Extraction of Pharmacogenomics Traits from Electronic Health Records

WILLIAM S. BUSH PHD MS

Assistant Professor
Institute for Computational Biology
Department of Epidemiology and Biostatistics
Case Western Reserve University

SCHOOL OF MEDICINE
CASE WESTERN RESERVE
UNIVERSITY

Overview

- Introduction to Electronic Medical Records
- The Process of "Electronic Phenotyping" and Challenges of Pharmacogenomics Phenotyping
- Example of Calcineurin-Inhibitor Toxicity in Heart Transplant Recipients
- Example of Statin Myotoxicity Detection

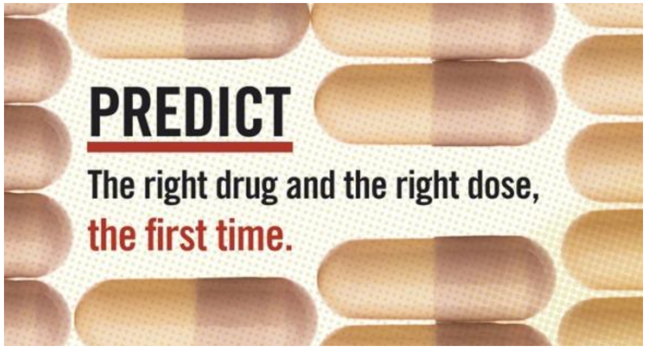
Electronic Medical/Health Records

- Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009
  - Incentivizes the nationwide adoption of EMR systems in the US

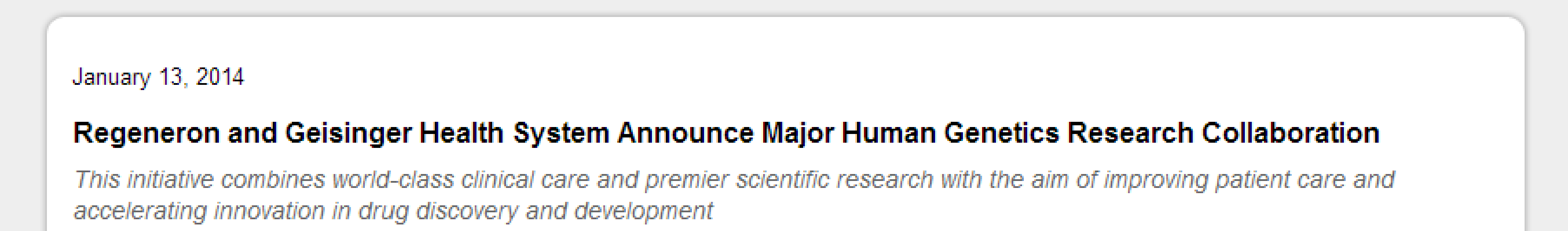
- While there are standards and commonalities, EMR systems are very heterogeneous
  - Epic Systems, Allscripts, Meditech, Cerner, IBM, McKesson, Siemens, GE Healthcare
  - **Epic** is most common at large medical centers

Benefits to Personalized Medicine

- Clinical Decision Support



- Research and Discovery

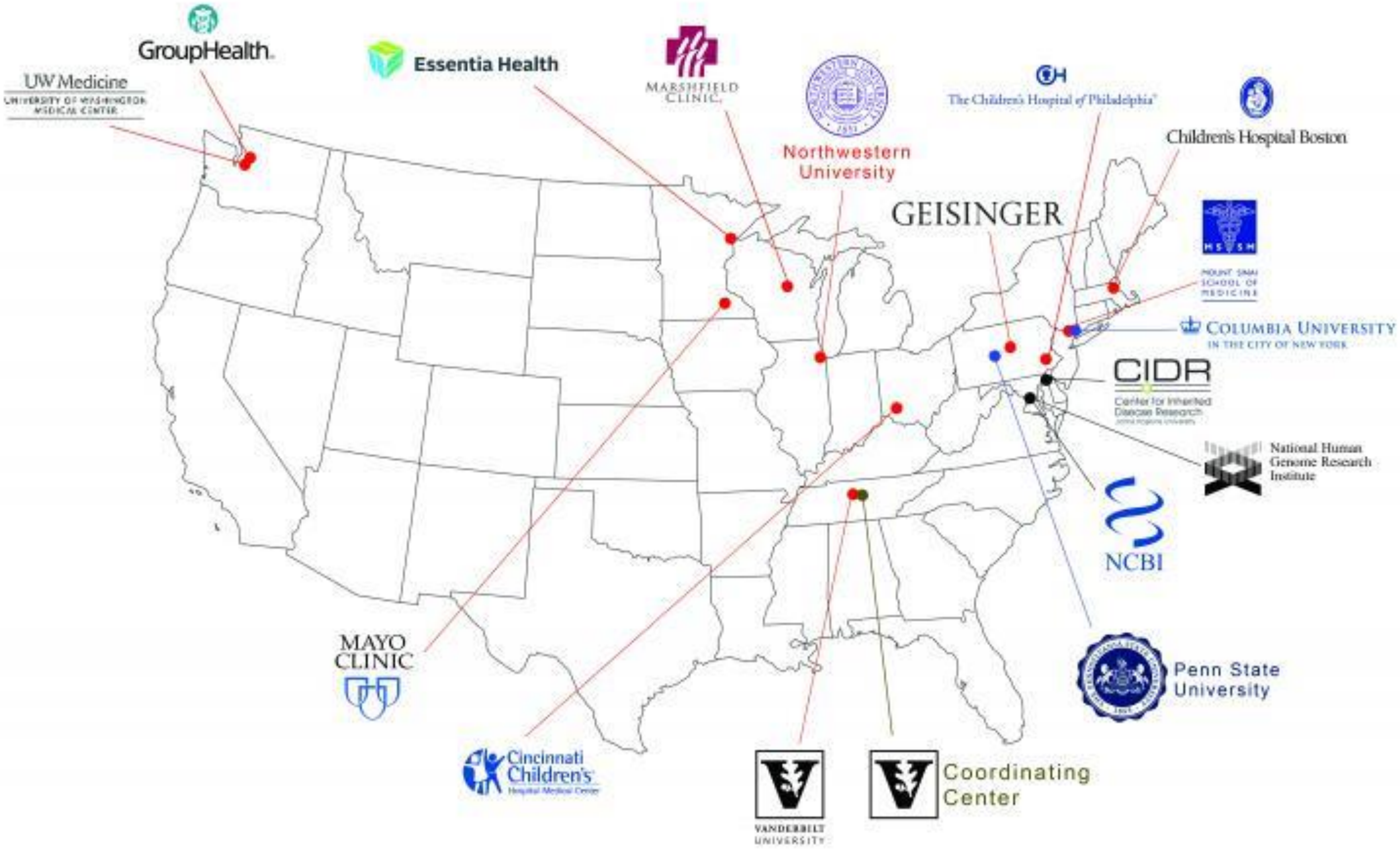


January 13, 2014

Regeneron and Geisinger Health System Announce Major Human Genetics Research Collaboration

This initiative combines world-class clinical care and premier scientific research with the aim of improving patient care and accelerating innovation in drug discovery and development

EMR-Linked Biorepositories



PMID: 24987407

EMR Components

- Structured Elements
- Pseudo-Structured Text
- Unstructured Text

- EMR elements can differ between systems and clinics

Structured Elements

- Demographic information
  - Date of birth, race/ethnicity, gender
- Vital Signs
  - Heart rate, blood pressure, height, weight, body temperature
- Some Laboratory Values
  - WBC, insulin, glucose, glomerular filtration rate (GFR)
- Billing and Procedure codes
  - ICD9/10, CPT, ICD-O-3

Pseudo-structured Text

- Free text documents with a loosely standardized format

- History and Physical Examination (HPE)

  "… approximately 10 months status post bilateral bunionectomy with metatarsal head resections 2 through 5 bilaterally. She denies any pain in her feet although she is a little upset that she has had a recurrence of her bilateral hallux valgus, left worse than the right.

  She does have some residual hallux valgus, worse on the left then on the right, but this is not uncommon following bunionectomy. We discussed the fact that her bunions were so bad to begin with, that her feet actually look pretty good."

- Problem List
  - Known significant medical conditions/diagnoses
  - Significant procedures
  - Allergies and Medications

Unstructured Text

- Clinical communications (phone calls, prescription refills, etc)
- Discharge notes
- Clinic-specific notes
- Some laboratory and procedure reports (CT scan, colonoscopy, etc)
- “Medical flotsam”

My EMR Existential Crisis

At some point in everyone's life, they realize that…

1. Their parents don't know everything.
2. Their doctor doesn't know everything.
3. Their medical record may be wrong.

EMRs are very useful, but they can be noisy and inaccurate.

Beware…

More structure DOES NOT imply higher accuracy

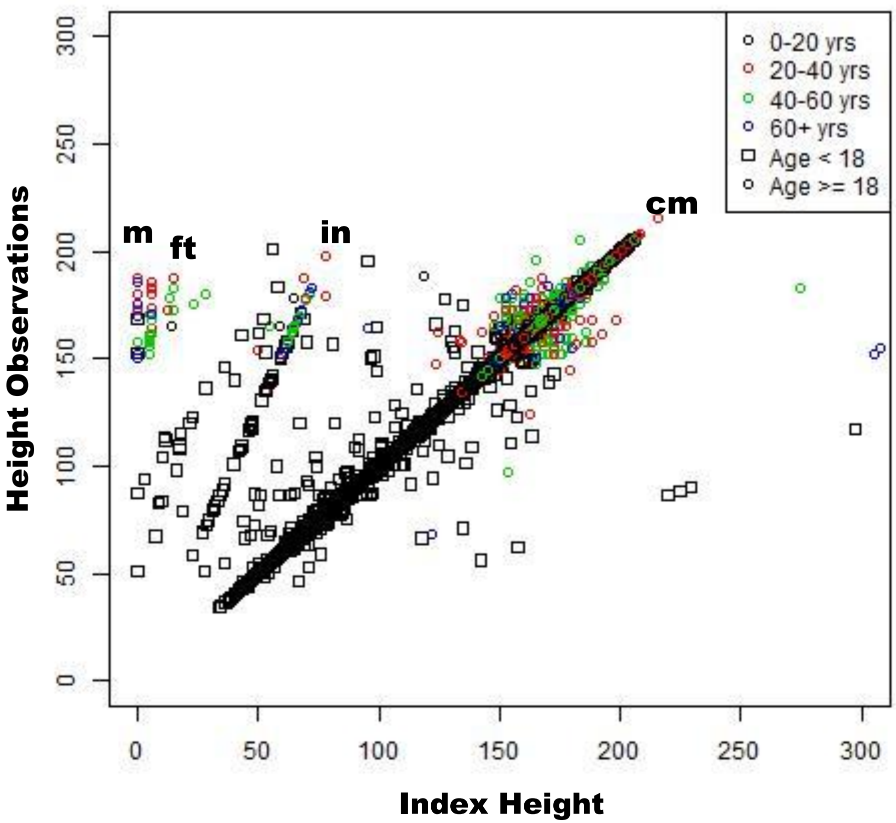
Problem Lists often include past conditions (prior to EMR or clinic visit)

In the US, ICD codes are used for… BILLING. Billing does not always reflect reality.

Example: Height and Weight

- The most ubiquitous measures reported in an EMR

Robert Goodloe



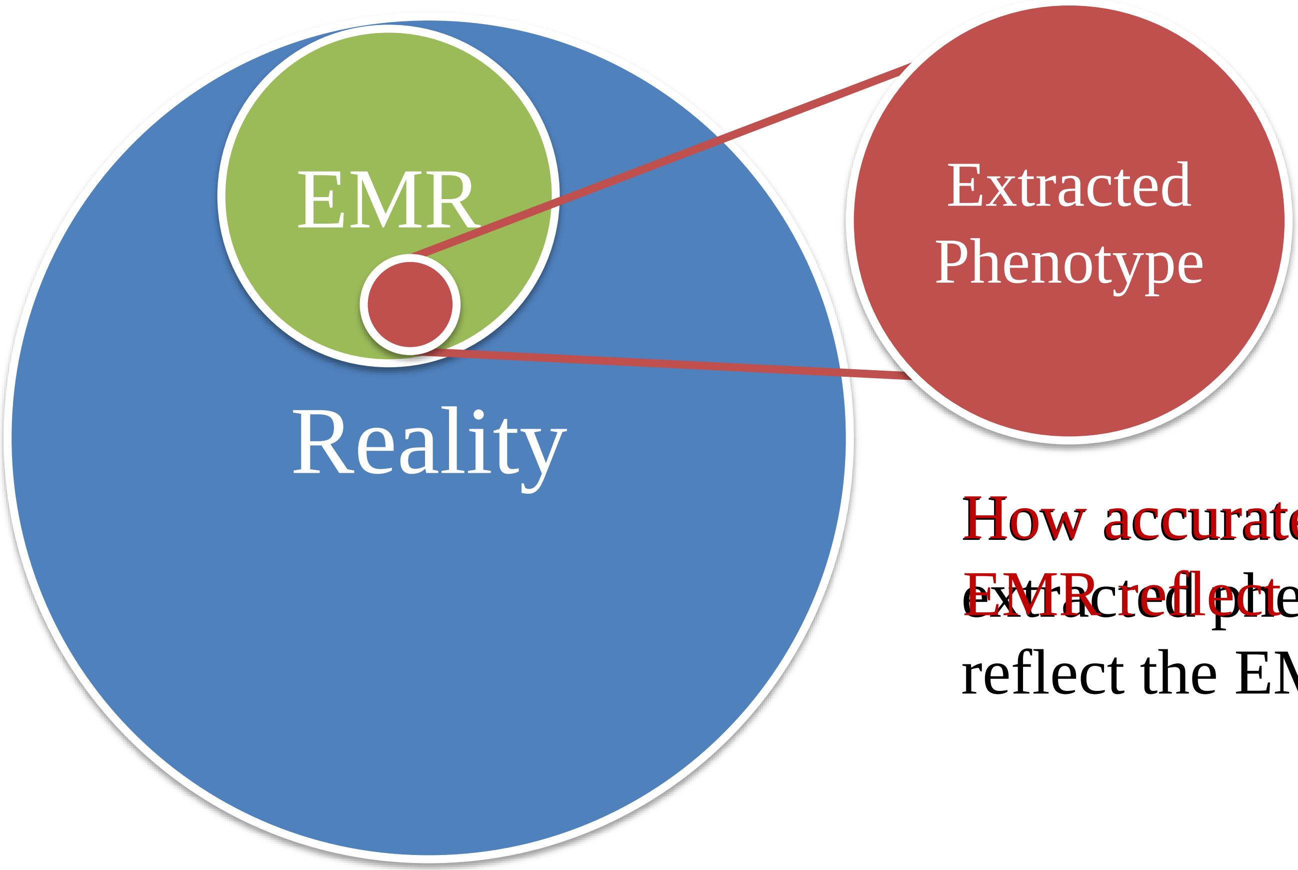
General Phenotyping Process

1. Consult with clinicians to understand the representation of the phenotype in the EMR
2. Develop an initial algorithm based on extracted EMR elements (unstructured text is more difficult)
3. Perform a manual review of algorithm-defined cases and controls and assess accuracy
4. Refine algorithm and repeat 2 & 3 until accuracy is acceptable

Algorithm Components

- Presence/absence of structured elements (codes)
- Ranges of quantitative values (labs, vital signs)
- Key word searches for inclusion/exclusion
- Natural language processing and/or concept identification
- Temporal sequencing

Algorithm Evaluation



How accurately does our extracted phenotype reflect the EMR?

How accurately does the EMR reflect reality?

Algorithm Evaluation

- Positive and Negative Predictive Values
- Consider that EMRs have a biased direction of reporting (better PPV than NPV for most conditions)
- Handling of outlier values

Covariates

- Each covariate must be treated as its own phenotype and requires algorithm development
- Height, Weight, BMI
- Smoking Status

Research and applications

ICD-9 tobacco smoking status

Laura K Wiley,[1,2] Anush

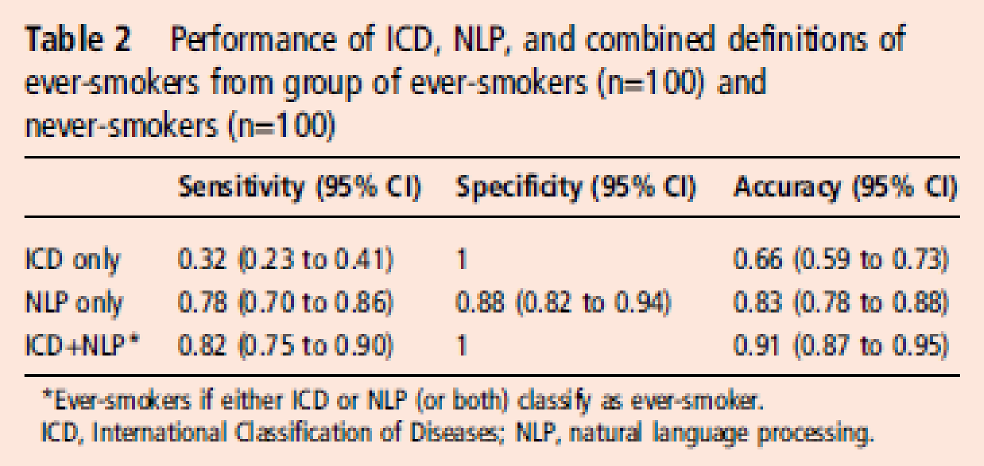


Table 2 Performance of ICD, NLP, and combined definitions of ever-smokers from group of ever-smokers (n=100) and never-smokers (n=100)

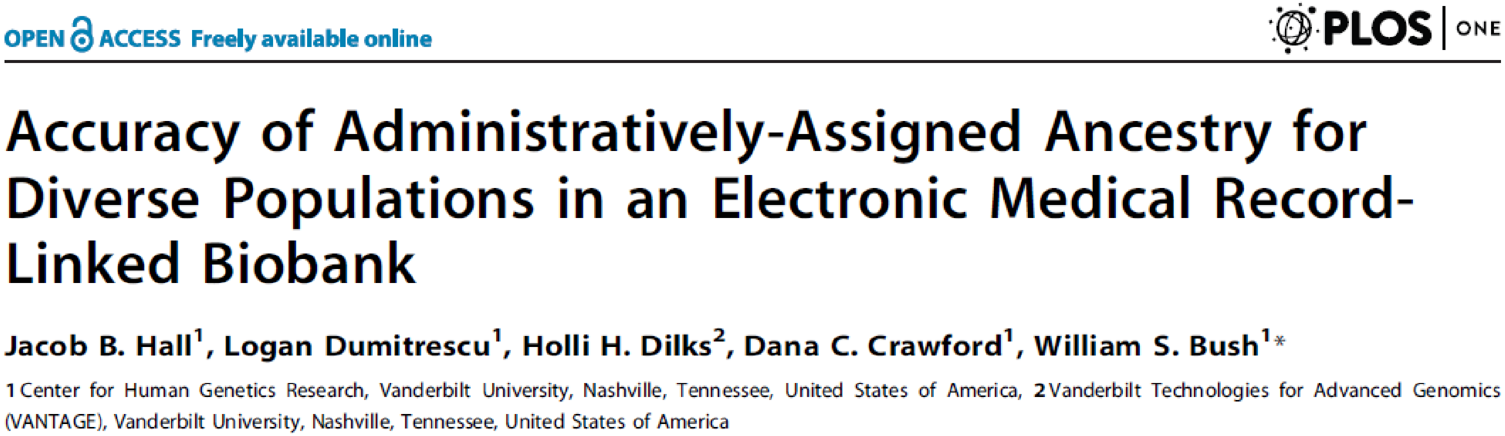
	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)
ICD only	0.32 (0.23 to 0.41)	1	0.66 (0.59 to 0.73)
NLP only	0.78 (0.70 to 0.86)	0.88 (0.82 to 0.94)	0.83 (0.78 to 0.88)
ICD+NLP*	0.82 (0.75 to 0.90)	1	0.91 (0.87 to 0.95)

*Ever-smokers if either ICD or NLP (or both) classify as ever-smoker.
ICD, International Classification of Diseases; NLP, natural language processing.

PMID: 23396545

Race and Genetic Ancestry

- Race may be defined by different criteria
- May not always be self report



OPEN ACCESS Freely available online

PLOS | ONE

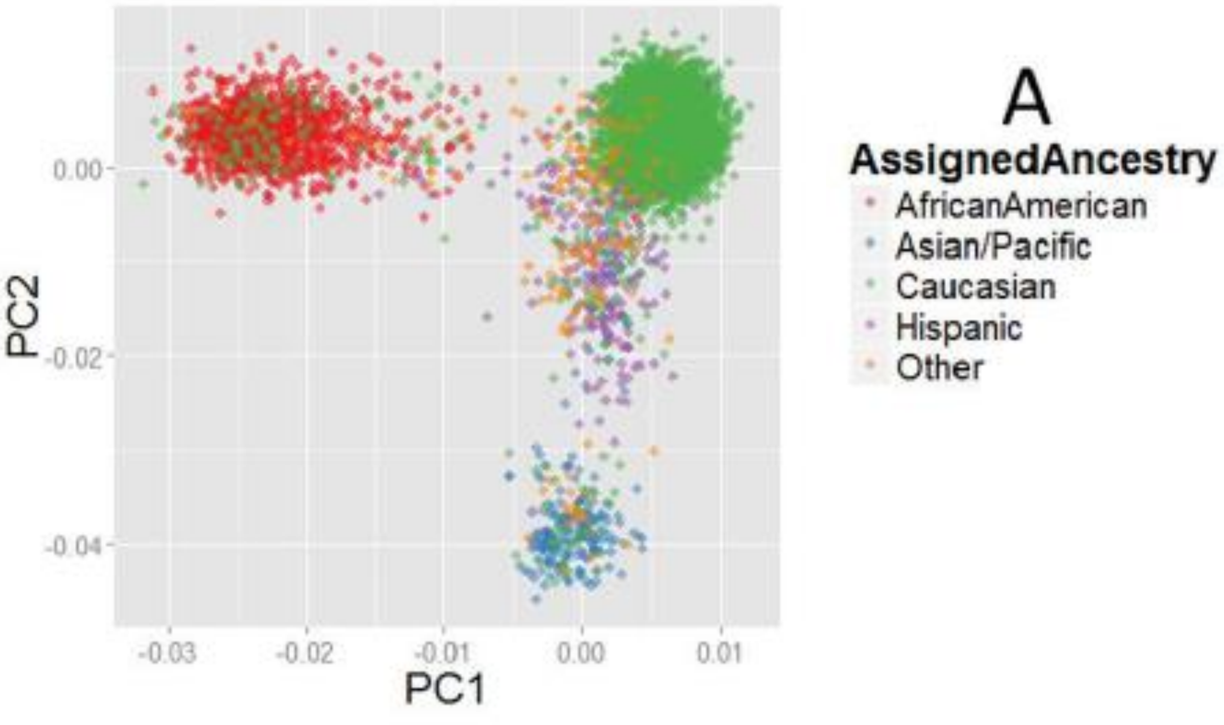
Accuracy of Administratively-Assigned Ancestry for Diverse Populations in an Electronic Medical Record-Linked Biobank

Jacob B. Hall[1], Logan Dumitrescu[1], Holli H. Dilks[2], Dana C. Crawford[1], William S. Bush[1]*

[1] Center for Human Genetics Research, Vanderbilt University, Nashville, Tennessee, United States of America, [2] Vanderbilt Technologies for Advanced Genomics (VANTAGE), Vanderbilt University, Nashville, Tennessee, United States of America

PMID: 24896101

Race and Genetic Ancestry



Jake Hall

- General agreement for European and African Descent populations

PMID: 24896101

Pharmacogenomic Phenotypes

- Dose Response
- Categorical Response/Non-Response
- Adverse Event Detection

- Medication information is typically not structured

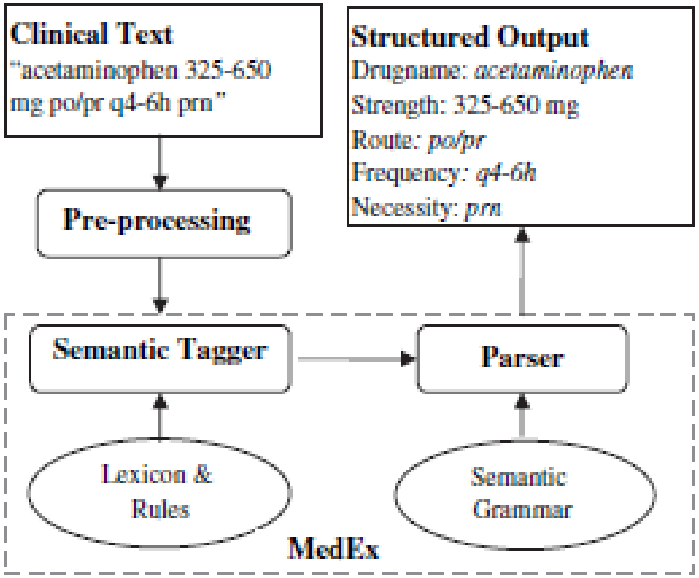
Medications

- Natural Language Processing

  “The patient takes atorvastatin 20mg tablets, ½ tablet daily”

[redacted]

MedEx: a medication inf
for clinical narratives



Clinical Text
"acetaminophen 325-650 mg po/pr q4-6h prn"

Pre-processing

Semantic Tagger

Lexicon & Rules

Parser

Semantic Grammar

MedEx

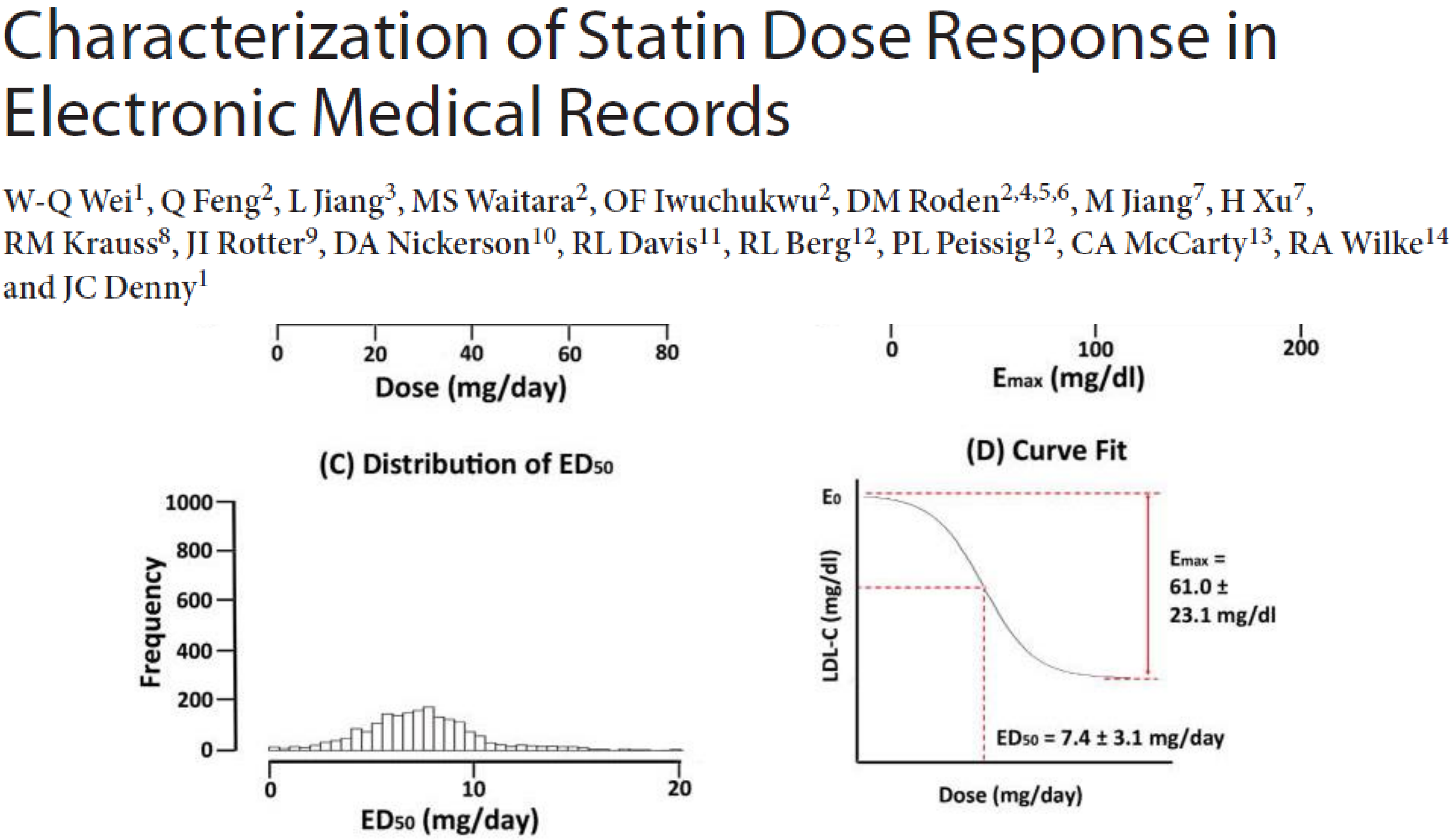
Structured Output
Drugname: *acetaminophen*
Strength: 325-650 mg
Route: *po/pr*
Frequency: *q4-6h*
Necessity: *prn*

Figure 1 An overview of the MedEx system.

Hua Xu,[1] Shane P Stenner,[1,2] Son Doan,[1] Kevin B Johnson,[1,3] Lemuel R Waitman,[1] Joshua C Denny[1,2]

PMID: 20064797

Simvastatin Dose-Response



Characterization of Statin Dose Response in Electronic Medical Records

W-Q Wei[1], Q Feng[2], L Jiang[3], MS Waitara[2], OF Iwuchukwu[2], DM Roden[2,4,5,6], M Jiang[7], H Xu[7], RM Krauss[8], JI Rotter[9], DA Nickerson[10], RL Davis[11], RL Berg[12], PL Peissig[12], CA McCarty[13], RA Wilke[14] and JC Denny[1]

Case Study: Calcineurin-inhibitor Toxicity

Work by Matt Oetjens, Will Bush, Russ Wilke, Josh Denny, Kelly Birdwell, and Dana Crawford

PMID: 24297552

Calcineurin-inhibitor Toxicity

- Given post-transplant to prevent organ rejection
  - Tacrolimus and cyclosporine
- Narrow therapeutic window
- Nephrotoxicity is a serious and common complication
- Serum creatinine and glomerular filtration rates (GFR) are monitored post-transplant to assess kidney function

Study Design

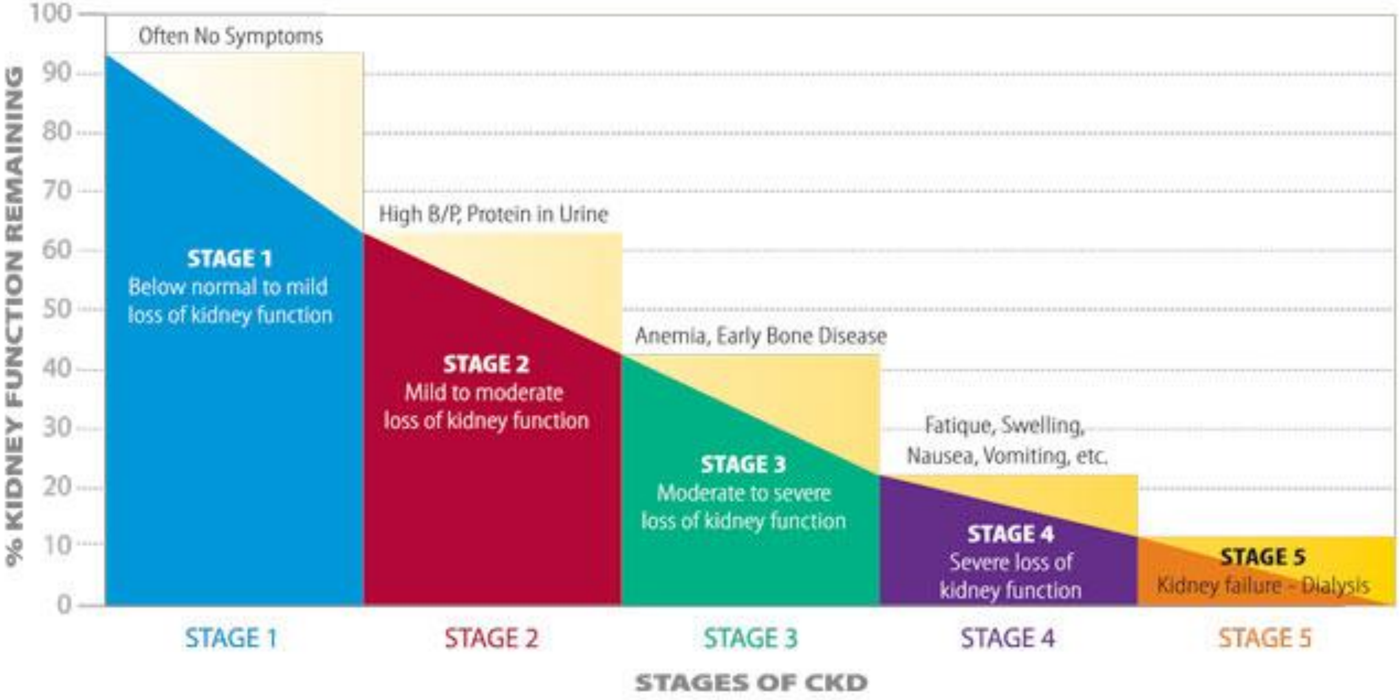
- Patients identified having:
  - Heart transplant documented with >= 3 ICD9 Code V42.1 (heart replaced by transplant) and/or one CPT Code 33945 (cardiectomy with heart transplant)
  - One or more mention of an immunosuppressant
  - Age > 15 at date of transplant
  - Available DNA

Clinical Covariates

- BMI
- Serum creatine
- Systolic and diastolic blood pressure and hypertension
  - Monthly medians
  - Relevant medication
- Chronic kidney disease (defined by ICD9 codes)
- Diabetes mellitus (defined by ICD9 codes)

Phenotype

- Chronic kidney disease is classified in 5 stages of severity (determined by estimated GFR)



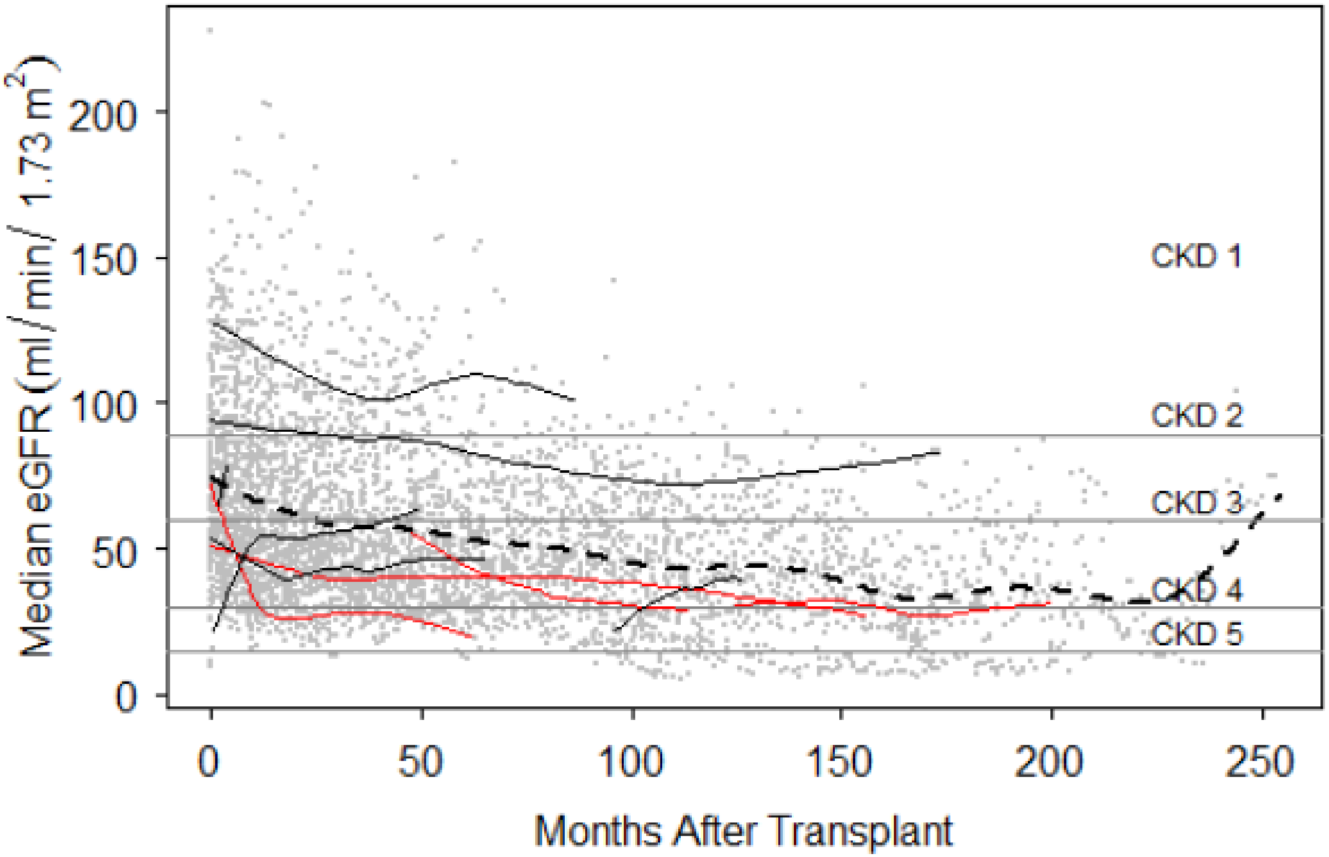
Phenotype

- eGFR is calculated

$$186 \times \text{Serum Creatinine}^{-1.154} \times \text{Age}^{-0.203} \times [1.212 \text{ if Black}] \times [0.742 \text{ if Female}]$$

- Severe Kidney Disease: Post-transplant eGFR < 30 mL/min/1.73m^2 for 3 consecutive months
- Time to development of severe nephrotoxicity clinically attributed to calcinurin inhibitor toxicity

Clinical Trajectories



Key Points

- Study population is under clinical surveillance
- Phenotype is based on established, repeated clinical measures
- Clear alternate endpoints (i.e. dialysis, death)
- Manual review required for some attributes (date of transplant)

Case Study: Statin Myotoxicity

Work by Laura Wiley, Jeremy Moretz, Josh Denny, Josh Peterson, and Will Bush

Statin Myotoxicity

- Statins are the most widely prescribed class of drugs in the world

- The most common side effect of statin use is muscle toxicity
- Ranges from mild muscle aches to rhabdomyolysis (rapid muscle breakdown)
- Myotoxicity rates are estimated between 9-20%
- May be a common reason for non-compliance

Dealing with a Nebulous Phenotype

- An adverse drug reaction that may not be reported, or if reported may not be well documented
- Lots of ancillary causes of general myopathy
- Investigation and resolution of symptoms can vary widely among providers
  - Some order a CK measurement
  - Some simply switch the statin
  - Some tell the patient "its all in their head"

Phenotyping Strategies

- ICD9 codes for myopathy related events
- Creatine kinase measures (indication of muscle breakdown)
  - Various ranges
  - Various exclusions (with troponin measures)
- Natural language processing

PMIDs: 22912565, 19476582, 23942138, 23530940, 23546564, 22195188

Defining a Gold Standard

- Selected all individuals who had mention of a statin
- Randomly selected 300 individuals (enriched for myopathy related terms found in the medication extraction process)
- All records were reviewed by two independent reviewers (grad student and clinical pharmacy resident)
- Flagged records considered myotoxic

Multiple Phenotyping Approaches

- Indications must be dated after first statin mention
- ICD9 – “toxic myopathy”, “poisoning by antilipemic drugs”, “other myopahies”, “myopathy unspecified”, etc
- Multiple Creatine Kinase measurement criteria
- Mention of statin in the drug allergy section of problem list

Gold Standard Characteristics

- Independent review agreement was strong (K = 0.986)
- Average of 2.6 statins per patient
- 1.7 myotoxic statin associations per patient

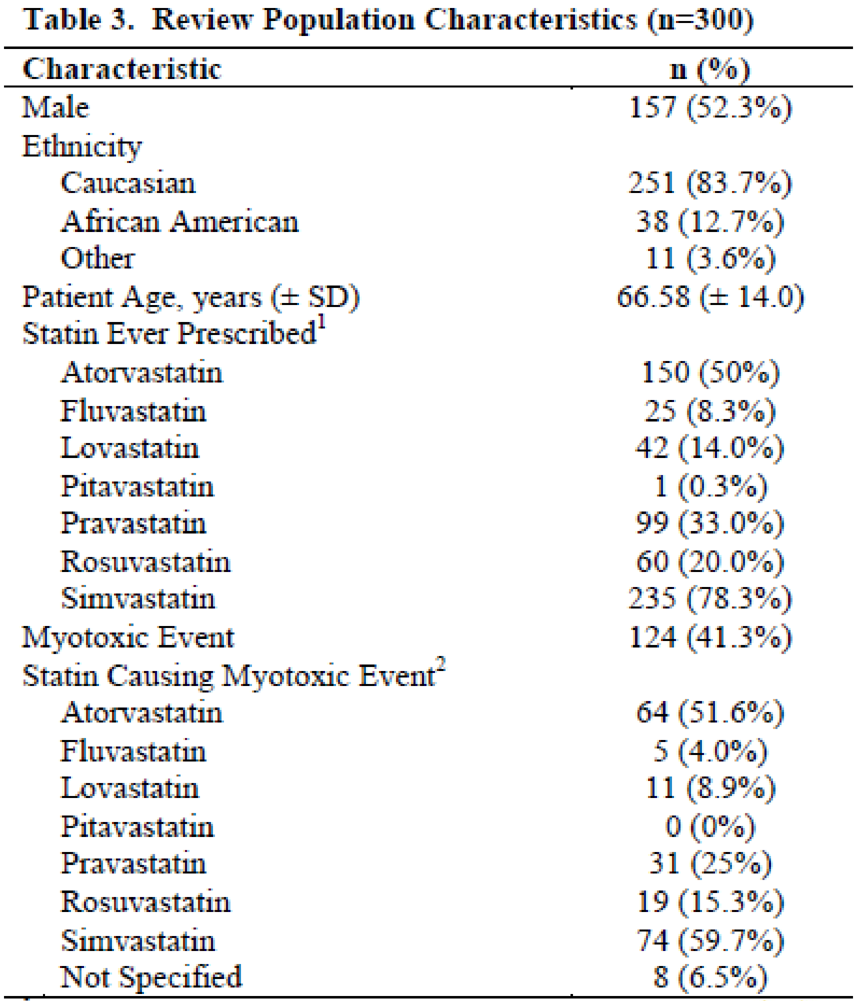
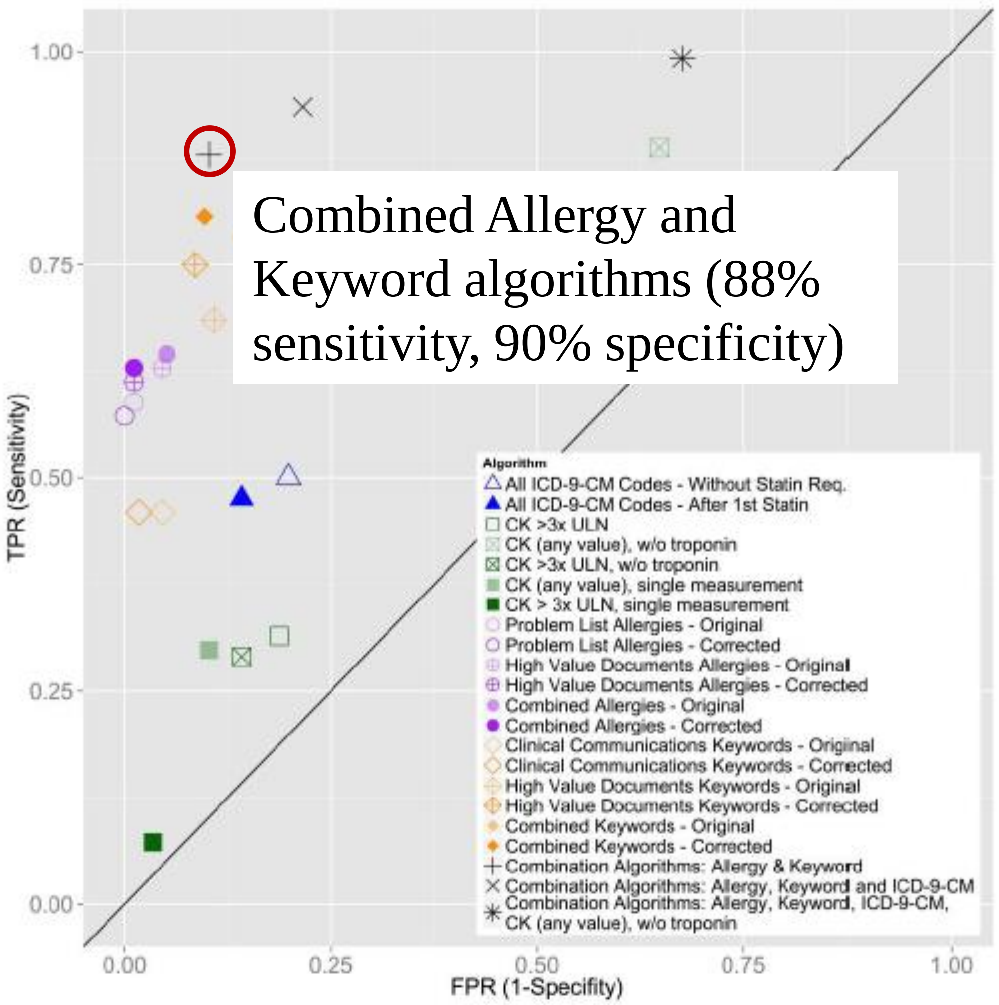


Table 3. Review Population Characteristics (n=300)

Characteristic	n (%)
Male	157 (52.3%)
Ethnicity	
Caucasian	251 (83.7%)
African American	38 (12.7%)
Other	11 (3.6%)
Patient Age, years (± SD)	66.58 (± 14.0)
Statin Ever Prescribed[1]	
Atorvastatin	150 (50%)
Fluvastatin	25 (8.3%)
Lovastatin	42 (14.0%)
Pitavastatin	1 (0.3%)
Pravastatin	99 (33.0%)
Rosuvastatin	60 (20.0%)
Simvastatin	235 (78.3%)
Myotoxic Event	124 (41.3%)
Statin Causing Myotoxic Event[2]	
Atorvastatin	64 (51.6%)
Fluvastatin	5 (4.0%)
Lovastatin	11 (8.9%)
Pitavastatin	0 (0%)
Pravastatin	31 (25%)
Rosuvastatin	19 (15.3%)
Simvastatin	74 (59.7%)
Not Specified	8 (6.5%)

Algorithm evaluation



Combined Allergy and Keyword algorithms (88% sensitivity, 90% specificity)

Key Points

- Study population was ill-defined (mention of a statin)
- Phenotype is based on non-structured attributes, free text
- Clinical course of action is not well-defined
- Confidence in algorithm performance, but pushes the boundaries of how well the EMR captures the phenotype

Conclusions

- EMRs contain a wealth of information, but it is not always easy to extract
- Understanding the clinical use of EMRs is critical
- Medication-related traits can be defined, but rely on pseudo/unstructured data elements
- Outcomes based on routine and consistent clinical measures are the low-hanging fruit.

Acknowledgements

- Bush Lab
  - Jake Hall
  - Laura Wiley
  - Alex Fish
  - Mike Sivley
- Crawford Lab
  - Matt Oetjens
  - Robert Goodloe
  - Eric Farber-Eger
- Josh Denny
- Josh Peterson
- Jeremy Moretz
- Kelly Birdwell