 Data Mining for Identifying Candidate Drivers of Drug Response in Heterogeneous Cancer – Sheida Nabavi/University of Connecticut

10:30 – 11:30 AM

OSUWMC Innovations Showcase | Moderator: Dr. Peter J. Embi, Ohio State University
Location: US Bank Theater

- William D. Smoyer, MD – Vice President and Director of Center for Clinical and Translational Research, Nationwide Children's Hospital Research Institute
- Randi Foraker, PhD – Assistant Professor, Division of Epidemiology, College of Public Health
- Wondwossen Gebreyes, DVM, PhD – Professor, Department of Veterinary Preventative Medicine, College of Veterinary Medicine
- Colleen Spees, PhD, MEd, RDN, FAND – Assistant Professor, Department of Medical Dietetics, College of Medicine

11:30 AM – 12:30 PM

Lunch and Special Sessions (locations as noted below)

- Trainees: Birds of a Feather (Location: Performance Hall & Potter Plaza)
- Training Directors: Annual Training Directors Meeting (Location: Barbie Tootle Room)
- NLM Program Staff Webinar (Location: Hays Cape Room)

12:45 – 1:55 PM

Open Mic Session X1: Translational Bioinformatics and Clinical Research Informatics | Moderator: Dr. Bill Hersh, OHSU (12 speakers, 5 minutes per speaker including questions)
Location: US Bank Theater

1. Building a Centralized Resource for Computational Venom Research – Joseph Romano/Columbia University
2. Master Regulators of Cancer Drug Sensitivity – Michael Sharpnack/The Ohio State University

3. Using Rigorous Multi-Target Drug Profiles to Explore Off-Target Pathways – Aurora Blucher/Oregon Health & Science University
4. Applications of Deep Learning to Genomic Data – Timothy Lee/Stanford University
5. Prediction of Reproductive Outcomes in Structural Translocation Carriers – Archana Shenoy/Stanford University
6. Computational Analysis of Association of ClinVar Variants with DNA Palindromes – Viji Avali/University of Pittsburgh
7. Personalized Modeling for Identifying Genomic and Clinical Factors in Chronic Pancreatitis – Joyeeta Dutta-Moscato/University of Pittsburgh
8. A Macrophage-Specific Gene Signature to Predict Response to Treatment – Yasmin Lyons/University of Texas MD Anderson Cancer Center
9. Subtyping of Supratentorial Pediatric Brain Tumors Using RNAseq Data – Wayne Liang/University of Washington
10. From Genetic Informatics to a Biological Model: Analysis of Genetic Variants of SLC5A – Jamie Fox/University of Wisconsin-Madison
11. Dental Plaque Meta-omics for Diagnosis of Oral and Systemic Disease – Timothy Rhoads/University of Wisconsin-Madison
12. Inferring Mechanistic Detail from Qualitative Biological Models – Michael Kochen/Vanderbilt University

2:00 – 3:00 PM

Parallel Paper Focus Session A (locations are noted below)
(3 papers at 12 minutes each plus 24 minutes for Q&A)

Focus Session A1 | Moderator: Dr. Robert El-Kareh, UCSD
Location: US Bank Theater

- Conserved Elongation Factor Spt5 Affects Antisense Transcription in Fission Yeast – Scott Kallgren/Harvard Medical School
- Genotype to Phenotype Relationships in Autism Spectrum Disorders – Jonathan Chang/Columbia University
- Longitudinal Metabolome Wide Association Study of Cognitive Decline in Healthy Adults – Burcu Darst/University of Wisconsin-Madison

Focus Session A2 | Moderator: Dr. Carol Friedman, Columbia University
Location: Cartoon Room

- Predicting Required Diagnostic Tests from Patient Triage Data – Haley Hunter-Zinck/Veterans Administration
- Classification of Literature Derived Drug Side Effect Relationships – Justin Mower/Baylor College of Medicine
- Assessing the Potential Risk in Drug Prescriptions During Pregnancy – Ferdinand Dhombres/National Library of Medicine

Focus Session A3 | Moderator: Dr. John Hurdle, University of Utah
Location: Traditions Room

- Uncertainty Quantification (UQ) in Breast and Ovarian Cancer Risk Prediction Based on Self-Reported Family History – Lance Pflieger/University of Utah
- Performance Drift in Clinical Prediction Across Modeling Methodologies – Sharon Davis/Vanderbilt University
- Sample-Specific Sparsity Adjustment Improves Differential Abundance Analysis of

16S rRNA Data – Liyang Diao/Yale University

3:00 – 3:30 PM | **Posters and Coffee Break**
Location: Near registration table, outside of The Great Hall Meeting Room

Topic 1 – Healthcare Informatics:

- Jeff Day/National Library of Medicine; Benjamin Slovis/Columbia University; Khoa Nguyen/Veterans Administration; Pamela Hoffman/Veterans Administration; Rajdeep Brar/Yale University; Paul Bennett/University of Wisconsin-Madison; Ross Lordon/University of Washington; Alex Cheng/Vanderbilt University; Le-Thuy Tran/University of Utah; Jiantao Bian/University of Utah; Adam Rule/University of California, San Diego; Juan Chaparro/University of California, San Diego; Charles Puelz/Rice University; Paul Varghese/Harvard Medical School; Nathan Bahr/Oregon Health & Science University; Scott Hebbring/University of Wisconsin-Madison

Topic 2 – Bioinformatics/Computational Biology:

- Mark Homer/Harvard Medical School; Alba Seco de Herrera/National Library of Medicine; Donghoon Lee/Yale University; Lucy Wang/University of Washington; Abigail Lind/Vanderbilt University; Geoffrey Schau/Oregon Health & Science University; Kelly Regan/The Ohio State University; Songjian Lu/University of Pittsburgh; Daniel McShan/University of Colorado

Topic 3 – Clinical Research Translational Informatics:

- Andrew Goldstein/Columbia University; Jessica Torres/Stanford University; Alejandro Schuler/Stanford University; Jodi Schneider/University of Pittsburgh; Yuzhe Liu/University of Pittsburgh; En-Ju Lin/The Ohio State University; Matthew Bernstein/University of Wisconsin-Madison; John Magnotti/Baylor College of Medicine; Sheida Nabavi/University of Connecticut

3:30 – 4:45 PM | **Plenary Session #2** | Moderator: Dr. Larry Hunter, University of Colorado
Location: US Bank Theater (12 minutes per presentation, 3 minutes for Q&A)

1. Predicting Drug Response Curves in a Large Cancer Cell Line Screen – Nathan Lazar/Oregon Health & Science University
2. Aggressive Glioblastoma Phenotype Evolves Over Decade-Long Growing Phase – Daniel Rosenbloom/Columbia University
3. Unsupervised Deep Learning Reveals Prognostically Relevant Subtypes of Glioblastoma – Jonathan Young/University of Pittsburgh
4. Computational Studies of Protein-Protein Interface Mutations – Jennifer Gaines/Yale University
5. Modeling of the Minimally Gained Significant Region of Trisomy 12 in Chronic Lymphocytic Leukemia – Zachary Abrams/The Ohio State University

4:45 – 9:30 PM | **Location: Columbus Zoo and Aquarium**

4:45 – 5:30 PM | **Transportation Time | Ohio Union → Columbus Zoo**
Complementary shuttle that will take guests from The Ohio Union to the Columbus Zoo and Aquarium for the reception and dinner.

5:45 – 6:30 PM | **Reception**

6:30 – 8:30 PM | **Dinner**

8:00 – 9:30 PM | **Transportation Time | Zoo → Columbus Hilton Downtown**
Complementary shuttle from the Columbus Zoo back to the conference hotel.

NLM Informatics Training Conference 2016
The Ohio State University

Tuesday, June 28, 2016

6:30 – 8:00 AM	**Transportation Time | Conference Hotel → Ohio Union** There will be a complementary shuttle from the Columbus Hilton Downtown to The Ohio Union.
7:00 – 7:55 AM	**Posters and Breakfast** *Location: Outside of The Great Hall Meeting Room*
8:00 – 9:05 AM	**Open Mic Session X2**: Healthcare and Public Health Informatics | Moderator: Dr. Patricia Brennan, University of Wisconsin-Madison (3-4 minutes per speaker followed by 1-2 minutes Q&A) *Location: US Bank Theater* 1. New Network-Based Tools for Integrated Analysis of Biomedical Data – Andrew Laitman/Baylor College of Medicine 2. Promoting Observational Learning of Nutrition Through a Mobile Health Application – Michelle Chau/Columbia University 3. Outpatient Clinical Decision Support Rule Analysis – Mujeeb Basit/Harvard Medical School 4. DXplain Mobile: An Assessment of a Smartphone-Based Expert Diagnostic System – Baker Hamilton/Harvard Medical School 5. Computing the Impact of the Medicare Shared Savings Program – Fabricio Kury/National Library of Medicine 6. Assessing the Accuracy of Computing Clinical Quality Measures in the Ophthalmology Domain – Olubumi Akiwumi/Oregon Health & Science University 7. Technical Barriers to Situational Awareness in Laboratory Testing – Argus Athana-Crannell/University of California, San Diego 8. Share Happiness is Doubled: Time-Dependent Analysis of Sentiment on an Online Forum – Rebecca Marmor/University of California, San Diego 9. Grocery Transaction Data: Novel Ways to Understand Dietary Quality of Obesogenic Family Environment – Valli Chidambaram/University of Utah 10. Understanding User Requirements for a Recipe Recommender System – Diane Walker/University of Utah 11. Building a Tool to Support Women Experiencing Menopause to Track Health and Symptoms – Uba Backonja/University of Washington 12. Identifying Patients with Amyotrophic Lateral Sclerosis using Veterans Health Administration Data – Jennifer Aucoin/Veterans Administration 13. Acceptance of a Risk Estimation Tool for Colorectal Cancer Screening

– Cherie Luckhurst/Veterans Administration

9:05 – 10:05 AM

Parallel Paper Focus Session B (papers at 12 minutes each plus 24 minutes for Q&A) (*Locations as noted below*)

Focus Session B1 | Moderator: Dr. Michael Krauthammer, Yale University
Location: US Bank Theater

- A Bioinformatics Approach to Identify Novel Drugs Against Liver Cancer – Tasneem Motiwala/The Ohio State University
- Signatures of Accelerated Somatic Evolution on a Genome-wide Scale – Kyle Smith/University of Colorado
- Identification and Validation of CNVs using WGS Data from 274 Individuals – David Jakubosky/University of California, San Diego

Focus Session B2 | Moderator: Dr. Harry Hochheiser, University of Pittsburgh
Location: Cartoon Room

- Computing Geographical Access to Hospitals in Two Countries – Fabricio Kury/National Library of Medicine
- Bursting the Information Bubble: Designing Inpatient-Centered Technology Beyond the Hospital Room – Andrew Miller/University of Washington
- User-Centered Design and Evaluation of RxMAGIC: A System for Prescription Management and General Inventory Control for Low-Resource Settings – Arielle Fisher/University of Pittsburgh

Focus Session B3 | Moderator: Dr. John Magnotti, Baylor College of Medicine
Location: Traditions Room

- Clinical Decision Support Anomaly Pathways – Steven Kassakian/Oregon Health & Science University
- Medical Entity Recognition: a Meta-Learning Approach with Selective Data Augmentation – Asma Ben Abacha/National Library of Medicine
- Untangling the Structure of High-Throughput Sequencing Data with veRitas – David Moskowitz/Stanford University

10:05 – 10:50 AM

Posters and Coffee Break
Location: Near registration table, outside of The Great Hall Meeting Room

Topic 1 – Healthcare Informatics:

- Jeff Day/National Library of Medicine; Benjamin Slovis/Columbia University; Khoa Nguyen/Veterans Administration; Pamela Hoffman/Veterans Administration; Rajdeep Brar/Yale University; Paul Bennett/University of Wisconsin-Madison; Ross Lordon/University of Washington; Alex Cheng/Vanderbilt University; Le-Thuy Tran/University of Utah; Jiantao Bian/University of Utah; Adam Rule/University of California, San Diego; Juan Chaparro/University of California, San Diego; Charles Puelz/Rice University; Paul Varghese/Harvard Medical School; Nathan Bahr/Oregon Health & Science University; Scott Hebbring/University of Wisconsin-Madison

Topic 2 – Bioinformatics/Computational Biology:

- Mark Homer/Harvard Medical School; Alba Seco de Herrera/National Library of Medicine; Donghoon Lee/Yale University; Lucy Wang/University of Washington; Abigail Lind/Vanderbilt University; Geoffrey Schau/Oregon Health & Science University; Kelly Regan/The Ohio State University; Songjian Lu/University of Pittsburgh; Daniel McShan/University of Colorado

Topic 3 – Clinical Research Translational Informatics:

- Andrew Goldstein/Columbia University; Jessica Torres/Stanford University;

	Alejandro Schuler/Stanford University; Jodi Schneider/University of Pittsburgh; Yuzhe Liu/University of Pittsburgh; En-Ju Lin/The Ohio State University; Matthew Bernstein/University of Wisconsin-Madison; John Magnotti/Baylor College of Medicine/Sheida Nabavi/University of Connecticut
10:50 – 12:05 PM	**Plenary Session 3 |** Moderator: Dr. Mark Craven, University of Wisconsin-Madison (12 minutes per presentation, 3 minutes for Q&A) *Location: US Bank Theatre* 1. Modeling Neutral Evolution at Small Scales – Aaron Wacholder/University of Colorado 2. EHR-Wide GxE Study using Smoking Information Extracted from Clinical Notes – Travis Osterman/Vanderbilt University 3. High-Throughput Machine Learning from Electronic Health Records – Ross Kleiman/University of Wisconsin-Madison 4. Comparison of Variant Annotation Tool Terminology using the Sequence Ontology – Nicole Ruiz-Schultz/University of Utah 5. Constructing a Biomedical Relationship Database from Literature using DeepDive – Emily Mallory/Stanford University
12:05 – 1:00 PM	**Lunch and Special Sessions** (*locations as noted below*) • Trainees: Birds of a Feather (*Location: Performance Hall & Potter Plaza*) • Grants Management/X-Train Webinar (*Location: Barbie Tootle Room*)
1:00 – 2:15 PM	**Career Transitions Panel:** Moderator: Dr. Doug Fridsma, AMIA Former NLM trainees talk about their experiences in making the transition from pre-doctoral or post-doctoral fellow to their first research position. *Location: US Bank Theater* 1. Sheida Nabavi/University of Connecticut 2. Nick Soulakis/Northwestern University 3. Songijan Lu/University of Pittsburgh 4. Mike Conway/University of Utah 5. Scott Hebbring/University of Wisconsin-Madison/Marshfield Clinic 6. Kavishwar Wagholikar/Harvard Medical School 7. Meredith Zozus/University of Arkansas
2:15 – 2:30 PM	**Closing Session and Awards** | Dr. Peter J. Embi and Dr. Valerie Florance *Location: US Bank Theater*
2:30 – 3:30 PM	**Transportation Time | Ohio Union → Columbus Hilton Downtown → CMH Airport** **There will be a complementary shuttle that will take guests from the conference venue back to the hotel and to the airport.**

TRAINING CONFERENCE ATTENDEES AND PRESENTERS			
Name	**Affiliation**	**Presentation Session**	**E-Mail Address**
Columbia University			
George Hripcsak	**Training Director**		hripcsak@columbia.edu
Mary Boland	Predoctoral		mb3402@cumc.columbia.edu
William Brown III	Predoctoral		wb2253@cumc.columbia.edu
Jonathan Chang	Predoctoral	Focus Presentation	jsc2197@columbia.edu
Michelle Chau	Predoctoral	Open Mic	mmc2106@cumc.columbia.edu
Sylvia Cho	Predoctoral		sc3901@cumc.columbia.edu
Carol Friedman	Faculty		cf9@cumc.columbia.edu
Andrew Goldstein	Postdoctoral	Poster Presentation	ag3304@cumc.columbia.edu
Silis Jiang	Predoctoral		syj2108@cumc.columbia.edu
Erik Ladewig	Predoctoral		el2707@cumc.columbia.edu
Mari Millery	Postdoctoral		mm994@cumc.columbia.edu
Joseph Romano	Predoctoral	Open Mic	jdr2160@cumc.columbia.edu
Daniel Rosenbloom	Postdoctoral	Plenary Presentation	daniel.rosenbloom@gmail.com
Benjamin Slovis	Postdoctoral	Poster Presentation	bhs2133@cumc.columbia.edu
Gulf Coast Consortium			
Lydia Kavraki	**Training Director**		kavraki@rice.edu
GCC: Baylor College of Medicine			
Lee Call	Predoctoral		lee.call@bcm.edu
Evan Jones	Predoctoral		evan.jones@bcm.edu
Sangbae Kim	Postdoctoral		sangbae.kim@bcm.edu
Andrew Laitman	Predoctoral	Open Mic	laitman@bcm.edu
John Magnotti	Postdoctoral	Poster Presentation	john.magnotti@bcm.edu
Justin Mower	Predoctoral	Focus Presentation	justin.mower@bcm.edu
Stephen Wilson	Predoctoral		stephen.wilson@bcm.edu
GCC: M.D. Anderson Cancer Center			
Han Chen	Postdoctoral		hanchen601@gmail.com
Yasmin Lyons	Postdoctoral	Open Mic	ymehta@manderson.org
GCC: Rice University			
Jayvee Abella	Predoctoral		jayvee.r.abella@gmail.com
Leo Elworth	Predoctoral		chilleo@gmail.com
Emily Hendryx	Predoctoral	Plenary Presentation	emily.hendryx@rice.edu
Matthew Pena	Postdoctoral		mipena@gmail.com
Charles Puelz	Predoctoral	Poster Presentation	cpuelz@rice.edu
Harvard University			
Alexa McCray	**Training Director**		alexa_mccray@hms.harvard.edu
Mujeeb Basit	Postdoctoral	Open Mic	mujeebbasit@gmail.com
Marzyeh Ghassemi	Predoctoral		mghassem@mit.edu
Baker Hamilton	Postdoctoral	Open Mic	baker215@gmail.com
Mark Homer	Postdoctoral	Poster Presentation	marklhomer@gmail.com
Scott Kallgren	Postdoctoral	Focus Presentation	scott@hms.harvard.edu
Tristan Naumann	Predoctoral		tjn@mit.edu
Frank Pandolfe	Postdoctoral		frank_pandolfe@hms.harvard.edu
Justin Rousseau	Postdoctoral	Plenary Presentation	justin_rousseau@hms.harvard.edu
Paul Varghese	Postdoctoral	Poster Presentation	paul_varghese@hms.harvard.edu
Kavishwar Wagholikar	Faculty		kwagholikar@mgh.harvard.edu
Jia Wang	Postdoctoral		jia_wang@hms.harvard.edu
National Library of Medicine			
Patricia Brennan	**NLM Director**		Patricia.brennan@nih.gov
Valerie Florance	**Extramural Programs Director**		florancev@mail.nih.gov
Christine Ireland	**Chief Committee Management Officer**		irelanc@mail.nih.gov
Hua-Chuan Sim	**Chief Program Officer**		simh@mail.nih.gov

Paul Fontelo	**Training Director**		pfontelo@mail.nih.gov
Asma Ben Abacha	Postdoctoral	Focus Presentation	Asma.benabacha@nih.gov
Jeff Day	Postdoctoral	Poster Presentation	Jeffrey.Day@nih.gov
Ferdinand Dhombres	Postdoctoral	Focus Presentation	Ferdinand.dhombres@nih.gov
Fabricio Kury	Postdoctoral	Focus Presentation/Open Mic	Fabricio.kury@nih.gov
Alba Seco de Herrera	Postdoctoral	Poster Presentation	albagarcia@nih.gov
The Ohio State University			
Ümit Çatalyürek	**Training Director**		umit@bmi.osu.edu
Peter Embi	**Training Director**		peter.embi@osumc.edu
Philip Payne	**Training Director**		philip.payne@osumc.edu
Zachary Abrams	Predoctoral	Plenary Presentation	zachary.abrams@osumc.edu
Travis Johnson	Predoctoral		travis.johnson@osumc.edu
En-Ju (Deborah) Lin	Postdoctoral	Poster Presentation	en-ju.lin@osumc.edu
Tasneem Motiwala	Postdoctoral	Focus Presentation	tasneem.motiwala@osumc.edu
Kelly Regan	Predoctoral	Poster Presentation	kelly.regan@osumc.edu
Michael Sharpnack	Predoctoral	Open Mic	michael.sharpnack@osumc.edu
Sara Sinicropi-Yao	Predoctoral		sara.sinicropi-yao@osumc.edu
Christina Yu	Predoctoral		christina.yu@osumc.edu
Mark Zucker	Predoctoral		mark.zucker@osumc.edu
Oregon Health & Science University			
William Hersh	**Training Director**		hersh@ohsu.edu
Olubumi Akiwumi	Predoctoral		akiwumi@ohsu.edu
Nathan Bahr	Predoctoral	Poster Presentation	baharn@ohsu.edu
Aurora Blucher	Predoctoral	Open Mic	blucher@ohsu.edu
Steven Chamberlin	Postdoctoral		chamberst@ohsu.edu
Karen Eden	Faculty		edenk@ohsu.edu
Julian Egger	Predoctoral		eggerj@ohsu.edu
Mark Engelstad	Postdoctoral		engelsta@ohsu.edu
Erin Hickman	Postdoctoral		hickmaer@ohsu.edu
Ilya Ivlev	Postdoctoral		ivlev@ohsu.edu
Steven Kassakian	Postdoctoral	Focus Presentation	kassakia@ohsu.edu
Nathan Lazar	Predoctoral	Plenary Presentation	lazar@ohsu.edu
Eric Leung	Predoctoral		leunge@ohsu.edu
Shannon McWeeney	Faculty		mcweeney@ohsu.edu
Thomas (Josh) Meyer	Postdoctoral		meyerjos@ohsu.edu
Smriti Rao	Predoctoral		rasm@ohsu.edu
Geoffrey Schau	Predoctoral	Poster Presentation	schau@ohsu.edu
Kristen Stevens	Predoctoral		stevenkr@ohsu.edu
Matthew Sundling	Postdoctoral		sundling@ohsu.edu
Dana Womack	Predoctoral		womacda@ohsu.edu
Stanford University			
Russ Altman	**Training Director**		russ.altman@stanford.edu
Hunter Boyce	Predoctoral		hboyce@stanford.edu
Diego Calderon	Predoctoral		dcal@stanford.edu
Timothy Lee	Postdoctoral	Open Mic	tklee@stanford.edu
Emily Mallory	Predoctoral	Plenary Presentation	emily.mallory@stanford.edu
David Moskowitz	Predoctoral	Focus Presentation	dmosk@stanford.edu
Steven Schaffert	Postdoctoral		nlennart@stanford.edu
Alejandro Schuler	Predoctoral	Poster Presentation	aschuler@stanford.edu
Anna Shcherbina	Predoctoral		annashch@stanford.edu
Archana Shenoy	Postdoctoral	Open Mic	shenoyas@stanford.edu
Erika Strandberg	Predoctoral		estrandb@stanford.edu
Jessica Torres	Predoctoral	Poster Presentation	jntorres@stanford.edu
Alice Yu	Predoctoral		xalliceyu@stanford.edu
Darvin Yi	Predoctoral		darvinyi@stanford.edu
University of California, San Diego			

Lucila Ohno-Machado	**Training Director**		lohnomachado@ucsd.edu
Argus Athana-Crannell	Predoctoral	Open Mic	argus@athanas.org
Juan Chaparro	Postdoctoral	Poster Presentation	jchaparro@ucsd.edu
Margaret Donovan	Predoctoral		mkrdonovan@gmail.com
Michelle Dow	Predoctoral		mdow@ucsd.edu
Robert El-Kareh	Faculty		relkareh@ucsd.edu
Kemal Eren	Predoctoral		keren@ucsd.edu
YingXiang Huang	Predoctoral		yi108@eng.ucsd.edu
David Jakubosky	Predoctoral	Focus Presentation	David.jakubosky@gmail.com
Zachary Lipton	Predoctoral	Plenary Presentation	zlipton@cs.ucsd.edu
Rebecca Marmor	Postdoctoral	Open Mic	ramarmor@ucsd.edu
Gilbert Ramirez	Postdoctoral		gilbert.ramirez@gmail.com
Adam Rule	Predoctoral	Poster Presentation	acrule@ucsd.edu
University of Colorado, HSC			
Lawrence Hunter	**Training Director**		larry.hunter@ucdenver.edu
KC Anderson	Predoctoral		kelsey.anderson@ucdenver.edu
Ben Garcia	Predoctoral		benjamin.garcia@ucdenver.edu
Cody Glickman	Predoctoral		Cody.glickman@ucdenver.eu
Kimberly Kanigel Winner	Postdoctoral		kimberly.kanigelwinner@ucdenver.edu
Will Kindel	Postdoctoral		kindel.will@gmail.com
Ettie Lipner	Postdoctoral		lipnere@NJHealth.org
Dan Mcshan	Postdoctoral		daniel.mcshan@ucdenver.edu
Brian Ross	Postdoctoral		brian.ross@ucdenver.edu
Michael Shaffer	Predoctoral		michael.shaffer@ucdenver.edu
Jenny Wen Shi	Postdoctoral		wen.2.shi@ucdenver.edu
Kyle Smith	Predoctoral	Focus Presentation	Kyle.s.smith@ucdenver.edu
Aaron Wacholder	Predoctoral	Plenary Presentation	Aaron.wacholder@ucdenver.edu
Elizabeth White	Postdoctoral		elizabeth.white@ucdenver.edu
Nicolle Witte	Predoctoral		nicolle.witte@ucdenver.edu
University of Pittsburgh			
Rebecca Jacobson	**Training Director**		rebeccaj@pitt.edu
Saja Al-Alawneh	Predoctoral		saa147@pitt.edu
John Aronis	Postdoctoral		jma18@pitt.edu
Viji Avali	Postdoctoral	Open Mic	viji.avali@pitt.edu
Luca Calzoni	Postdoctoral		lucacalzoni@pitt.edu
Rafael Ceschin	Predoctoral		rcc10@pitt.edu
Michael Ding	Predoctoral		dingm@pitt.edu
Amie Draper	Predoctoral		ajd109@pitt.edu
Joyeeta Dutta-Moscato	Predoctoral	Open Mic	jod30@pitt.edu
Arielle Fisher	Predoctoral	Focus Presentation	Arf56@pitt.edu
Harry Hochheiser	Faculty		harryh@pitt.edu
Andy King	Predoctoral		ajk77@pitt.edu
Erich Kummerfeld	Postdoctoral		ekummerfeld@pitt.edu
Yuzhe Liu	Predoctoral	Poster Presentation	y.liu@pitt.edu
Sam Rosko	Predoctoral		scr25@pitt.edu
Jodi Schneider	Postdoctoral	Poster Presentation	jos188@pitt.edu
Jonathan Young	Postdoctoral	Plenary Presentation	Jdy10@pitt.edu
University of Utah			
Julio Facelli	**Training Director**		julio.facelli@utah.edu
John Hurdle	**Training Director**		john.hurdle@utah.edu
Jiantao Bian	Predoctoral	Poster Presentation	jiantao.bian@utah.edu
Samuel Brady	Postdoctoral		samuel.brady@utah.edu
Valli Chidambaram	Predoctoral	Open Mic	u0758965@utah.edu
Dedra Cutler	Postdoctoral		dedra.cutler@utah.edu
Rolando Hernandez	Predoctoral		u0977066@utah.edu
Albert Park	Postdoctoral		alpark1216@gmail.com
Lance Pflieger	Predoctoral	Focus Presentation	l.pflieger@utah.edu

Casey Rommel	Predoctoral		casey.rommel@utah.edu
Nicole Ruiz-Schultz	Predoctoral	Plenary Presentation	nicole.ruiz@utah.edu
Jianyin Shao	Postdoctoral		jianyin.shao@utah.edu
Michael Sinclair	Postdoctoral		michael.sinclair@utah.edu
Teresa Taft	Predoctoral		teresa.taft@utah.edu
Le-Thuy Tran	Postdoctoral	Poster Presentation	ltran@cs.utah.edu
Diane Walker	Predoctoral	Open Mic	diane.walker@utah.edu
Rosalie Waller	Predoctoral		rosalie.waller@utah.edu
Charlene Weir	Faculty		charlene.weir@utah.edu
University of Washington			
George Demiris	**Training Director**		gdemiris@u.washington.edu
Uba Backonja	Postdoctoral	Open Mic	backonja@uw.edu
Shefali Haldar	Predoctoral		shaldar@uw.edu
Ryan James	Predoctoral		rcjames@uw.edu
Will Kearns	Predoctoral		kearnsw@uw.edu
Laura Kneale	Predoctoral	Plenary Presentation	lkneale@uw.edu
Wayne Liang	Postdoctoral	Open Mic	waynehl@uw.edu
Ross Lordon	Predoctoral	Poster Presentation	rlordon@uw.edu
Sean Mikles	Predoctoral		smikles@uw.edu
Andrew Miller	Postdoctoral	Focus Presentation	millerad@uw.edu
Carolyn Paisie	Postdoctoral		capaisie@uw.edu
Donahue Smith	Predoctoral		dssmith@uw.edu
Lucy Wang	Predoctoral	Poster Presentation	lucylw@uw.edu
University of Wisconsin-Madison			
Mark Craven	**Training Director**		craven@biostat.wisc.edu
Lauren Baker	Predoctoral		iwicki@wisc.edu
Paul Bennett	Predoctoral	Poster Presentation	psbennett@wisc.edu
Matthew Bernstein	Predoctoral	Poster Presentation	matthewb@cs.wisc.edu
Anthony Cesnik	Predoctoral		cesnick@wisc.edu
Vincent Chen	Postdoctoral		vbchen@wisc.edu
Alex Cobain	Predoctoral		cobain@cs.wisc.edu
Sam Condon	Predoctoral		sgcondon@wisc.edu
Burcu Darst	Predoctoral	Focus Presentation	bdarst@wisc.edu
Jamie Fox	Postdoctoral	Open Mic	fox.jaime@mcrf.mfldclin.edu
Scott Hebbring	Faculty		Hebbring.scott@mcrf.mfldclin.edu
Ross Kleiman	Predoctoral	Plenary Presentation	rkleiman@cs.wisc.edu
Enaida Mendonca	Faculty		emendonca@wisc.edu
Katie Overmyer	Postdoctoral		kappacoo@gmail.com
Timothy Rhoads	Postdoctoral	Open Mic	trhoads@wisc.edu
Yuriy Sverchkov	Postdoctoral		yuriy.sverchkov@wisc.edu
Lindsay Traeger	Postdoctoral		Lltraeger@wisc.edu
Vanderbilt University			
Cindy Gadd	**Training Director**		cindy.gadd@vanderbilt.edu
Alex Cheng	Predoctoral	Poster Presentation	alex.cheng@vanderbilt.edu
Sharon Davis	Predoctoral	Focus Presentation	sharon.e.davis@vanderbilt.edu
Dara Eckerle Mize	Postdoctoral		dara.e.mize@vanderbilt.edu
Michael Greer	Predoctoral		michael.j.greer@vanderbilt.edu
Morgan Harrell	Postdoctoral		morgan.harrell@vanderbilt.edu
Gretchen Jackson	Faculty		gretchen.jackson@vanderbilt.edu
Michael Kochen	Predoctoral	Open Mic	michael.a.kochen@vanderbilt.edu
Matthew Lenert	Predoctoral		matthew.c.lenert@vanderbilt.edu
Abigail Lind	Predoctoral	Poster Presentation	abigail.l.lind@vanderbilt.edu
Travis Osterman	Postdoctoral	Plenary Presentation	travis.j.osterman@vanderbilt.edu
Jamie Robinson	Postdoctoral		jamie.r.robinson@vanderbilt.edu
Sara Savage	Postdoctoral		sara.r.savage@vanderbilt.edu
Bryan Steitz	Predoctoral		bryan.steitz@vanderbilt.edu
Linda Zhang	Predoctoral		lin.zhang@vanderbilt.edu

Yale University			
Cynthia Brandt	**Training Director**		cynthia.brandt@yale.edu
Michael Krauthammer	**Training Director**		mkrauthammer@yale.edu
Richard Shiffman	**Training Director**		richard.shiffman@yale.edu
Rajdeep Brar	Postdoctoral	Poster Presentation	rajdeep.brar@yale.edu
Liyang Diao	Postdoctoral	Focus Presentation	Liyang.diao@yale.edu
Christopher Fragoso	Predoctoral		christopher.fragoso@yale.edu
Jennifer Gaines	Predoctoral	Plenary Presentation	jennifer.gaines@yale.edu
Mohammed Khan	Postdoctoral		mohammed.khan@yale.edu
Michael Klein	Predoctoral		michael.klein@yale.edu
Donghoon Lee	Predoctoral	Poster Presentation	donghoon.lee@yale.edu
Kevin Lopez	Predoctoral		Kevin.Lopez@yale.edu
Kyle McGregor	Postdoctoral		kyle.mcgregor@yale.edu
Mate Nagy	Predoctoral		mate.nagy@yale.edu
Alexandra Signoriello	Predoctoral		alexandra.signoriello@yale.edu
Peter Williams	Predoctoral		peter.williams@yale.edu
Veterans Administration			
Steven Brown	**Training Director**		steven.brown@va.gov
Jennifer Aucoin	Postdoctoral	Open Mic	jennifer.aucoin@va.gov
Pamela Hoffman	Postdoctoral	Poster Presentation	pamela.hoffman@va.gov
Haley Hunter-Zinck	Postdoctoral	Focus Presentation	Haley.hunter-zinck@va.gov
Cherie Luckhurst		Open Mic	cherie.luckhurst@va.gov
Khoa Nguyen		Poster Presentation	khoa.nguyen6@va.gov
Geoffrey Tso	Postdoctoral	Plenary Presentation	Geoffrey.tso@va.gov
Career Transitions Panel			
Mike Conway	University of Utah	Podium Presentation	mike.conway@utah.edu
Scott Hebbring	U. Wisconsin-Madison	Podium & Poster Presentation	Hebbring.scott@mcrf.mfldclin.edu
Songjian Lu	University of Pittsburgh	Podium & Poster Presentation	songjian@pitt.edu
Sheida Nabavi	University of Connecticut	Podium & Poster Presentation	nabavi@engr.uconn.edu
Nick Soulakis	Northwestern University	Podium Presentation	Nicholas.soulakis@northwestern.edu
Kavishwar Wagholikar	Harvard University	Podium Presentation	kwagholikar@mgh.harvard.edu
Meredith Zozus	University of Arkansas	Podium Presentation	mzozus@uams.edu

Programmatic Administrative Contacts			
Institution	**Name**	**Title**	**Email Address**
Columbia University	Marina Bonanno	Graduate Program Manager	mmb2058@cumc.columbia.edu
	Luz-Raquel Perez	Department Administrator	rp302@cumc.columbia.edu
Harvard University	Aimee Smith	Program Coordinator	aimee_smith@hms.harvard.edu
	Katherine Flannery	Program Manager	katherine_flannery@hms.harvard.edu
Ohio State University	James Gentry	Education Program Manager	James.Gentry@osumc.edu
Oregon Health & Science University	Diane Doctor	Ed. Programs Coordinator	doctord@ohsu.edu
	Andrea Ilg	Ed. Programs Administrator	ilgan@ohsu.edu
	Lynne Schwabe	Administrative Assistant	schwabel@ohsu.edu
Gulf Coast Consortia	Melissa Glueck	Keck Center Associate Director	glueck@rice.edu
	Karen Ethun	GCC Executive Director	kethun@rice.edu
Stanford University	Mary Jeanne Oliva	Student Services Officer	mjoliva@stanford.edu
UC San Diego	Hailey Marshall	Division Coordinator	hlmarshall@ucsd.edu
University of Colorado	Elizabeth Wethington	Lead Program Administrator	elizabeth.wethington@ucdenver.edu
	Kathy Thomas	Administrative Coordinator	kathy.r.thomas@ucdenver.edu
University of Pittsburgh	Toni Porterfield	Training Program Manager	tls18@pitt.edu
University of Utah	Angela Matthes	Department Manager	angela.matthes@utah.edu
	Linda Galbreath	Grants and Contracts Officer	linda.galbreath@hsc.utah.edu
University of Washington	Laura Brewsaugh	Graduate Program Advisor	lorab2@uw.edu
	Heidi Kelm	Department Administrator	heidi5@uw.edu
University of Wisconsin-Madison	Karen Nafzger	Program Administrator	nafzger@biotech.wisc.edu
	Louise Pape	Program Coordinator	lpape@wisc.edu

Vanderbilt University	Claudia McCarn	Program Manager	claudia.mccarn@vanderbilt.edu
	Rischelle Jenkins	Training Program Coordinator	rischelle.jenkins@vanderbilt.edu
Yale University	Leigh Clemens	Financial Assistant IV	leigh.clemens@yale.edu
	Robin Einbinder	Assistant Administrator	robin.einbinder@yale.edu
	Lisa Sobel	Graduate Registrar	lisa.sobel@yale.edu

PLENARY/FOCUS SESSION PRESENTATIONS (Listed Alphabetically by Presenter)

Lipton, Zachary	**University of California, San Diego**	**Learning to Diagnose with LSTM Recurrent Neural Networks**	**20**
Mallory, Emily	**Stanford University**	**Constructing a Biomedical Relationship Database from Literature using DeepDive**	**37**
Miller, Andrew	**University of Washington**	**Bursting the Information Bubble: Designing Inpatient-Centered Technology Beyond the Hospital Room**	**32**
Moskowitz, David	**Stanford University**	**Untangling the Structure of High-Throughput Sequencing Data with veRitas**	**34**
Motiwala, Tasneem	**The Ohio State University**	**A Bioinformatics Approach to Identify Novel Drugs Against Liver Cancer**	**30**
Mower, Justin	**Baylor College of Medicine**	**Classification of Literature Derived Drug Side Effect Relationships**	**24**
Osterman, Travis	**Vanderbilt University**	**EHR-Wide GxE Study using Smoking Information Extracted from Clinical Notes**	**35**
Pflieger, Lance	**University of Utah**	**Uncertainty Quantification (UQ) in Breast and Ovarian Cancer Risk Prediction Based on Self-Reported Family History**	**25**
Rosenbloom, Daniel	**Columbia University**	**Aggressive Glioblastoma Phenotype Evolves Over Decade-Long Growing Phase**	**27**
Rousseau, Justin	**Harvard Medical School**	**Data in Emergency Department Provider Notes at Time of Image Order Entry**	**19**
Ruiz-Schultz, Nicole	**University of Utah**	**Comparison of Variant Annotation Tool Terminology using the Sequence Ontology**	**36**
Smith, Kyle	**University of Colorado**	**Signatures of Accelerated Somatic Evolution on a Genome-wide Scale**	**30**
Tso, Geoffrey	**Veterans Administration**	**Automatic Detection of Drug-Drug Interactions Between Clinical Practice Guidelines**	**21**
Wacholder, Aaron	**University of Colorado**	**Modeling Neutral Evolution at Small Scales**	**35**
Young, Jonathan	**University of Pittsburgh**	**Unsupervised Deep Learning Reveals Prognostically Relevant Subtypes of Glioblastoma**	**28**

POSTER PRESENTATIONS (Listed Alphabetically By Presenter)

Presenter	Institution	Title	Page
Bahr, Nathan	**Oregon Health & Science University**	**#115 – Teamwork Behaviors of Emergency Medical Service Teams in Pediatric Simulations**	**45**
Bennett, Paul	**University of Wisconsin-Madison**	**#106 – Improving and Applying Medical High-Throughput Machine Learning**	**40**
Bernstein, Matthew	**University of Wisconsin-Madison**	**#307 – Standardizing Sample-Specific Metadata in the Sequence Read Archive**	**54**

Bian, Jiantao	**University of Utah**	**#110 – Automatic Identification of High Impact Articles in PubMed to Support Clinical Decision-Making**	**42**
Brar, Rajdeep	**Yale University**	**#105 – A Multi-Axial Based Knowledge Management System for Alerts**	**40**
Chaparro, Juan	**University of California, San Diego**	**#112 – Prospective Study of a Kawasaki Disease Natural Language Processing Tool**	**43**
Cheng, Alex	**Vanderbilt University**	**#108 – Quantifying Burden of Treatment in Patients with Breast Cancer**	**41**
Day, Jeff	**National Library of Medicine**	**#101 – Movement Disorders Journal: Testing an App to Track Parkinson's Symptoms**	**38**
Goldstein, Andrew	**Columbia University**	**#301 – Informatics Approaches for Evidence Appraisal and Synthesis**	**51**
Hebbring, Scott	**University of Wisconsin-Madison**	**#116 – Large-Scale Family Cohorts Linked to Electronic Health Records**	**45**
Hoffman, Pamela	**Veterans Administration**	**#104 – Designing a Telehealth Training Curriculum using a Telemental Health Model**	**39**
Homer, Mark	**Harvard Medical School**	**#201 – Predicting Accidental Falls in People Aged 65 Years and Older**	**46**
Lee, Donghoon	**Yale University**	**#203 – The Epigenomic Landscape of Aberrant Splicing in Cancer**	**47**
Lin, En-Ju	**The Ohio State University**	**#306 – Understanding Clinical Trial Patient Screening from the Coordinator's Prospective**	**53**
Lind, Abigail	**Vanderbilt University**	**#205 – Conserved Transcriptional Regulators Control Divergent Toxin Production in Fungi**	**48**
Liu, Yuzhe	**University of Pittsburgh**	**#305 – Impact of Missing Data on Automatic Learning of Clinical Guidelines**	**53**
Lordon, Ross	**University of Washington**	**#107 – Assessing the Delay in Communication Regarding Digital Inpatient Documentation**	**41**
Lu, Songjian	**University of Pittsburgh**	**#208 – Signal-Oriented Pathway Analyses Reveal a Signaling Complex as a Synthetic Lethal Target for p53 Mutations**	**49**
Magnotti, John	**Baylor College of Medicine**	**#308 – Causal Inference During Multisensory Speech Perception**	**54**
McShan, Daniel	**University of Colorado**	**#209 – Towards a Knowledge-Base for Biochemical Reasoning**	**50**
Nabavi, Sheida	**University of Connecticut**	**#309 – Data Mining for Identifying Candidate Drivers of Drug Response in Heterogeneous Cancer**	**55**
Nguyen, Khoa	**Veterans Administration**	**#103 – Medication Use Among Veterans Across Health Care Systems**	**39**
Puelz, Charles	**Rice University**	**#113 – Modeling of Hypoplastic Left Heart Syndrome for Improved Decision Support**	**44**

Regan, Kelly	**The Ohio State University**	**#207 – Analysis of Orphan Disease Gene Networks to Enable Drug Repurposing**	**49**
Rule, Adam	**University of California, San Diego**	**#111 – Design Thinking in Radiation Oncology**	**43**
Schau, Geoffrey	**Oregon Health & Science University**	**#206 – Determining Gene Expression Trends using Single-Cell RNA-seq with CREoLE**	**48**
Schneider, Jodi	**University of Pittsburgh**	**#304 – Acquiring and Representing Drug-Drug Interaction Knowledge and Evidence**	**52**
Schuler, Alejandro	**Stanford University**	**#303 – Predicting Heterogeneous Causal Treatment Effects for First-Line Antihypertensives**	**52**
Seco de Herrera, Alba	**National Library of Medicine**	**#202 – Content-Based fMRI Activation Maps Retrieval**	**46**
Slovis, Benjamin	**Columbia University**	**#102 – Design of a Prescription-Based Laboratory Result Notification System**	**38**
Torres, Jessica	**Stanford University**	**#302 – Using Wearable Technology to Aid in the Classification of Different Cardiac Arrhythmias**	**51**
Tran, Le-Thuy	**University of Utah**	**#109 – Evaluating the Use of an Automated Section Identifier for Focused Information Extraction Tasks on a VA Big Data Corpus**	**42**
Varghese, Paul	**Harvard Medical School**	**#114 – Taxonomic Classification of HIT Hazards Associated with EHR Implementation: Initial and Stabilization Phases**	**44**
Wang, Lucy	**University of Washington**	**#204 – Identifying and Resolving Inconsistencies in Biological Pathway Resources**	**47**

OPEN MIC PRESENTATIONS X1 AND X2 (Listed Alphabetically by Presenter)

Presenter	Institution	Title
Akiwumi, Olubumi	**Oregon Health & Science University**	**Assessing the Accuracy of Computing Clinical Quality Measures in the Ophthalmology Domain**
Athana-Crannell, Argus	**University of California, San Diego**	**Technical Barriers to Situational Awareness in Laboratory Testing**
Aucoin, Jennifer	**Veterans Administration**	**Identifying Patients with Amyotrophic Lateral Sclerosis using Veterans Health Administration Data**
Avali, Viji	**University of Pittsburgh**	**Computational Analysis of Association of ClinVar Variants with DNA Palindromes**
Backonja, Uba	**University of Washington**	**Building a Tool to Support Women Experiencing Menopause to Track Health and Symptoms**
Basit, Mujeeb	**Harvard Medical School**	**Outpatient Clinical Decision Support Rule Analysis**
Blucher, Aurora	**Oregon Health & Science University**	**Using Rigorous Multi-Target Drug Profiles to Explore Off-Target Pathways**
Chau, Michelle	**Columbia University**	**Promoting Observational Learning of Nutrition**

		Through a Mobile Health Application
Chidambaram, Valli	**University of Utah**	**Grocery Transaction Data: Novel Ways to Understand Dietary Quality of Obesogenic Family Environment**
Dutta-Moscato, Joyeeta	**University of Pittsburgh**	**Personalized Modeling for Identifying Genomic and Clinical Factors in Chronic Pancreatitis**
Fox, Jamie	**University of Wisconsin-Madison**	**From Genetic Informatics to Biological Model: Analysis of Genetic Variants of SLC5A**
Hamilton, Baker	**Harvard Medical School**	**DXplain Mobile: An Assessment of a Smart Phone-Based Expert Diagnostic System**
Kochen, Michael	**Vanderbilt University**	**Inferring Mechanistic Detail from Quantitative Biological Models**
Kury, Fabricio	**National Library of Medicine**	**Computing the Impact of the Medicare Shared Savings Program**
Laitman, Andrew	**Baylor College of Medicine**	**New Network-Based Tools for Integrated Analysis of Biomedical Data**
Lee, Timothy	**Stanford University**	**Applications of Deep Learning to Genomic Data**
Liang, Wayne	**University of Washington**	**Subtyping of Supratentorial Pediatric Brain Tumors using RNAseq Data**
Luckhurst, Cherie	**Veterans Administration**	**Acceptance of a Risk Estimation Tool for Colorectal Cancer Screening**
Lyons, Yasmin	**University of Texas MD Anderson Cancer Center**	**A Macrophage-Specific Gene Signature to Predict Response to Treatment**
Marmor, Rebecca	**University of California, San Diego**	**Share Happiness is Doubled: Time-Dependent Analysis of Sentiment on an Online Forum**
Rhoads, Timothy	**University of Wisconsin-Madison**	**Dental Plaque Meta-Omics for Diagnosis of Oral and Systemic Disease**
Romano, Joseph	**Columbia University**	**Building a Centralized Resource for Computational Venom Research**
Sharpnack, Michael	**The Ohio State University**	**Master Regulators of Cancer Drug Sensitivity**
Shenoy, Archana	**Stanford University**	**Prediction of Reproductive Outcomes in Structural Translocation Carriers**
Walker, Diane	**University of Utah**	**Understanding User Requirements for a Recipe Recommender System**

Evaluating Publically Available Personal Health Records for Home Health

Authors: Laura Kneale, Yong Choi, Sean Mikles, George Demiris, University of Washington

Abstract: Personal health records (PHRs) were designed to encourage patient engagement. Frequent utilizers of the healthcare system, such as homebound older adults receiving home health services, may benefit from PHRs; however, PHRs have not been evaluated for use in home health. We identified existing PHRs using MyPHR.com, a systematic literature review, and Healthit.gov. We identified the similarities and differences between PHR functionality with the purpose to evaluate how the existing systems would benefit home health clients.

97 PHRs were initially identified, and 22 PHRs met our inclusion criteria. Our preliminary findings suggest that significant gaps exist across the PHRs. For example, only 2 (9.1%) PHRs provided role-based proxy access for informal caregivers, 6 (27.3%) allowed users to upload PDF documents from previous clinical encounters, and 4 (18.2%) were flexible in allowing consumers to choose what data elements to track (e.g. weight, diet, clinical values, etc.). In addition, we are currently assessing the PHRs' usability from a home health client, informal caregiver, and home health nurse perceptive.

We suggest that available PHRs may be difficult to implement in home health. In this talk, we will provide recommendations to improve utility, and ultimately utilization, of PHRs with home health clients.

Data in Emergency Department Provider Notes at Time of Image Order Entry

Authors: Justin Rousseau, Ivan Ip, Ali Raja, Vlad Valtchinov, Ramin Khorasani, Harvard Medical School, Brigham and Women's Hospital

Abstract: Objective: Identify opportunities to improve the communication between ordering providers and radiologists at the time of image ordering, which currently is insufficient, posing a patient safety concern. **Materials and Methods**: We evaluated observational data documented in electronic health record (EHR) notes prior to image ordering from 666 consecutive Emergency Department encounters over an 18-month study period for adult patients with headaches during which head CT was performed. We compared relevant concepts specific to headache extracted via ontology-based natural language processing of notes to image order requisitions. **Results**: History of present illness (HPI) was initially submitted in 33.9% and completed in 23.4% of encounters prior to image ordering. The number of concepts specific to headache per note was significantly greater than the number of indications per image order requisition (median 3 vs. 1; $p<0.0001$). There was no significant difference between the number of concepts in HPIs completed prior to image ordering compared to those completed after image ordering ($p=0.07$). **Discussion**: EHR documentation provides a source of valuable information that could be used in an automated fashion to facilitate and enhance the imaging ordering process. **Conclusion**: Future work is needed to assess the utility of EHR data prior to image ordering.

Pediatric ECG Feature Identification

Authors: Emily P Hendryx[1], Craig G Rusin[2], Beatrice M Riviere[1]
[1] Rice University, Houston, TX, [2] Baylor College of Medicine and Texas Children's Hospital, Houston, TX

Abstract: Since each part of the electrocardiogram (ECG) corresponds to a different stage in the cardiac cycle, tracking changes in individual ECG features over time can help physicians gain further insight into changes in a patient's clinical status. However, expecting physicians to fully analyze ECG subtleties in real time while analyzing the rest of the presented patient data is impractical, especially over longer periods of time. The goal of this work, therefore, is to automate the ECG feature-identification process on a beat-by-beat basis.

While some algorithms for identifying individual ECG features exist, these methods typically rely on specific timing thresholds and are derived from adult data. To better serve the pediatric population – specifically those with congenital heart disease – we are developing a library of key pediatric ECG morphologies using data collected from the bedside monitors at Texas Children's Hospital. Key morphologies for the library are identified via the CUR matrix factorization. This beat selection leads to the definition of morphology classes to be used in conjunction with dynamic time warping in identifying individual ECG features in unlabeled beats. The labeled features can then be considered in the development of predictive models for real-time clinical decision support.

This research was funded by a training fellowship from the Gulf Coast Consortia, on the Training Program in Biomedical Informatics, National Library of Medicine (NLM) T15LM007093, PD – Lydia E. Kavraki.

Learning to Diagnose with LSTM Recurrent Neural Networks

Authors: Zachary C Lipton, David C Kale, Charles Elkan, Randall Wetzell, University of California, San Diego

Abstract: Clinical medical data, especially in the ICU, consist of multivariate time series of observations. For each patient visit, sensor data and lab test results are recorded in the patient's electronic health record. While potentially containing a wealth of insights, the data is difficult to mine effectively, owing to varying length, irregular sampling and missing data. Recurrent neural networks, particularly those using Long Short-Term Memory (LSTM) hidden units, are powerful and increasingly popular models for learning from sequence data. They effectively model varying length sequences and capture long-range dependencies. We present the first study to empirically evaluate the ability of LSTMs to recognize patterns in multivariate time series of clinical measurements. Specifically, we consider multilabel classification of diagnoses, training a model to classify 128 diagnoses given 13 frequently but irregularly sampled clinical measurements. First, we establish the effectiveness of a simple LSTM network for modeling clinical data. Then we demonstrate a straightforward and effective training strategy in which we replicate targets at each sequence step. Trained only on raw time series, our models outperform several strong baselines, including a multilayer perceptron, recognizing diabetic ketoacidosis, idiopathic scoliosis, asthma and brain neoplasms all with AUC > .85 and F1 > .5.

Automatic Detection of Drug-Drug Interactions Between Clinical Practice Guidelines

Authors: Geoffrey J Tso[1,2], Samson W Tu[2], Mark A Musen[2], Mary K Goldstein[1,2] ,
[1]Dept. of Veterans Affairs VA Palo Alto Health Care System, Palo Alto, CA; [2]Stanford University School of Medicine, Stanford, CA

Abstract: Since many patients have multiple chronic conditions, they are commonly prescribed many medications that can potentially have clinically significant drug-drug interactions (DDI). However, these DDIs are rarely discussed in clinical practice guidelines (CPG). Knowing potential interactions between treatment plans is important in point of care clinical decision making and in clinical decision support (CDS) systems for patients with multiple chronic conditions. In this study, we describe and validate a method for automatically detecting DDIs between CPG recommendations. The system extracts drug and drug class recommendations from narrative CPGs, normalizes the terms, creates a mapping of drugs and drug classes, and then identifies occurrences of DDIs between CPG pairs. We analyzed 75 CPGs written by national organizations in the United States that discuss outpatient management of common chronic diseases. Using a reference list of 360 high risk and clinically significant DDIs as determined by an expert panel, our preliminary analysis identifies 108 of these DDIs in 38 CPG pairs (18 unique CPGs). Four of the CPGs contained specific discussion about these possible high risk DDIs. This study identifies important gaps in CPGs and provides a method to prevent clinically significant DDIs in a CDS system supporting multiple chronic conditions.

Conserved Elongation Factor Spt5 Affects Antisense Transcription in Fission Yeast

Authors: Scott P Kallgren[1], Ameet Shetty[2], Burak H Alver[1], Peter J Park[1], Fred Winston[2]
[1]Department of Biomedical Informatics, [2]Department of Genetics, Harvard Medical School

Abstract: Spt5 is the only transcription elongation factor conserved in all three domains of life, but its molecular mechanisms are not yet thoroughly studied genomically. From an inducible depletion strain, we sequenced nascent transcripts (NET-seq), mature mRNA (RNA-seq), and RNA polymerase II-associated chromatin (ChIP-seq) to elucidate general effects of Spt5 on transcription. These show an increase in 5' CDS antisense transcription by RNA-seq and a general accumulation of RNA Pol II at the 5' ends of genes by exogenous spike-in-normalized ChIP-seq. We are currently analyzing NET-seq to determine: 1) whether novel antisense transcripts are resulting from new transcription or aberrant decay, 2) whether antisense transcript accumulation affects sense transcription, and 3) how Spt5 affects RNA Pol II pausing positions and magnitude. These results will provide insight into how Spt5 functions to facilitate RNA Pol II elongation in diverse organisms.

Genotype to Phenotype Relationships in Autism Spectrum Disorders

Authors: Jonathan Chang, Columbia University, Sarah R Gilman, Columbia University, Andrew H Chiang, Columbia University, Stephan J Sanders, UCSF & Dennis Vitkup, Columbia University

Abstract: Autism spectrum disorders (ASDs) are characterized by phenotypic and genetic heterogeneity. Our analysis of functional networks perturbed in ASD suggests that both truncating and nontruncating *de novo* mutations contribute to autism, with a bias against truncating mutations in early embryonic development. We find that functional mutations are preferentially observed in genes likely to be haploinsufficient. Multiple cell types and brain areas are affected, but the impact of ASD mutations appears to be strongest in cortical interneurons, pyramidal neurons and the medium spiny neurons of the striatum, implicating cortical and corticostriatal brain circuits. In females, truncating ASD mutations on average affect genes with 50–100% higher brain expression than in males. Our results also suggest that truncating *de novo* mutations play a smaller role in the etiology of high-functioning ASD cases. Overall, we find that stronger functional insults usually lead to more severe intellectual, social and behavioral ASD phenotypes.

Longitudinal Metabolome Wide Association Study of Cognitive Decline in Healthy Adults

Authors: Burcu F Darst, Ronald Gangnon, Joshua J Coon, Sterling C Johnson, Corinne D Engelman, University of Wisconsin, Madison

Abstract: Despite being the sixth leading cause of death in the US and its steadily increasing prevalence, little is known about the cause of late onset Alzheimer's disease (AD). Several metabolomics studies of AD were recently published, but the examination of metabolomic profiles prior to AD diagnosis is important to distinguish predictive versus diagnostic profiles, since the disease process itself influences metabolites. Using longitudinal plasma samples from the Wisconsin Registry for Alzheimer's Prevention (WRAP), a cohort study enrolling initially asymptomatic participants enriched with a parental history of AD, metabolomic profiles were quantified using mass spectrometry for 28 participants showing cognitive decline and 55 matched cognitively stable participants. A metabolome-wide association study (MWAS) was performed using conditional random effects logistic regression models with strata for gender and age, which participants were matched on. Of the 615 metabolites tested, 20 met statistical significance after adjusting for multiple testing, 10 of which were amino acids that all showed decreased levels in cases. This aligns with recent research suggesting that a lack of essential amino acids could lead to neuronal death in the hippocampus, a hallmark characteristic of AD. Further research is necessary to determine the role amino acids play in the onset of AD.

Predicting Required Diagnostic Tests from Patient Triage Data

Authors: Haley Hunter-Zinck, Stephan Gaehde, Department of Veterans Affairs, VA Boston Healthcare System

Abstract: Emergency departments are continuously working to increase patient satisfaction and reduce length of stay. Laboratory tests or imaging procedures are often ordered only after evaluation of the patient by a provider. Accurate prediction of complaint specific diagnostic testing has the potential to allow testing to be initiated immediately after patient triage and reduce length of stay. We investigated whether we could predict patients' ordered tests from data collected at triage. Using the National Hospital Ambulatory Medical Care Survey from the Centers for Disease Control and Prevention, we extracted from approximately 20,000 patient visits information that would be available upon triage or from previous medical history as well as procedures ordered during the visit. Using a multivariate machine learning framework, we assessed prediction performance and the relative importance of each data feature.

Prediction performance varied greatly depending on the test but mostly due to its frequency of administration. For example, we predicted the order of a complete blood count, administered in 44% of sampled visits, with 78% accuracy. Several variables were important for prediction across all procedures, including arrival by ambulance, acuity score, age, and injury. Overall, we have adequate information in triage data alone to predict relatively common test ordering.

Classification of Literature Derived Drug Side Effect Relationships

Authors: Justin Mower[1,2], Devika Subramanian[3], Trevor Cohen[1,2]
[1] Baylor College of Medicine, Houston, TX, [2] University of Texas Health Science Center at Houston, Houston, TX, [3] Rice University, Houston, TX

Abstract: Adverse drug events (ADEs) are one of the leading causes of preventable patient morbidity and mortality. An important aspect of post-marketing drug surveillance involves identifying potential side-effects utilizing ADE reporting systems and/or Electronic Health Records. Due to the inherent noise of these data, identified drug/ADE associations must be manually reviewed by domain experts – a human-intensive process that scales poorly with large numbers of possibly dangerous associations and rapid growth of biomedical literature.

Consequently, recent work has employed scalable Literature Based Discovery methods, which exploit implicit relationships between biomedical entities within the literature to assist in identifying plausible drug/ADE connections. We extend this work by evaluating machine learning classifiers applied to high-dimensional vector representations of relationships extracted from the literature by the SemRep Natural Language Processing system, as a means to identify true drug/ADE connections. Evaluating against a manually curated reference standard, we show that applying a classifier to such representations improves performance over previous approaches. These trained systems are able to reproduce outcomes of the extensive manual literature review process used to create the reference standard, paving the way for assisted, automated review as an integral component of the pharmacovigilance process.

This research was funded by a training fellowship from the Gulf Coast Consortia, on the Training Program in Biomedical Informatics, National Library of Medicine (NLM) T15LM007093, PD – Lydia E. Kavraki.

Assessing the Potential Risk in Drug Prescriptions During Pregnancy

Authors: Ferdinand Dhombres, Vojtech Huser, Olivier Bodenreider, National Library of Medicine

Abstract: Background: Eighty percent of the pregnant women in the US have at least one drug prescription during pregnancy. In 2015 the FDA introduced new drug labeling regulations, with narrative summaries describing the risk and supporting evidence. **Objectives**: To assess the potential risk in drug prescriptions during pregnancy, with respect to the new FDA standard. **Methods**: As a proxy for the FDA standard, we used narrative recommendations from a reference textbook (Briggs, 10th ed. 2015). We analyzed claims data of 159.7M patients from 2003 to 2014. We identified pregnant women by procedure codes for delivery and extracted prescriptions 270 days before delivery. We used the RxNorm API to relate drugs from claims data to the reference. **Results**: Of the 15,815,624 systemic drugs prescribed to 3,741,743 pregnant women, 93% were covered by the reference. The distribution among 6 broad categories was: “compatible with pregnancy” or “probably compatible” (41.2%), “low risk” (16.2%), “moderate risk” (39.3%), and “high risk” or contraindicated (3.29%). Interestingly, a majority of the risk assessment was supported by evidence from human data. **Conclusions**: This investigation demonstrates the feasibility of assessing the potential risk in drug prescriptions during pregnancy, with respect to the new FDA standard, as well as stronger evidence.

Uncertainty Quantification (UQ) in Breast and Ovarian Cancer Risk Prediction Based on Self-Reported Family History

Authors: Lance Pflieger and Julio C Facelli, Department of Biomedical Informatics, University of Utah

Abstract: Risk prediction models, such as BRCAPRO, BOADICEA and Claus, have been developed in order to identify patients' risk of developing Hereditary Breast and Ovarian Cancer. These models assume that patient family health history is accurate and complete; however, family history information collected in a typical clinical setting is known to be imprecise. Using UQ methodologies, we show substantial uncertainty in risk classifications. For our analysis, we generated binomial distributions using family history accuracies found in the literature. These distributions were used in Monte Carlo simulation to reclassify the lifetime risk of a known pedigree into risk categories defined by the American Cancer Society. We found, on average, that up to 55% of high-risk pedigrees are misclassified into lower risk categories, with large disparities between best- and worst- case accuracy scenarios. Risk was frequently misclassified into a lower risk category as self-reported specificities are generally higher than sensitives. Our work implies that; i) UQ of the risk prediction needs to be considered when recommending a course of action; ii) better family history collection tools are needed to decrease uncertainty. This study provides a generalizable method for UQ that can applied to other biomedical fields that use predictive modeling.

Performance Drift in Clinical Prediction Across Modeling Methodologies

Authors: Sharon E Davis, Thomas A Lasko, Guanhua Chen, and Michael E Matheny, Vanderbilt University

Abstract: Integrating prediction models into real-time electronic health record decision support can enhance patient and provider decision-making. However, model accuracy can degrade over time as clinical practice and patient populations change, limiting the utility and impact of such models. We explore whether and how modeling methodologies exacerbate or alleviate performance drift by comparing temporal performance of models developed using common statistical and machine learning techniques. We modeled acute kidney injury among hospitalized patients in a national dataset of admissions to Veterans Affairs facilities (n=1,841,951). Admissions in 2003 served as the development cohort, and we assessed performance within 3-month quarters in 2004-2012. Across all models, discrimination was maintained and calibration declined during validation years 1 and 3. The event rate and case mix drifted over time, while predictor-outcome associations did not. We hypothesize that settings with pronounced association drift may lead to differential calibration drift across models and are implementing parallel analysis modeling hospital mortality and readmission to assess performance in cohorts affected by different combinations of event rate, case mix, and association drift. Understanding methods-based differences in performance drift may inform implementation strategies balancing the need to maintain acceptable levels of calibration and efficient use of analytic resources.

Sample-Specific Sparsity Adjustment Improves Differential Abundance Analysis of 16S rRNA Data

Authors: Liyang Diao[1], Glen Satten[2], Hongyu Zhao[3]
[1]Yale University, Department of Medical Informatics, [2]Centers for Disease Control and Prevention, [3]Yale University School of Public Health, Department of Biostatistics

Abstract: The analysis of microbiome data presents many statistical challenges, especially when the data are very sparse. Although various methods have been proposed to normalize data and address data sparsity, their performance is less than satisfactory. While adjusting counts with a simple pseudocount is a relatively common practice, its effects have not been studied in highly sparse data, where they might affect downstream results the most.

We propose two methods to adjust highly sparse data, and compare the performance of these against fixed pseudocount adjustments, specifically focusing on how downstream results are affected by the adjustments combined with various library size normalization methods. We find that our proposed sample-specific adjustment methods can outperform the pseudocount method in both simulated and experimental data sets, improving the ability of researchers to find true differentially abundant bacteria in 16S rRNA data.

Predicting Drug Response Curves in a Large Cancer Cell Line Screen

Authors: Nathan H Lazar, Mehmet Gonen, Shannon McWeeney, Adam Margolin, Kemal Sonmez, Oregon Health & Science University

Abstract: Precision oncology aims to improve cancer patient outcomes by tailoring treatment to an individual patient's tumor. In order to find genetic markers that predict response, several large cancer cell line (CCL) screens have been performed measuring the growth of CCLs when treated with a panel of drugs at varying doses. The current computational tools used in this area reduce these data to a single value indicating response for each CCL/drug combination. This simplification eliminates a large amount of the experimental data, cannot produce measures of uncertainty and consequently shows poor agreement across studies.

My method uses a three-dimensional tensor factorization framework to predict the full dose-response curve for each CCL/drug combination. Mutation, copy number and expression data for CCLs as well as target and structural features for drugs are used as predictors and parameters are estimated using Bayesian variational approximation. When applied to the largest data set of this type (907 cell lines, 545 drugs and 16 doses) the method can accurately predict responses for CCLs and drugs that are not included in the training set. Additionally, by using sparsity-inducing priors the model can highlight relationships between the CCL genomics and drug features that govern response.

Aggressive Glioblastoma Phenotype Evolves Over Decade-Long Growing Phase

Authors: Daniel I S Rosenbloom, Jiguang Wang, Erik Ladewig, Sakellarios Zairis, Raul Rabadan, Columbia University Medical Center

Abstract: Longitudinal studies of tumor genomics have revealed that tumor evolution rarely follows a linear order of mutation accrual. Instead, lesions observed at later timepoints can lose mutations relative to earlier timepoints, suggesting that these later lesions are evolutionary "throwbacks" that diverged from an initial clone years before diagnosis. We developed an evolutionary model to quantify this process and estimate timing of events in tumorigenesis. Applying our model to whole-exome sequences of 92 glioblastoma patients, we found that half (45/92) exhibit genetically distinct diagnosis and relapse samples, with no shared subclonal mutations. Genetic substitution rates among these patients were remarkably consistent, with a median [interquartile range] of 0.028 [0.018 – 0.041] substitutions per megabase-year. Most strikingly, the common ancestor of diagnosis and relapse samples was estimated to have preceded diagnosis by over a decade in most patients (median 12.6 years, IQR 7.2 – 22.6 years). This long divergence time, coupled with mutational patterns observed in *EGFR*, *TP53*, *PDGFRA*, and other known driver genes, suggests that accumulation of driver alterations in glioblastoma occurs over a decade(s)-long growing phase. This phase results in a diverse population, each clone capable of experiencing a unique set of genetically driven expansions.

Unsupervised Deep Learning Reveals Prognostically Relevant Subtypes of Glioblastoma

Authors: Jonathan D Young, Chunhui Cai, Xinghua Lu, University of Pittsburgh

Abstract: Understanding the cellular signal transduction pathways that drive cells to become cancerous is fundamental to developing personalized cancer therapies that decrease the morbidity and mortality of cancer. The purpose of this study was to develop an unsupervised deep learning model for finding meaningful, lower-dimensional representations of cancer gene expression data. Ultimately, we hope to use these representations to reveal hierarchical relationships (pathways) involved in cancer pathogenesis.

We downloaded 7,528 gene expression samples (each with 15,404 features) across 17 different cancer types from TCGA. We developed a python deep learning library, which included an unsupervised implementation of a Stacked Restricted Boltzmann Machine (SRBM) – Deep Autoencoder (DA).

Extensive model selection identified a promising hidden layer architecture for this dataset. Logistic regression to predict the pathological N-stage of the samples, using the final hidden layer representations as input, performed better than a proportionally random or tissue-type based classifier. Consensus clustering of the low-dimensional representations allowed for more robust clustering than clustering the high-dimensional input data. Consensus clustering of glioblastoma samples across all models identified 6 clusters with differential prognosis. Numerous novel and previously reported glioblastoma subtype-specific genes were found to be significantly correlated with each glioblastoma subtype.

An SRBM-DA deep learning model can be trained to represent meaningful abstractions of cancer gene expression data that provide novel insight into patient survival. Ultimately, deep learning and consensus clustering revealed a subclass of the proneural glioblastoma subtype that was enriched with G-CIMP phenotype samples and demonstrated improved prognosis.

Computational Studies of Protein-Protein Interface Mutations

Authors: Jennifer C Gaines, Corey S O'Hern, and Lynne Regan, Yale University

Abstract: Computational methods are invaluable for assessing the significance of patient DNA variants uncovered in clinical DNA sequencing. Despite major advances, current approaches have found limited success in predicting the change in binding due to mutations at protein-protein interfaces. Here, we implement a hard-sphere model for amino acid structure to study natural and designed protein-protein interfaces. We show that a hard-sphere model of amino acids can recapitulate the side chain dihedral angle distributions for amino acids at natural protein-protein interfaces. In addition, we calculate the packing fraction in naturally occurring interfaces and find that it is comparable to dense random packing in protein cores. We then study the effects of mutations at protein-protein interfaces using a dataset of experimentally studied interface mutations. Our model will enable the prediction of the change in binding energy due to mutations at protein-protein interfaces, many of which are involved in disease onset and progression.

Modeling of the Minimally Gained Significant Region of Trisomy 12 in Chronic Lymphocytic Leukemia

Authors: Zachary Abrams[1], Lynne Abruzzo[2], Kevin Coombes[1], Philip Payne[1]
[1]Department of Biomedical Informatics; [2]Department of Pathology, The Ohio State University

Abstract: Chromosomal abnormalities, gains and losses, are among the strongest independent predictors of rapid disease progression and inferior survival in chronic lymphocytic leukemia (CLL). One common CLL cytogenetic aberration is trisomy 12 (tr12), with the gaining of an additional copy of chromosome 12 (c12). This aberration is difficult to model genetically so the underlying genetic drivers in tr12 CLL cases are unknown.

We utilized a lab-developed karyotype parsing and modeling system, the loss-gain-fusion model, which transforms text-based karyotype data into a binary vector for large-scale analysis. We observed 776 CLL patients' karyotypes to determine if there are differentially gained regions on c12.

We counted gains by breaking c12 into individual cytogenetic bands, then measuring if there were particular sub-bands with higher gains higher. We identified band 12q24 as the most gained region on c12 (gained in 22.8% of the population) compared to the rest of c12 (gained in 21.8% of patients). This suggests 12q24 may be the minimally required c12 gain to drive CLL progression.

In 20 cases where the only cytogenetic aberration was tr12 we looked at the mRNA expression profile and mapped c12's location on each RNA transcript. We then measured if 12q24's protein coding genes were differentially overexpressed compared to other c12 regions. Thus we identified genes that are overexpressed in tr12 that potentially isolate the minimally gained region on c12 related to CLL progression.

A Bioinformatics Approach to Identify Novel Drugs Against Liver Cancer

Authors: Tasneem Motiwala[1], Kelly Regan[1], Ryan Reyes[2], Samson T Jacob[2], Philip R O Payne[1]
[1]Biomedical Informatics, The Ohio State University, Columbus, Ohio, [2]Molecular Virology, Immunology and Medical Genetics, The Ohio State University, Columbus, Ohio

Abstract: The high cost and relative inefficiency of traditional drug discovery approaches have led to a growing interest in drug repositioning. By identifying new indications for existing drugs, drug repurposing offers promise in reducing cost, decreasing drug development timeframe and improving success rates in the clinic. Further, it is an important advancement for diseases like liver cancer that that do not respond well to standard therapy and are in urgent need for effective therapy. Here, through a connectivity-mapping approach, we identified novel drugs for use as first-line therapy in the treatment of hepatocellular carcinoma (HCC) or following progression on sorafenib. Connectivity mapping uses pattern-matching algorithms to compare genome-wide gene expression changes related to biological states of interest: e.g. tumor vs. normal, or drug-resistant vs. sensitive cells against a database of gene expression signatures of various cell lines with drug or gene perturbations. Using this approach, we have identified several drugs that could potentially reverse the gene expression signature of primary HCC and/or sorafenib resistance. Two of the drug hypotheses tested in *in vitro* growth inhibition and colony formation assays validate the specificity of the prediction. Currently, work is underway to explore the mechanisms of the therapeutic effects of these drugs.

Signatures of Accelerated Somatic Evolution on a Genome-wide Scale

Authors: Kyle S Smith, Debashis Gosh, University of Colorado, Anschutz Medical Campus

Abstract: Using a computational method called SASE-hunter we identified a novel signature of accelerated somatic evolution (SASE) marked by a significant excess of somatic mutations localized in a genomic locus, and prioritized those loci that carried the signature in multiple cancer patients. Detection of clinically relevant signatures of somatic evolution in the promoters of known cancer genes in lymphoma raised testable hypotheses whether SASE could be detected in other cancer types as well, and whether these signatures could be detected in non-coding regions outside gene promoters. The current SASE-hunter method is insufficient to meet the need, and a genome-wide assessment requires development of a novel algorithm, which is more advanced than the original SASE-hunter and has sufficient statistical power to detect SASEs at a genome-wide scale. SASE-mapper is a powerful tool for the identification of SASEs on a genome-wide scale. In addition to those signatures of accelerated somatic evolution previously discovered by SASE-hunter, SASE-mapper identifies many regions in the non-coding regions of the genome outside of promoters associated with alterations in gene expression and clinical outcomes. SASE-mapper is written in Python 2.7 and available at http://github.com/kylessmith/SASE-mapper.

Identification and Validation of CNVs using WGS Data from 274 Individuals

Authors: David Jakubosky, Christopher DeBoever, Angelo Arias, Hiroko Matsui, Naoki Nariai, Agnieszka D'Antonio-Chronowska, He Li, Kelly A Frazer, University of California, San Diego

Abstract: Copy number variants (CNVs) are an important source of inter---individual genetic variation and contribute to quantitative traits and complex diseases. Algorithms utilizing discordant and split read pair information are used to identify smaller CNVs (50bp---3kb) and those using read depth discover larger CNVs (>=2Kb). Thus, a combination of approaches must be used for CNV discovery, adding complexity to obtaining a complete set of CNVs and data quality control. Here we use high read--- depth (40X) whole genome sequence (WGS) data to call CNVs in 274 individuals, of which 195 are in families (including 30 trios and 25 sets of monozygotic twins) and 79 are unrelated to anyone else in the collection. We found 16013 CNVs, with a minor allele frequency > 1% and ranging in length from 50bp to 209kb (median length = 3049bp). Based on segregation analysis and concordance between twins we estimate that ~80% of the multi---allelic CNVs and ~99% of the biallelic CNVs are valid. Using transcriptome data generated from induced pluripotent stem cells derived from 215 of these individuals, we found 422 genes with significant CNV associations, including 180 genes with CNV lead variants. We demonstrate that high quality CNVs can be called using high read---depth WGS data.

Computing Geographical Access to Hospitals in Two Countries

Authors: Fabrício S P Kury, Raymonde C Uy, Jessica Faruque, Paul Fontelo
Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health

Abstract: Geographical access to hospitals, here defined as the time it takes to drive a car from a person's residence to the nearest hospital, has controversial association with healthcare utilization and outcomes. In this study we demonstrate how to use hospital data, Census data, and modern online-based Geographical Information System (GIS) APIs to compute, with high precision, the percentage of the population that has geographical access to hospitals in two countries: USA and Brazil. We review the availability of data for each country, the magnitude of the computation task, and how we used cloud computing to deliver results in feasible time. We analyze the sociodemographic and economic characteristics of the served and underserved populations under several time thresholds, filter hospitals according to the types of services they provide, and correlate the size of population covered with the volume of utilization of each hospital. We demonstrate that the vast majority of the population resides very near at least one hospital, that this concentration is sharper in Brazil, and how the numbers change after filtering hospitals. We display highly detailed zoom-able maps and demonstrate how misleading their appearance might be. We conclude by reviewing prominent limitations for these analyses in the case of each country.

Bursting the Information Bubble: Designing Inpatient-Centered Technology Beyond the Hospital Room

Authors: Andrew D Miller, Ari Pollack, Wanda Pratt, University of Washington

Abstract: Although hospital care is carefully documented and electronically available, few information systems exist for patients and families to use while inpatient. We present findings from three participatory design sessions conducted with 13 former patients, their parents, and clinicians from a large children's hospital. Participants discussed challenges they faced getting information while in the hospital, and then designed possible technological solutions. Participants created 9 designs aimed at extending parents' access to and involvement in patients' care.

Participants' designs showed how information technology can allow parents and children to disseminate information from within the hospital room, access information from the hospital room remotely, establish collaborative communication with the clinical care team, and learn about their child's care throughout the hospital stay. For example, two child participants envisioned a communicator watch that their parents would use to talk with clinicians remotely. A parent/clinician team proposed a shared calendar for parents and clinicians to use throughout the stay. Several parent-designed solutions focused on simplifying intake, reducing repetitive questions and allowing parents and children to add information proactively.

These designs show that patients and caregivers can be more than recipients of health information; they can produce, aggregate, and learn information throughout a hospital stay.

User-Centered Design and Evaluation of RxMAGIC: A System for Prescription Management and General Inventory Control for Low-Resource Settings

Authors: Arielle M Fisher, Lauren Jonkman, Gerald P Douglas, University of Pittsburgh

Abstract: The availability of healthcare services in low-resource settings is limited due to health, economic, and education disparities in underserved populations. Free clinics are critical in providing primary care and pharmaceutical services to these patients, however they represent an understudied work environment in healthcare. In addition to service-related challenges, such as difficulty in obtaining essential medicines, free clinics are burdened with distinctive organizational challenges.

Ensuring an uninterrupted drug supply is essential to providing healthcare in these settings. Accurate information on current stock counts is necessary to minimize stockouts and wastage due to expiry. Informatics tools have tremendous potential to assist healthcare workers and enhance process efficiency if designed to support user workflow.

We developed a system for Prescription Management and General Inventory Control (RxMAGIC) at the Birmingham Free Clinic (BFC) in Pittsburgh, PA, a walk-in clinic that serves medically vulnerable populations. A mixed-methods approach was employed to identify and quantify process inefficiencies in the dispensary. RxMAGIC is a modular, problem-driven solution designed to mitigate workflow challenges and improve pharmacist efficiency by streamlining the dispensing process and improving inventory control. Although RxMAGIC was developed in the context of the BFC, we believe it may alleviate similar medication management challenges in developing countries.

Clinical Decision Support Anomaly Pathways

Authors: Steven Z Kassakian, David A Dorr, Oregon Health and Science University

Abstract: Clinical decision support (CDS) tools are designed to aid decision making with the ultimate goal of improving health outcomes. CDS is a central part of electronic health record (EHR) systems and has been shown to improve a multitude of outcomes. However, in some clinical practice situations, CDS many not improve outcomes and may have detrimental effects on decision making through the increasingly recognized phenomena of alert fatigue. In many situations, the proper functioning of CDS tools is essential to providing appropriate care and their dysfunction may result in poor care and in some cases harm to patients. The tools are usually built around a complex series of logic based on variables in the EHR. Little is known regarding how to appropriately monitor and detect when CDS tools are not functioning as intended. The field of anomaly detection is focused on finding patterns in data which do not conform to historical or predicted patterns. By applying methods of anomaly detection from other domains, we are exploring the ability to detect broken CDS tools. Our preliminary results have discovered multiple CDS tools that are no longer functioning as designed. Most importantly, we are elucidating the pathways through which these CDS tools fail.

Medical Entity Recognition: a Meta-Learning Approach with Selective Data Augmentation

Authors: Asma Ben Abacha and Dina Demner-Fushman, National Library of Medicine

Abstract: With the increasing number of annotated corpora for supervised medical entity recognition (MER), it becomes interesting to study the combination and augmentation of these corpora for the same annotation task. Combing annotated corpora such as clinical texts or scientific articles is a challenging task since it generally drops the classification performance for supervised systems. We study the combination of different corpora for MER by using a meta-learning classifier that combines the results of individual conditional random fields (CRF) models trained on different corpora. We propose selective data augmentation approaches and compare them with several meta-learning algorithms and baselines. We evaluate our approach using four sub-classifiers trained on four heterogeneous corpora: i2b2, SemEval, Berkeley and NCBI. We show that despite the high disagreements between the individual CRF models on the four test corpora, our selective data augmentation approach improves performance on all test corpora and outperforms the simple combination of individual corpora. Our results confirm that the agreement between label predictions of the pairwise models is an effective metric in selecting relevant sources for data augmentation when used with reliability indicators such as the class balance of each corpus.

Untangling the Structure of High-Throughput Sequencing Data with veRitas

Authors: David M Moskowitz, William J Greenleaf, Stanford University

Abstract: High-throughput sequencing offers unprecedented power in describing genomic and epigenomic changes in biological processes, but effective interpretation requires accounting for variance associated with batches, RNA degradation, and other technical details. In this talk, I will introduce veRitas, a method combining principal component analysis with feature selection to elucidate confounding and technical artifacts. This approach additionally assesses differential expression without parametric assumptions, in contrast to existing methods, which are specific to RNA-seq.

Modeling Neutral Evolution at Small Scales

Authors: Aaron Wacholder, David D Pollock, University of Colorado, Anschutz Medical Campus

Abstract: Developing precise models of neutral genomic evolution will enable sensitive detection of selection in the genome, and thus of function. A large body of research demonstrates that, at the megabase scale, neutral substitution rates are strongly dependent on genomic context, such as the recombination rate, replication timing, and chromatin structure. However, a large fraction of regional substitution rate variation occurs at much smaller scales, and the nature of this variation is largely unknown. Investigation of local substitution rate variation has been hindered because low substitution counts in small regions prevents accurate direct estimation of substitution rates.

We developed a model of substitution rates for each substitution type in 1000 bp windows across the genome, accounting for changes over time and effects at different spatial scales. Applying this model to a whole-genome alignment of the great apes, we find strong effects at all spatial scales that differ across time and among substitution types. We identify a major change in the kilobase-scale substitution process between the human-gorilla and human-chimpanzee divergence, while larger-scale substitution processes have remained relatively stable. These findings provide the starting point for a precise time and space dependent model of neutral substitution rates.

EHR-Wide GxE Study using Smoking Information Extracted from Clinical Notes

Authors: Travis J Osterman, Lisa Bastarache, Wei-Qi Wei, Jonathan D Mosley, Joshua C Denny, Vanderbilt University

Abstract: Genotype by environment interaction (GxE) studies provide a method to assess whether genomic and environmental effects are additive or whether there is an additional interaction. We describe here a GxE study to investigate associations between tobacco exposure and genetic risk across 105 diseases.

Patients were identified from Vanderbilt University Medical Center's (VUMC) de-identified DNA biobank (BioVU) which is linked to electronic health record data. Approximately 15,000 individuals with exome array data were selected for this analysis. Tobacco exposure was ascertained by a novel natural language processing algorithm. Phenotypes were determined by International Classification of Disease 9 (ICD-9) codes.

We analyzed 1750 SNP-phenotype pairs previously reported in the NHGRI catalog. To test for smoking x SNP interaction, we used a logistic regression with age, gender, pack years, SNP, and pack years x SNP terms. We calculated p-values for the smoking x SNP interaction term, controlling for the remaining covariates.

Smoking was strongly associated with a number of expected phenotypes such as lung cancer. The SNP x smoking interaction p-value was <0.05 for 57 SNP-phenotype pairs. Evidence of interaction was seen in several cancers, including lung, breast, and prostate cancer. Three cardiovascular phenotypes demonstrated interaction: Ischemic heart disease, hypertension, and aortic aneurysm.

High-Throughput Machine Learning from Electronic Health Records

Authors: Ross Kleiman[1,2,†], Paul Bennett[1,2,†], Peggy Peissig[3], Zhaobin Kuang[1], James Linneman[3], Scott Hebbring[3], Michael Caldwell[3], David Page[1,2]
[1]Department of Computer Sciences, University of Wisconsin, Madison, [2]Computation and Informatics in Biology and Medicine, [3]Marshfield Clinic, Marshfield, WI; [†] Co-First Author

Abstract: The use of Electronic Health Record (EHR) systems has increased dramatically in recent years. This vast digitization of medical data allows for new ways to predict diseases that were not possible with paper charts. While prior work has focused on predicting individual diseases, our research builds thousands of models to predict nearly every diagnosis (ICD-9 code) a patient could receive. This high-throughput machine learning approach yields inference on the health landscape of both individual patients and patient populations. Integral in our approach is the use of a dynamic control matching scheme that, for each diagnosis, automatically selects appropriate case and control patients using minimal hand tuning. Across the nearly 4,000 models, we observe a mean AUC of 0.8026±0.0619 predicting 1 month prior to diagnosis, and a mean AUC of 0.7585±0.0631 predicting 6 months prior to diagnosis. Furthermore, we break down our results across 15 major disease categories including pregnancy complications and diseases of the circulatory system. This work opens a potential pathway to pan-diagnostic decision support. Instead of only targeting a small number of well-understood diseases, this research shows machine learning techniques can be used to help predict the broad spectrum of diagnoses a patient may receive.

Comparison of Variant Annotation Tool Terminology using the Sequence Ontology

Authors: Nicole Ruiz-Schultz, Barry Moore, Shawn Rynearson, Karen Eilbeck, University of Utah

Abstract: Analysis of next-generation sequencing (NGS) data involves multiple steps including base calling, quality assessment, read alignment, variant calling, variant annotation, and variant prioritization. Variant annotation is the step in sequence data analysis of determining the effect of a sequence variant with regards to the features of a reference sequence. Many open source and commercial tools are available to perform this step, with differing sets of effects annotated and differing terminology used. These differences can make comparing variant annotations from different tools challenging and in some cases, a one-to-one comparison cannot be made. The goal of this project was to present a comparison of terms used by variant annotation tools, utilizing the Sequence Ontology to map between terms.

Terms from VAAST, VEP, ANNOVAR, Jannovar, Seattleseq, SnpEff and VAT were mapped to the SO if not already using the terminology. Prior to the start of this project, VAAST and VEP used SO terms. SnpEff and Jannovar adopted SO usage during the project. We will present the scope of annotation for each tool, the concordance and discordance between the terms. SO is increasingly used to standardize terms from variant annotation tools currently available so results can be easily compared.

Constructing a Biomedical Relationship Database from Literature using DeepDive

Authors: Emily K Mallory, Ce Zhang, Christopher Re, Russ B Altman, Stanford University

Abstract: A complete repository of biomedical relationships is key for understanding cellular processes, human disease and drug response. After decades of experimental research, the majority of the discovered biomedical relationships exist solely in textual form in the literature. While curated databases have experts manually annotate relevant relationships or interactions from text, these databases struggle to keep up with the exponential growth of the literature. DeepDive is a trained system for extracting information from a variety of sources, including text. In this work, we developed multiple entity and relationship application tasks to extract biomedical relationships from full text articles. Each relationship extractor identified candidate relations using co-occurring entities within an input sentence. Using a set of generic feature patterns, DeepDive computed a probability that an individual candidate relation was a true relationship based on the sentence. For extracting gene-gene relationships, our system achieved 76% precision and 49% recall in extracting direct and indirect interactions. For randomly curated extractions, the system achieved between 62% and 83% precision based on direct or indirect interactions, as well as sentence-level and document-level precision. In addition, we developed extractors for gene-disease and gene-drug relationships. This work represents the first application of DeepDive to the biomedical domain.

Poster #101: Movement Disorders Journal: Testing an App to Track Parkinson's Symptoms

Authors: Jeff Day, Jeff Baldwin, Omar Ahmad, Mark Hallett, John Harrington, Anne Altemus, and Codrin Lungu, National Library of Medicine

Abstract: Neurologists use patient histories to assess the symptom patterns and severity of Parkinson's Disease in order to adjust medications. However, patient recall can be imprecise with only two or three yearly visits. We have designed an iPad app to help patients track their symptoms and medications, and we will test compliance in the recording of data between the app and standardized paper forms. Twelve Parkinson's patients scheduled for the placement of Deep Brain Stimulation (DBS) were recruited for this study, and randomized into two groups: a group of six patients who will receive an iPad, and another group of six patients who will receive paper forms to record their data. Each patient will begin the study after DBS placement and be followed for three months. We will analyze the frequency of patient-recorded data as a test for compliance, and use surveys to evaluate patient satisfaction for both groups. Surveys and patient interviews will provide insight into user experience with the app, which can inform design strategies for mobile technology built for movement disorder patients.

Poster #102: Design of a Subscription-Based Laboratory Result Notification System

Authors: Benjamin H Slovis, Hojjat Salmasian, Gilad Kuperman, David K Vawdrey, Columbia University

Abstract: **Background:** The delayed review of laboratory results is potentially harmful. Established processes (e.g. phone-calls) provide notification of critical laboratory values, however evidence suggests that physician awareness of non-critical and normal results affect clinical decision-making. Many HIT tools have demonstrated improved physician response time to laboratory results, yet continued utilization and enhancements are rare, with an overall lack of provider control. Specifically, few studies have documented subscription-based notifications. **Objective:** We propose a tool to provide physicians with near-real-time notification of laboratory results through text-page and email, via subscription at the time of order-entry. Needs-assessments will include evaluation of the extent to which current processes delay hospital care, resulting in clinician dissatisfaction. Preferred methods of notification and notification utility for specific laboratory tests will also be assessed. **Methods:** A physician-observer will document current processes, and promote dialog with clinical house-staff at an urban academic hospital regarding barriers to appropriate results-review. A survey will be distributed to determine the perceived usefulness of subscribed laboratory notifications. **Significance:** Our long-term objective is to develop a subscription-based notification system to reduce time between available results and physician awareness. We expect physician interest and encouraging survey results. Such a tool has the capacity to potentially improve the quality of clinical care.

Poster #103: Medication Use Among Veterans Across Health Care Systems

Authors: Khoa A Nguyen, Alan J Zillich, Susan Perkins, David Haggstrom,
Dept. of Veterans Affairs Richard L Roudebush VA Medical Center, Indianapolis, IN; Purdue University College of Pharmacy

Abstract: Dual health care system use is becoming a common type of care for most Veterans. The VA is implementing a nationwide health information exchange (HIE) program called the Virtual Lifetime Electronic Record (VLER), which allows providers to access and share patient information among each other. Because there is a lack of information about the use of medications across dual systems of care, the objective of this study is to describe the prevalence of medication dispensing across VA and non-VA health care systems prior to enrollment in VLER.

In this retrospective cohort study, we examined outpatient dispensing during a two-year window prior to VLER enrollment. Data were extracted from the VA Pharmacy Benefits Management system and a regional HIE. Medication source was assessed at the subject level, and categorized as VA source, non-VA source, or both. We then compared the mean number of prescriptions as well as overall and pairwise differences in medication dispensing.

Out of 52,444 Veterans included in our study, 17.4% of subjects (n=9,123) obtained medications outside the VA including prescriptions for antibiotics, antineoplastics, and anticoagulants. Subjects receiving medication from both sources appeared to have more complex medical needs, as reflected by their higher overall mean number of medications.

Poster #104: Designing a Telehealth Training Curriculum using a Telemental Health Model

Authors: Pamela Hoffman, Rhonda Johnston, Cindy Brandt, and Linda Godleski, Department of Veterans Affairs, VA Connecticut Healthcare System; VA Telehealth Service

Abstract: Problem: Telehealth is a well-established modality for treating patients at a distance and improving access to care. Few studies have been published on training in telehealth specialties. **Approach**: We propose a standard curriculum on telehealth, based on a current telemental health training model. Our innovative curriculum follows a strategic outline: Background, evidence base, legal and regulatory concerns, emergency procedures, applications for an encounter, and case simulations. **Outcomes**: This model curriculum has been implemented, live and remotely to over 4800 participants in 2 VA facilities and 3 training programs, with very positive effect. Participant satisfaction is consistently over 80% and learners' impressions of competence invariably increase. **Future steps**: This innovative model is a way to standardize training efforts in telehealth. Virtual and remote training in telehealth will extend access to knowledge and subsequent services to patients nationwide.

Poster #105: A Multi-Axial Based Knowledge Management System for Alerts

Authors: Rajdeep Brar[1], Richard Shiffman[1]
[1] Yale Center for Medical Informatics, New Haven, CT

Abstract: Background: The phenomenon of alert fatigue can have serious negative implications in regard to workflow, user satisfaction, clinical effectiveness, as well as patient safety. Knowledge organization models that can categorize clinical alerts in a comprehensive and useful way for curation and update are needed. **Hypothesis:** A multi-axial based knowledge organization model for alerts can help target areas for quality improvement and patient safety. **Methods**: The 546 alerts in Yale's instance of Epic™ will be manually categorized according to function, IOM quality heading, medical specialty, care setting, and additional groupings with perceived utility. Alert firing and override statistics will be monitored. **Results:** A comprehensive set of alert categories has been identified. Additional categories will be added to the initial set for model enrichment. We plan to improve alert use and override statistics by targeting poorly performing alerts based on category. **Conclusions:** We believe this approach will be useful when maintaining existing clinical alerts and when building new ones. Statistics will be computed on each category, e.g., frequency of firing and action by user, and then used to garner insights into whether certain categories of alerts are performing as expected. We will then use those insights to target alerts for sensitivity/specificity adjustment or retirement.

Poster #106: Improving and Applying Medical High-Throughput Machine Learning

Authors: Paul Bennett[1,2,†], Ross Kleiman[1,2,†], Peggy Peissig[3], Zhaobin Kuang[1], James Linneman[3], Scott Hebbring[3], Michael Caldwell[3], David Page[1,2]

[1]Department of Computer Sciences, University of Wisconsin, Madison, [2]Computation and Informatics in Biology and Medicine, [3]Marshfield Clinic, Marshfield, WI [†] Co-First Author

Abstract: In recent years, many healthcare professionals and researchers have become keenly interested in predicting disease risk using electronic medical record data. Using highly parallelized computing, we built predictive models for nearly every diagnosis (ICD-9 code) a patient could receive. These models achieved a mean AUC of 0.8026±0.0619 predicting diagnoses 1 month in advance and a mean AUC of 0.7585±0.0631 predicting diagnoses 6 months in advance. Given the tremendous breadth of this work, we are presented with many new challenges. Our research helps address the difficult task of appropriately matching cases to controls across thousands of diagnoses with particular emphasis on case-control matching for pregnancy complication prediction. Furthermore, we examine novel applications unique to high-throughput prediction. We perform a simulated prospective study across all diagnoses predicted and then bi-cluster patients and diseases based on the model scores. We also investigate using model scores as a feature set for predicting hospital readmission. Our research represents a new direction in medical machine learning and completes several necessary steps in improving and applying this high-throughput method of diagnosis prediction.

Poster #107: Assessing the Delay in Communication Regarding Digital Inpatient Documentation

Authors: Ross Lordon, Thomas Payne, University of Washington

Abstract: Within the past decade, healthcare records generally have transitioned from paper to digital formats. Unfortunately, this new method is time consuming[1]. A study in 2012 reported physicians were spending 49% of their workday using a computer and 70% of this time was spent performing documentation[2]. An unintended consequence concerns the delay between when patients are seen during rounds and when their encounter note is written and signed by their physician. The encounter note is the central location of critical care information. Within certain popular EHRs, an encounter note is not viewable by others until it is signed. This delay may cause communication errors, delay in care, or other unintended consequences.

We conducted a prospective observational study of physician teams within a county safety net hospital. Physicians recorded the time each patient was seen during rounds. Timestamps documenting when notes were signed in the EHR were obtained from a clinical data repository. The gap in documentation was calculated by determining the difference between these times. 212 patient encounters were analyzed and the average documentation gap was 5.4 hours with a maximum of 17.3 hours. An opportunity exists to improve the digital documentation process, potentially allowing physicians to be more efficient.

1) Cusack CM, Hripcsak G, Bloomrosen M, Rosenbloom ST, Weaver CA, Wright A, Vawdrey DK, Walker J, Mamykina L. The future state of clinical data capture and documentation: a report from AMIA's 2011 Policy Meeting. J Am Med Inform Assoc. 2013 Jan 1;20(1):134- 40. doi: 10.1136/amiajnl-2012-001093. Epub 2012 Sep 8. PubMed PMID: 22962195; PubMed Central PMCID: PMC3555335.

2) Oxentenko AS, Manohar CU, McCoy CP, Bighorse WK, McDonald FS, Kolars JC, Levine JA. Internal medicine residents' computer use in the inpatient setting. J Grad Med Educ.2012 Dec;4(4):529 32. doi:10.4300/JGME-D-12-00026.1.

Poster #108: Quantifying Burden of Treatment in Patients with Breast Cancer

Authors: Alex C Cheng, Mia A Levy, Vanderbilt University

Abstract: Chronic disease decreases a patient's quality of life through the direct effect of illness, as well as the burden of treatment imposed to counteract illness. While burden of illness is well studied, the burden of treatment is not as well understood or monitored. We developed a method to quantify one dimension of the burden of treatment based on patient encounters with the healthcare system. Specifically, we tracked the total time spent in appointments and admissions, waiting time, and travel time to the medical center. We applied this method to a population of stage I-III breast cancer patients at Vanderbilt University Medical Center.

We were able to differentiate burden of treatment for patients with stage I-III cancer in the first 18 months after diagnosis. As hypothesized, stage III patients had the greatest treatment burden, followed by stage II patients and stage I patients. Future work will evaluate the reproducibility and generalizability of this method for quantifying burden of treatment across other clinical settings and chronic diseases. This approach may enable identification of high-risk groups that could benefit from interventions to decrease patient work and improve outcomes.

Poster #109: Evaluating the Use of an Automated Section Identifier for Focused Information Extraction Tasks on a VA Big Data Corpus

Authors: Le-Thuy T Tran, Guy Divita, Marjorie H Carter, Matthew H Samore, Adi V Gundlapalli, University of Utah School of Medicine and VA Salt Lake City Health Care System

Abstract: The Veterans Health Information Systems and Technology Architecture (VistA)/CPRS **(**Computerized Patient Record System) is an electronic medical record of the VA enterprise-wide health information system. The large numbers of clinical notes stored in VistA/CPRS are a valuable information extraction resource for detecting patient care and treatment patterns, risks and outcomes of diseases, or adverse events. For efficiently mining these data, we have developed an automated section identifier based on an ontology of clinical document sections to preprocess the clinical notes for further focused information extraction. The identifier was first trained on a set of 1000 documents and then used to identify a fine level of clinical note sections in a corpus of about one million records derived from VistA. The information from this preprocessing step is stored for future efficient access to a specific content of the notes.

We evaluate the use of our developed automated section identifier for focused information extraction tasks including extracting vital signs data, retrieving patient-reported symptoms, and identifying risk and evidence of homelessness among Veterans.

Poster #110: Automatic Identification of High Impact Articles in PubMed to Support Clinical Decision-Making

Authors: Jiantao Bian[1], Siddhartha Jonnalagadda[2], Gang Luo[1], Guilherme Del Fiol[1]
[1]University of Utah, [2]Northwestern University

Abstract: Objectives: Researchers have been trying to make PubMed more useful for supporting clinicians' decision making. We aim to help clinicians find studies with high clinical impact. **Materials and Methods:** Our overall method is based on machine learning algorithms with a variety of features including Altmetric score (tracks online popularity of scientific work), journal impact factors, study registration in ClinicalTrials.gov, publication in PubMed Central, article age, study sample size, comparative study, citation count, number of comments on PubMed and study quality (according to a state-of-the-art machine learning classifier developed by Kilicoglu et al.). The algorithms were developed and evaluated with a gold standard composed of 502 high impact clinical studies that are referenced in 11 clinical guidelines from various diseases. **Results:** Among Naïve Bayes, support vector machine (SMO), and decision tree (J48) with default parameters in Weka, Naïve Bayes performed best. It outperformed the baseline in terms of top 20 precision (mean =34% vs. 12%), mean average precision (mean = 24% vs. 5%) and mean reciprocal rank (mean = 0.78 vs. 0.18). **Conclusions:** Preliminary results show that the high impact Naïve Bayes classifier using a variety of features is a promising approach to identifying high impact studies for clinical decision support.

Poster #111: Design Thinking in Radiation Oncology

Authors: Adam Rule, Erin Gillespie, Nadir Weibel, Todd Pawlicki, University of California, San Diego

Abstract: Radiation oncologists routinely use weekly chart rounds to check quality of care with other clinicians. However, there is sparse evidence that chart rounds improve patient outcomes. Moreover, recent studies found just 4-12% of treatment plans were modified at typical chart rounds. This low rate has been attributed to limited time for discussing patient cases (just 3 minutes at many practices) and many cases being review after treatment begins.

To redesign chart rounds, we assembled a team of radiation oncologists, physicists, and designers at UC San Diego for two half-day workshops. The participants used design thinking to guide the workshops, which encourages thoroughly defining the problem before brainstorming solutions.

In the first workshop, participants identified four goals of chart rounds (quality assurance, decision support, education, and team building) and identified three areas for redesign. (How might we document and disseminate informal peer review? How might we ensure participants feel time spent on peer review is well spent? How might we facilitate a culture of collaboration, safety, and team building?) During the second workshop, participants brainstormed solutions to these prompts including an email review system that supports more focused and flexible forms of review. This design is currently being prototyped for testing.

Poster #112: Prospective Study of a Kawasaki Disease Natural Language Processing Tool

Authors: Juan D Chaparro, Chu-Nan Hsu, Zach Meyers, Adriana Tremoulet. University of California, San Diego

Abstract: Kawasaki Disease (KD) is a rare pediatric febrile syndrome consisting of prolonged fever and five clinical symptoms. Nearly 20% of children with KD develop coronary artery aneurysms if left untreated. However, diagnosis is often delayed due to lack of a diagnostic test and overlap with other febrile syndromes, thus there is a need for improved diagnostic tools.

KD-NLP is a natural language processing tool to identify patients with high-suspicion for KD using provider notes from the Emergency Department (ED). We recently published the development and testing of this tool using retrospective ED notes from patients with KD and febrile patients. The tool identifies the presence/absence of the five signs of KD in the narrative text and classifies patients on these findings.

We will implement this tool into a live electronic health record system to 1) prospectively determine the sensitivity/specificity of KD-NLP in a low prevalence population and 2) to evaluate the feasibility of KD- NLP in providing clinical decision support in a time frame that can affect medical decision making.

We are integrating the KD-NLP tool into the Epic ASAP module at Rady Children's Hospital San Diego and will begin data collection, but are also considering integration in non-pediatric emergency departments.

Poster #113: Modeling of Hypoplastic Left Heart Syndrome for Improved Decision Support

Authors: Charles Puelz[1], Beatrice Rivière[1], Craig G Rusin[2]
[1] Rice University, Houston, TX, [2] Baylor College of Medicine and Texas Children's Hospital, Houston, TX

Abstract: Babies born with congenital heart defects often require immediate surgery and many hours of critical care in the hospital. Their hemodynamic state pre- and post-surgery is complex, abnormal, and extremely challenging to manage. Indeed, all vital signs may indicate stability and yet the patient falls into unexpected cardiac arrest. Currently, our research focuses on a class of defects generally identified by a severely underdeveloped left ventricle called Hypoplastic Left Heart Syndrome (HLHS).

The purpose of our work is to develop a clinical decision support tool, based on a computational fluid dynamics model of the entire circulatory system, to aid clinicians in providing critical care to HLHS patients. This tool predicts blood pressure and flow waveforms in peripheral arteries and veins, and allows for the incorporation of measured patient data for simulations and model validation. Our goal is for clinicians to use this tool for insight into the complex hemodynamics of HLHS, and in turn to improve the care provided to these patients at the bedside.

This research was funded by a training fellowship from the Gulf Coast Consortia, on the Training Program in Biomedical Informatics, National Library of Medicine (NLM) T15LM007093, PD – Lydia E. Kavraki.

Poster #114: Taxonomic Classification of HIT Hazards Associated with EHR Implementation: Initial and Stabilization Phases

Authors: Paul Varghese, Adam Wright, David Bates, Harvard Medical School

Abstract: Data that describe the nature, magnitude and frequency of these EHR safety concerns remain scarce, with a limited number of studies focused upon mining patient safety incident reporting databases. By using both traditional in-hospital patient safety monitoring system reports and previously unexamined hospital information services customer complaint reports during a large-scale implementation of EHR at an academic medical center, we are in the process of 1) categorize the types of hazards using AHRQ hazard criteria; 2) assessing type and severity of patient harm (actual and potential) in both the initial phase (3 months) and subsequent stabilization phase.

Poster #115: Teamwork Behaviors of Emergency Medical Service Teams in Pediatric Simulations

Authors: Nathan Bahr, Jeanne-Marie Guise, Paul N Gorman, Oregon Health and Science University

Abstract: Teamwork can determine patient outcomes during prehospital care. In this work, we describe behaviors that appear to distinguish high-performing teams from low-performing teams and may contribute to improved outcomes.

Forty Emergency Medical service teams were recruited to participate in 4 pediatric simulations. Simulation performance and outcomes were assessed independently by a domain expert by counting and classifying observed errors and using the Clinical Teamwork Scale (CTS). Teams were classified as high-performing and low performing based on this assessment and selected two for analysis. To identify behaviors, the simulations were recorded, transcribed, and coded according to team communication patterns (speaker-listener interactions), task focus (task relevance of dialog content), and verbal behaviors (apparent purpose of speech act, e.g. query, inform, direction, acknowledge, etc.).

In the high-performing team, the leader called the Person in Charge (PIC), provided other members with situational assessments, clear goals, and directions to reach those goals. In the low-performing team, the PIC exhibited a preference to summarizing the situation and stating their own actions over directing others. We hypothesize that this behavior may be a silent cry for help, in which the PIC becomes lost and needs support from their teammates.

Poster #116: Large-Scale Family Cohorts Linked to Electronic Health Records

Authors: Scott J Hebbring[1, 2], Xiayuan Huang[2], John Mayer[1], Zhan Ye[1], David Page[2], (1) Marshfield Clinic and (2) University of Wisconsin Madison

Abstract: Challenges in population-based genetic research have resulted in a re-awakening of family-based studies. However, significant difficulties arise when identifying the most interesting diseases and families for genetic research. Use of large patient populations linked to an electronic health record (EHR) may alleviate such challenges. Using readily available basic demographic data in an EHR, we identified over 173,368 families including 8,242 families of twins from Marshfield Clinic. With these large cohorts of families all linked to extensive health records, thousands of diseases may be studied simultaneously by phenome-wide approaches. Studies in twins suggest that few diseases are random events and that family relationships are extremely important in predicting disease risk. With our novel phenome-wide methodologies highly translatable to other EHR systems, this study may pave the way for biotechnologically smart EHR systems that integrate family data to generate personalized family histories in real-time for the prediction, prevention, and treatment of many diseases and advancement of "precision medicine." Lastly, this study provides an intriguing perspective for the future of genetic epidemiologic research. Specifically, the future when large patient populations with sequenced genomes are unified by familial relationships in an integrated EHR system.

Poster #201: Predicting Accidental Falls in People Aged 65 Years and Older

Authors: Mark L Homer[1,2], Nathan P Palmer[1,2], Kenneth D Mandl[1,2]
[1]Computational Health Informatics Program, Boston Children's Hospital, [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA

Abstract: More than half a million people over 65 years of age accidentally fall every year in the United States alone. To help tackle the problem, we develop a predictive analytics model based upon machine learning (logistic regression with LASSO) to estimate each individual's unique risk of falling by looking at their past insurance claims. During testing, our predictive model successfully risk stratified people, where those in the highest stratum had greater than 15 times the risk than those in the lowest stratum (34.7% vs. 1.7%). Next steps include better modeling techniques and running a prospective study.

Poster #202: Content-Based fMRI Activation Maps Retrieval

Authors: Alba G Seco de Herrera, L Rodney Long, Sameer Antani, National Library of Medicine

Abstract: Functional Magnetic Resonance Imaging (fMRI) is a powerful tool used in the study of brain function. It can non-invasively detect signal changes of cerebral blood flow in areas of the brain where neuronal activity is varying. Statistical analysis of fMRI data is used to locate brain activity and generate brain activation maps. These maps are used to determine how a task is correlated with particular perceptual or cognitive state that is encoded by active brain regions.

Neuroimaging data sharing is becoming increasingly common. Currently, some efforts have been made to develop fMRI repositories. However, there is a need for content-based (CB-) fMRI retrieval methods that can retrieve studies relevant to a "query" brain activation. One approach is to take into account the full spatial pattern of brain activity to retrieve similar activity maps. This approach could also be extended to support cognitive state-based retrieval.

This work present an approach for CB-fMRI activations maps retrieval which return activation maps that have similar activation patterns to the given one. The proposed method develops a similarity score that matches map activation maps.

Poster #203: The Epigenomic Landscape of Aberrant Splicing in Cancer

Authors: Donghoon Lee, Jing Zhang, Mark B Gerstein, Yale University

Abstract: Nearly all protein-coding genes undergo alternative RNA splicing, which provides an important mean to expand transcriptome diversity beyond the scope of genomic information. While splicing is an elaborate process, it can be prone to errors that could become pathogenic. Unsurprisingly, aberrant splicing, which collectively refers to splicing events that could confer risk of a disease, is often implicated in cancer.

Recent studies have revealed splicing regulation is characterized by increased levels of nucleosome density and positioning, DNA methylation, and distinct histone modification patterns. However, most studies on aberrant splicing have largely focused on identifying genomic- and transcriptomic-level variations within splice sites, cis-acting splicing regulatory elements, and trans-acting splicing factors. The extent, nature, and effects of epignomic dysregulation in aberrant splicing remain unsolved.

By systematically profiling the epigenomic landscape of aberrant splicing using transcriptomic and epigenomic data from the ENCODE and the Epigenome Roadmap projects, we aimed to (1) identify chromatin status and distinct epigenetic signatures that characterize aberrant splicing in cancer, (2) classify aberrant splicing by different class of epigenomic dysregulation, and (3) elucidate the role of epigenomic control in aberrant splicing. The proposed study will significantly advance our understanding of epigenomic contribution to aberrant splicing in cancer.

Poster #204: Identifying and Resolving Inconsistencies in Biological Pathway Resources

Authors: Lucy L Wang, John Gennari, Neil Abernethy, University of Washington

Abstract: Biological pathways provide a high-level view of biological and disease processes, and have become a popular tool for studying genetic and molecular interactions. Many pathway knowledge bases exist providing complementary information; there have been attempts to integrate these resources to improve our analysis and understanding of biology. However, the same biological processes are represented differently in different resources, as each resource makes its own choices in knowledge representation. There is currently no accepted standardized way to integrate such data. A method is needed to access the collective knowledge of all these different data sources.

In order to merge information across pathway knowledge bases, inconsistencies must be identified and understood. Inconsistencies are found in 1) entity annotation, 2) entity existence, 3) reaction semantics, 4) reaction and entity granularity, 5) asserted level of information, and 6) external references. We identified these types of inconsistencies in several human pathway resources: HumanCyc, KEGG, PANTHER, and Reactome. We also provide recommendations for aligning pathways between resources, thereby providing biologists new ways to use and interpret the existing knowledge. This in turn is essential for furthering our understanding of biology and pathology, paving the way to advances in pathway analysis and drug target identification.

Poster #205: Conserved Transcriptional Regulators Control Divergent Toxin Production in Fungi

Authors: Abigail L Lind, Timothy D Smith, Ana M Calvo, and Antonis Rokas, Vanderbilt University and Northern Illinois University

Abstract: Filamentous fungi produce diverse secondary metabolites (SMs) essential to their ecology and adaptation. Fungal SMs have a double-edged impact on humans; some are carcinogenic toxins found in contaminated food supplies, while others, such as lovastatin and penicillin, have been repurposed as successful therapeutics. SMs play crucial roles in fungal ecology; lovastatin and penicillin, for example, are both antimicrobial compounds that provide their producers with a competitive advantage. In fungi, SMs are extremely diverse; each SM is typically produced by only a handful of species. The production of SMs is triggered by both biotic and abiotic factors and is controlled by widely conserved transcriptional regulators. To understand how the transcriptional regulators of SM regulate such divergent pathways under different conditions, we examined the genome-wide regulatory role of several master SM regulators in different fungal species and in different environmental conditions. Our findings indicate that master SM regulators undergo rapid transcriptional rewiring and interact with multiple abiotic signals to control SM production.

Poster #206: Determining Gene Expression Trends using Single-Cell RNA-seq with CREoLE

Authors: Geoffrey F Schau, Andrew Adey, Oregon Health and Science University

Abstract: Single-cell RNA-sequencing (scRNA-seq) is widely used to recapitulate gene expression trends through developmental time of heterogeneous biological tissue. Although several methods have sought to estimate pseudo-temporal expression trends, a number of technical limitations presented by scRNA-seq remain, including high expression variability and drop-out measurements, complicating trend estimation. We hypothesize that consensus estimation made by iteratively sub sampling expression profiles of individual cells will yield a smoother, more biologically accurate expression trend less susceptible to technical noise. To address this need, we have developed CREoLE, Consensus Representative Estimation of Lineage Expression, a general purpose algorithm designed to appropriately scale the dimensionality of scRNA-seq data, establish a branching lineage pathway substructure, and produce smooth, high-resolution gene expression trends through each developmental lineage.

Our analysis includes a comparison of current methods to CREoLE on both simulated as well as publicly available scRNA-seq data. In the simulation studies, we examined the impact of varying levels of artificial noise and drop out measurements. In these cases, CREoLE returns similar estimations at all evaluated noise levels and recapitulates published expression trends from literature, supporting our hypothesis that trend smoothing is feasible by calculating consensus estimation. CREoLE is implemented in R and is publicly available on GitHub.

Poster #207: Analysis of Orphan Disease Gene Networks to Enable Drug Repurposing

Authors: Kelly Regan, Zachary Abrams, Philip R O Payne, Department of Biomedical Informatics, The Ohio State University

Abstract: Over 7,000 orphan diseases have been described, while treatments exist for fewer than 400 due to their limited prevalence, lack of research resources and reduced commercial potential. Thus, drug repurposing represents an ideal alternative in order to circumvent the high costs and inefficiencies of the current drug discovery pipeline. Previous research has shown that disparate orphan diseases are highly connected through genetic mechanisms. Connectivity mapping is a computational drug repurposing system that exploits the observation that changes in gene expression patterns can reflect different conditions in human cells, such as exposure to drugs, gene-modifying agents and disease processes. We obtained orphan disease-gene relationship data from the Orphan Disease Network and Orphanet databases. Functional implications (e.g. GOF/LOF status) of orphan disease gene mutations were confirmed using the OMIM database. We focused on disease-causing germline mutation genes corresponding to reduced gene protein product and/or function in order to align with LINCS gene knock-down perturbation experiments. This study represents the first systematic application of gene expression-based connectivity mapping of orphan diseases for drug repurposing and to recapitulate known disease-disease relationships. Using network community detection algorithms, we have identified novel drug candidates for a subset of highly connected orphan disease network modules.

Poster #208: Signal-Oriented Pathway Analyses Reveal a Signaling Complex as a Synthetic Lethal Target for p53 Mutations

Authors: Songjian Lu, Chunhui Cai, Gonghong Yan, Zhuan Zhou, Yong Wan, Lujia Chen, Vicky Chen, Gregory F Cooper, Lina M. Obeid, Yusuf A Hannun, Adrian V Lee and Xinghua Lu, University of Pittsburgh

Abstract: The multi-omics data from The Cancer Genome Atlas (TCGA) provide an unprecedented opportunity to investigate cancer pathways and therapeutic targets through computational analyses. In this study, we developed a signal-oriented computational framework for cancer pathway discovery. First, we identify transcriptomic modules that are abnormally expressed in multiple tumors, such that genes in a module are most likely regulated by a common aberrant signal. Then, for each transcriptomic module, we search for a set of somatic genome alterations (SGAs) that perturbs the signal regulating the transcriptomic module. Computational evaluations indicate that our methods can identify pathways perturbed by SGAs. In particular, our analyses revealed that SGAs affecting *TP53*, *PTK2*, *YWHAZ*, and *MED1* perturb a set of signals that promote cell proliferation, anchor-free colony formation, and epithelial-mesenchymal transition (EMT). We further demonstrate that these proteins form a signaling complex that mediates these oncogenic processes in a coordinated fashion. These findings lead the hypothesis that disrupting the complex could be a novel therapeutic strategy for treating tumors with these genomic alterations. Finally, we show that disrupting the signaling complex by knocking down *PTK2*, *YWHAZ*, or *MED1* attenuates and reverses oncogenic phenotypes caused by mutant p53 in a "synthetic lethal" fashion. This signal-oriented framework for searching pathways and therapeutic targets is applicable to all cancer types, and thus potentially could have a broad impact on precision medicine in cancer.

Poster #209: Towards a Knowledge-Base for Biochemical Reasoning

Authors: McShan, Daniel and Hunter, L, University of Colorado-Denver

Abstract: KaBOB is knowledge-integration framework focused on genes and proteins, intended to support mechanistic explanations of experimental results in genomics, transcriptomics and proteomics. Extending it to include metabolic information would facilitate analysis of metabolomic datasets as well. Potential metabolomic knowledge sources for integration include HumanCyc with 1826 metabolites, ChEB with 3947 "human metabolites", and the Human metabolome database (HMDB) wth 29289 "endogenous" human metabolites.

HMDB has an order of magnitude more metabolites than HumanCyc or ChEBI largely because it curates not only small molecules but lipids, which are important in metabolism and signalling. HMDB provides cross references to HumanCyc (1174) and ChEBI (2791). Of these, only 1064 are cross-referenced to both; 1767 are in ChEBI, not HumanCyc, and 235 are in HumanCyc, not ChEBI. However, HMDB is not a superset of these other two data sources. Compared to what they self report, 36% (652/1826) metabolites are in HumanCyc but not in HMDB, and 29% (1156/3947) are in ChEBI but not HMDB.

In order to create a comprehensive knowledge-base of metabolites, each of these sources must be integrated. To do so, the KaBOB framework requires that each knowledge source be converted into a formal semantic relationship grounded in Open Biomedical Ontologies and expressed in the Semantic Web standard OWL language. Future work involves semantic mappings for each of the sources, and a set of queries demonstrated the ability to access knowledge seamlessly from all of them simultaneously.

Poster #301: Informatics Approaches for Evidence Appraisal and Synthesis

Authors: Andrew D Goldstein, Eric Venker, Chunhua Weng, Columbia University

Abstract: Clinical evidence should be valid, applicable, and synthesized. Unfortunately, bias, error, misconduct, and underreporting harm validity. Applicability is often inadequately defined and validated. Synthesis can be sporadic, redundant, or lacking rigor, completeness, or timeliness. Underlying these issues is the volume, disorganization, and under-appraisal of evidence. We surveyed the informatics literature addressing these issues, and defined knowledge gaps and intervention opportunities.

We first conducted a scoping review of articles focused on evidence appraisal and synthesis in 8 biomedical informatics journals. The search yielded 838 citations; 53 were included, representing 0.2% of all 24813 citations. Interventions included classifiers (60%), ontologies (17%), and social computing (9%). For classifiers, articles were predominantly validation studies, not broad implementations. For ontologies and social computing, articles were predominantly perspective pieces. Generally, appraisal tools had descriptive, not critical functions, and synthesis tools were aimed at search and inclusion, not subsequent synthesis processes.

Next, we are conducting a scoping review of articles focused on evidence appraisal in the broader biomedical literature to develop a conceptual framework, identify barriers, and propose informatics solutions. Initial analysis demonstrates that appraisal is not systematic, formal, or integrated into the scientific corpus and that existing attempts at solving this are problematic.

Poster #302: Using Wearable Technology to Aid in the Classification of Different Cardiac Arrhythmias

Authors: Jessica N Torres, Euan Ashley, Stanford University

Abstract: Cardiovascular diseases such as Atrial fibrillation (AF) and hypertrophic cardiomyopathy (HCM) increase the risk of stroke, heart failure, and even sudden death. The largest obstacle to early AF and HCM detection is its tendency to be intermittent and asymptomatic. Current clinical practices fails to capture latent risk situations such as changes in magnitude or variability over time or under specific conditions. Wearable technology affords the opportunity to continuously monitor patients through wireless medical sensors or mobile biosensors. This massive amount of real-time biometric data may hold invaluable clues for improving human health. In our study, we use a Samsung Simband device, a health-focused wearable technology, to monitor patient's physiological characteristics. Here, we present methods to process optical high-intensity LEDs technology known as photoplethysmography (PPG) signal for 1) estimating heart rate in the high intensity motion and 2) AF and HCM arrhythmia detection and classification. We find that knowledge gained from this application can lead to a better understanding of how new wearable technologies can be used to classify abnormal cardiac arrhythmias.

Poster #303: Predicting Heterogenous Causal Treatment Effects for First-Line Antihypertensives

Authors: Alejandro Schuler, Nigam Shah, Stanford University

Abstract: Hypertension (high blood pressure) is an overwhelmingly prevalent risk factor for negative cardiovascular outcomes, including heart disease and stroke. Despite being treatable, many patients struggle to control their hypertension. This is partly because there is considerable heterogeneity in patient responses to different classes of antihypertensive drugs. Although the different classes of hypertensive drugs are equally effective at a population level, it is not currently known which specific patients will respond better to which antihypertensives. We use statistical learning to predict patients' individual blood pressure responses to antihypertensive treatments using only their medical histories up to the point of their first prescription. To avoid confounding, we employ a sophisticated method of causal inference called a causal forest, which is conceptually a form of data-driven stratified matching. Our analysis is performed on the OHDSI common data model, which will enable us to validate our findings across multiple sites.

Poster #304: Acquiring and Representing Drug-Drug Interaction Knowledge and Evidence

Authors: Jodi Schneider and Richard D Boyce, University of Pittsburgh

Abstract: Potential drug-drug interactions (PDDIs) are a significant source of preventable drug-related harm. Poor quality evidence on PDDIs, combined with prescribers' general lack of PDDI knowledge, results in thousands of preventable medication errors each year. One contributing factor is that PDDI knowledge lacks a standard computable format. To address this, we are researching efficient strategies for acquiring and representing PDDIs knowledge, focusing on assertions and their supporting evidence.

We are acquiring knowledge from several sources. First, we have transformed 410 assertions and 519 evidence items from prior work. Second, we are examining FDA-approved drug labels, and so far annotators have identified 609 evidence items relating to pharmacokinetic PDDIs from 27 FDA-approved drug labels. Third, annotators have found 230 assertions of drug-drug interactions in 158 non-regulatory documents, including full text research articles.

We are building a two-layer evidence representation, with both generic and domain-specific layers. The generic layer reuses the Micropublications Ontology to annotate assertions and their supporting data, methods, and materials. For the domain-specific component we are building DIDEO–the Drug-drug Interaction and Drug-drug Interaction Evidence Ontology. DIDEO adds specific knowledge, such as the study types required to establish a given type of PDDI. The current version of DIDEO has 385 subclass axioms, and reuses formalized knowledge items, including from the Drug Ontology, Chemical Entities of Biological Interest, the Ontology of Biomedical Investigations, and the Gene Ontology.

Poster #305: Impact of Missing Data on Automatic Learning of Clinical Guidelines

Authors: Yuzhe Liu, Vanathi Gopalakrishnan, University of Pittsburgh

Abstract: Many machine learning algorithms ignore data with missing values. When learning on retrospective clinical data where missing values are common, discarding incomplete entries may significantly reduce the sample size or bias the resulting complete dataset. In our dataset used to learn clinical guidelines for imaging use in pediatric cardiomyopathy, eliminating patients with missing data reduces the dataset size by half. Recent work has shown success using machine learning techniques like decision trees, k-nearest neighbors, and self organizing maps to impute missing data in several real world datasets. We are investigating the impact of various imputation methods on the performance of our Bayesian rule learning technique for discovery of clinical guidelines. We compared the performance of mean value, k-nearest neighbor, and decision tree imputation as well as using indicator variables for missingness against performance on a complete dataset after deleting samples with missing values.

Poster #306: Understanding Clinical Trial Patient Screening from the Coordinator's Perspective

Authors: En-Ju D Lin[1], Stephen Johnson[2], Albert M Lai[1],
[1]Department of Biomedical Informatics, The Ohio State University; [2]Weill Cornell Medical Center

Abstract: Clinical research is crucial for generating evidence and providing effective treatments for patients. However, clinical trials are lengthy and expensive processes that often fail. Slow recruitment has been cited as a primary reason for the failure of clinical trials. Currently, clinical research coordinators typically perform the time consuming process of manually comparing a patient's, frequently complex, clinical history against a series of eligibility criteria. To address the challenges in recruitment, we plan to develop an automated approach to support pre-screening patients into clinical trials using data from the electronic health records (EHR). We first want to understand how clinical research coordinators identify and pre-screen patients for clinical trials, their needs and their experience with using EHR in the screening process. We conducted semi-structured interviews with 16 clinical trial coordinators at two large academic research medical centers. The interview covered four aspects: screening productivity, the use of EHR, eligibility criteria and language, and attitude towards automation. Using a conventional content analysis approach, two authors (EL and SJ) coded all transcripts and analyzed the concepts arose from the interviews. We have identified current needs and important considerations for moving towards automation.

Poster #307: Standardizing Sample-Specific Metadata in the Sequence Read Archive

Authors: Matthew N Bernstein[1] and Colin N Dewey[1,2,3]
[1]Department of Computer Sciences; [2]Department of Biostatistics and Medical Informatics; [3]Center for Predictive Computational Phenotyping, University of Wisconsin, Madison

Abstract: The NCBI's Sequence Read Archive (SRA) promises great biological insight if one could analyze the data in the aggregate; however, the data remains largely underutilized, in part, due to the unstructured nature of the metadata associated with each sample. The rules governing submissions to the SRA do not dictate a standardized set of terms that should be used to describe the biological samples from which the sequencing data are derived. As a result, the metadata include many synonyms, spelling variants, and references to outside sources of information. For these reasons, it remains difficult to query the database for biological samples that have certain targeted attributes such as specific diseases, tissues, or cell-types. In this poster, I describe our current effort in mapping each biological sample to terms in standardized ontologies. More specifically, we are developing a computational pipeline that automatically associates with each sample in the SRA database a set of terms in the Open Biomedical Ontologies.

Poster #308: Causal Inference During Multisensory Speech Perception

Authors: John Magnotti[1], Genevera Allen[2], and Michael Beauchamp[1]
[1] Baylor College of Medicine, Houston, TX, [2] Rice University, Houston, TX

Abstract: Speech is the primary form of human communication and is fundamentally multisensory: we seamlessly integrate visual information from a talker's facial movements and auditory information from the talker's voice. Integrating information across senses is especially important to counteract ubiquitous hearing loss during normal aging and is clinically relevant for the impaired language abilities observed in autism, schizophrenia, dyslexia, and stroke.

A first step toward eliminating multisensory integration deficits is a computational understanding of multisensory speech perception. Current computational models are based around the assumption that humans automatically integrate all available information from a talker's voice and face. Daily experiences and laboratory data, however, show that humans are selective in which information they choose to combine, and that this selection varies greatly from person to person. To solve this selection problem, we developed a novel graphical model based on the general idea of causal inference.

We applied our causal inference model to speech perception data from healthy individuals (N=265). Our model outperformed state-of-the-art Bayesian perceptual models, providing a more accurate computational framework for the study of multisensory speech perception. Measuring parameter differences across individuals and clinical groups can give us insight into the underlying reasons for measured differences in face-to-face communication.

This research was funded by a training fellowship from the Gulf Coast Consortia, on the Training Program in Biomedical Informatics, National Library of Medicine (NLM) T15LM007093, PD – Lydia E. Kavraki.

#309: Data Mining for Identifying Candidate Drivers of Drug Response in Heterogeneous Cancer

Author: Sheida Nabavi, University of Connecticut

Abstract: With advances in technologies, huge amounts of multiple types of high-throughput genomics data are available. These data have tremendous potential to identify new and clinically valuable biomarkers to guide the diagnosis, assessment of prognosis, and treatment of complex diseases. Integrating, analyzing, and interpreting big and noisy genomics data to obtain biologically meaningful results, however, remains highly challenging. Mining genomics datasets by utilizing advanced computational methods can help to address these issues.

To facilitate the identification of a short list of biologically meaningful genes as candidate drivers of anti-cancer drug resistance from an enormous amount of heterogeneous data, we employed statistical machine-learning techniques and integrated genomics datasets. We developed a computational method that integrates gene expression, somatic mutation, and copy number aberration data of sensitive and resistant tumors. In this method, an integrative method based on module network analysis is applied to identify potential driver genes. We applied this method to the ovarian cancer data from the cancer genome atlas. The method yields a short list of aberrant genes that also control the expression of their co-regulated genes. The final result contains biologically relevant genes, such as COL11A1, which has been recently reported as a cis-platinum resistant biomarker for ovarian carcinoma.