High-Throughput Machine Learning from EHR Data

David Page

Department of Biostatistics & Medical Informatics, and

Center for Predictive Computational Phenotyping (CPCP)

University of Wisconsin-Madison

Acknowledgements

NIH BD2K Center for Predictive Computational Phenotyping

Ross Kleiman

Paul Bennett

Michael Caldwell

Scott Hebbring

Miron Livny

Peggy Peissig

Vitor Santos Costa

Humberto Vidaillet

Wisconsin Genomics Initiative

Marshfield Clinic

CPCP

NIH National Institutes of Health

Turning Discovery Into Health

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The Electronic Health Record (EHR)

Demographics

ID	Year of Birth	Gender
P1	3.22.1963	M

Diagnoses

ID	Date	Diagnosis	Sign/Symptom
P1	6.2.1990	427.69 (PVC)	Palpitations

The Electronic Health Record (EHR)

Demographics

ID	Year of Birth	Gender
P1	3.22.1963	M

Diagnoses

ID	Date	Diagnosis	Sign/Symptom
P1	7.3.1997	Elevated BP	

The Electronic Health Record (EHR)

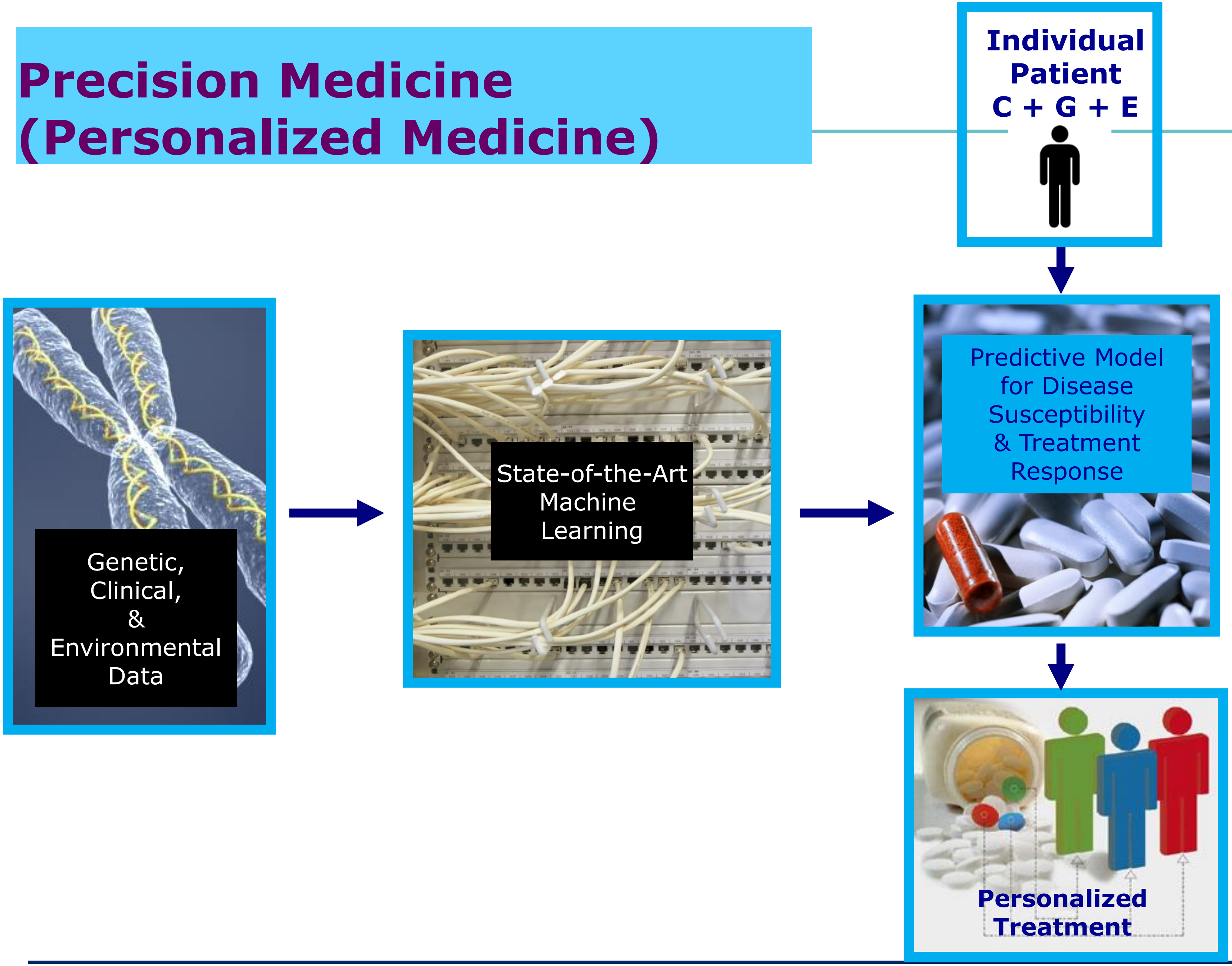
Demographics

ID	Year of Birth	Gender
P1	3.22.1963	M

Diagnoses

ID	Date	Diagnosis	Sign/Symptom
P1	9.1.1998	Atrial Fibrillation	Shortness of Breath

Precision Medicine (Personalized Medicine)



Individual Patient
C + G + E

Genetic, Clinical, & Environmental Data

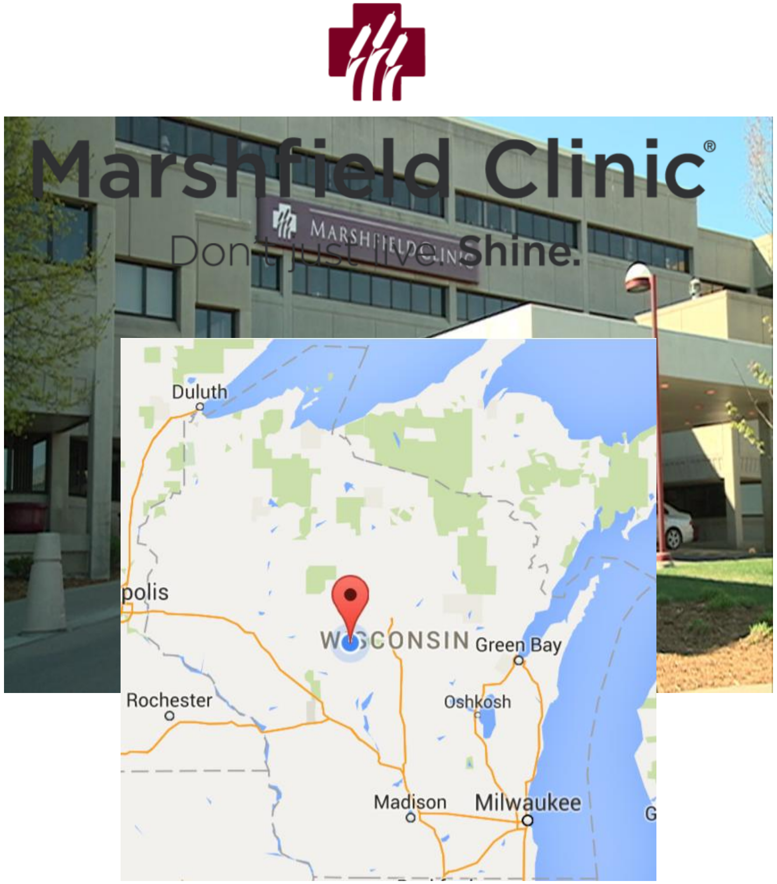
State-of-the-Art Machine Learning

Predictive Model for Disease Susceptibility & Treatment Response

Personalized Treatment

Marshfield Clinic EMR

- Marshfield Clinic
  - Health system in North Central Wisconsin
- 1.5M Patient Records spanning 40 years
  - Demographics
  - Diagnoses (ICD-9)
  - Labs
  - Procedures
  - Vitals



Electronic Health Record (EHR)

PatientID	Gender	Birthdate
P1	M	3/22/63

PatientID	Date	Physician	Symptoms	Diagnosis
P1	1/1/01	Smith	palpitations	hypoglycemic
P1	2/1/03	Jones	fever, aches	influenza

PatientID	Date	Lab Test	Result
P1	1/1/01	blood glucose	42
P1	1/9/01	blood glucose	45

PatientID	SNP1	SNP2	…	SNP500K
P1	AA	AB		BB
P2	AB	BB		AA

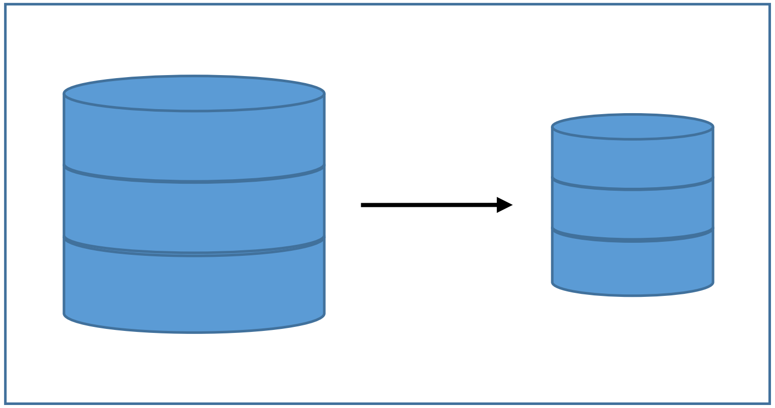
PatientID	Date Prescribed	Date Filled	Physician	Medication	Dose	Duration
P1	5/17/98	5/18/98	Jones	prilosec	10mg	3 months

Vision

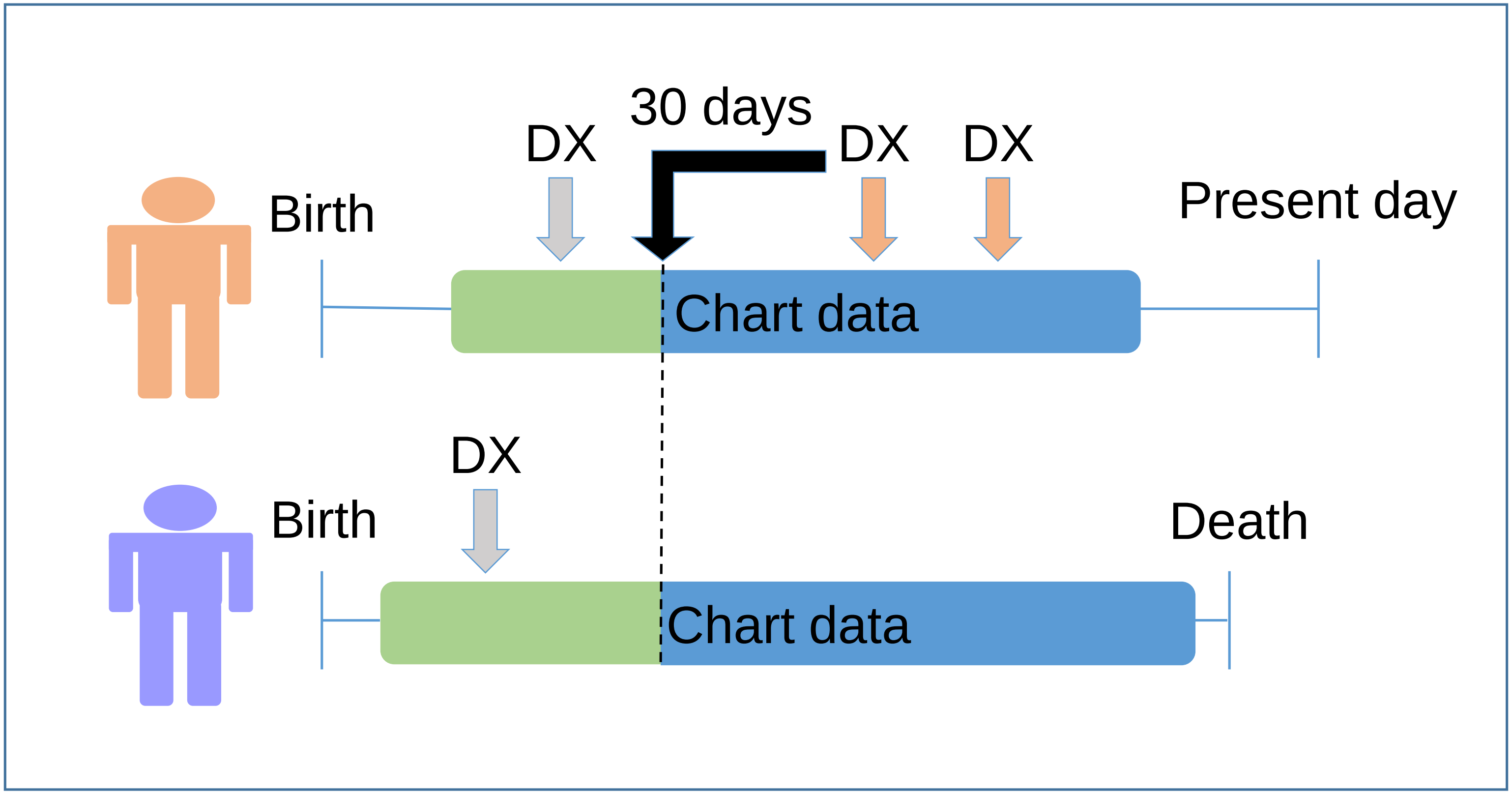
- Build predictive models for every diagnosis, every procedure, response to every drug, at press of a button.
- Translate the most accurate models into the clinic, whether as decision support algorithms or lessons for clinicians, FDA, etc.

Data Cleaning

- Originally 1.5M patients
- Remove Infrequent Patients
  - 4 diagnoses and 2 encounters
- 1.1M patients remained (~73%)

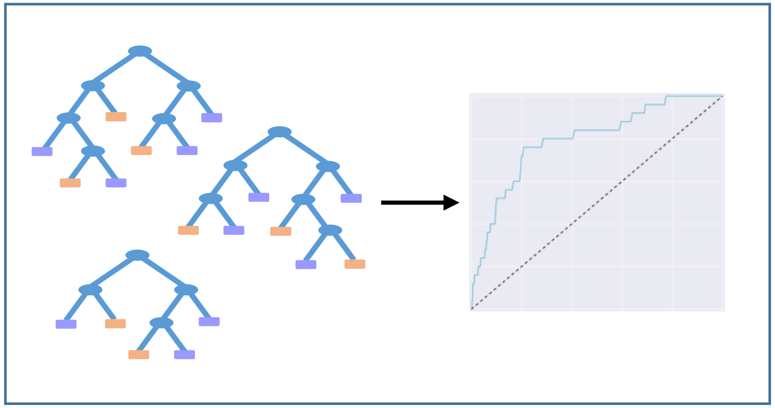


Case Control Matching

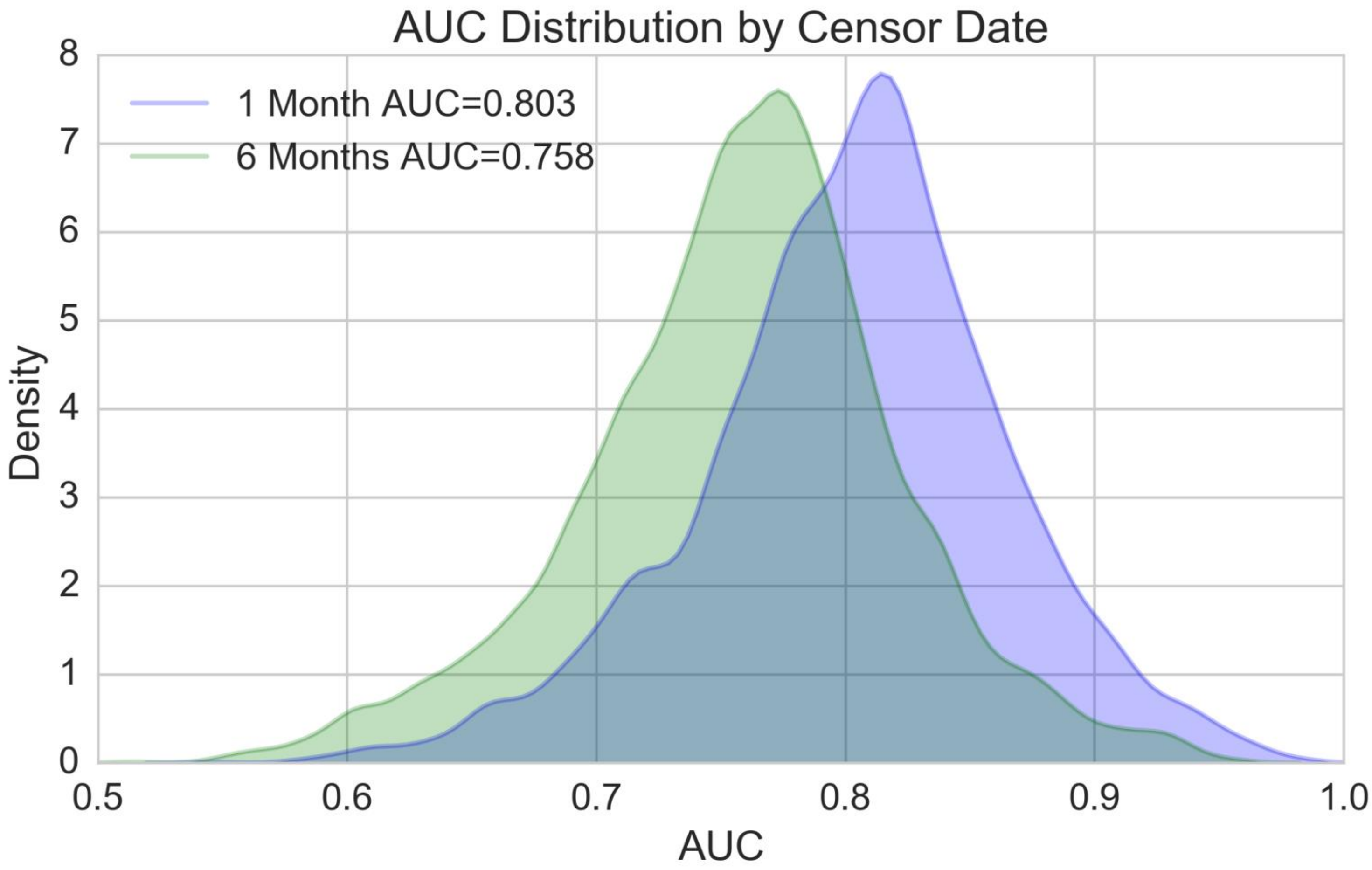


Model Construction and Evaluation

- Model nearly every ICD-9 code
  - At least 500 pairs
  - Exclude symptoms
- Build random forest model
- Evaluate models via AUC-ROC

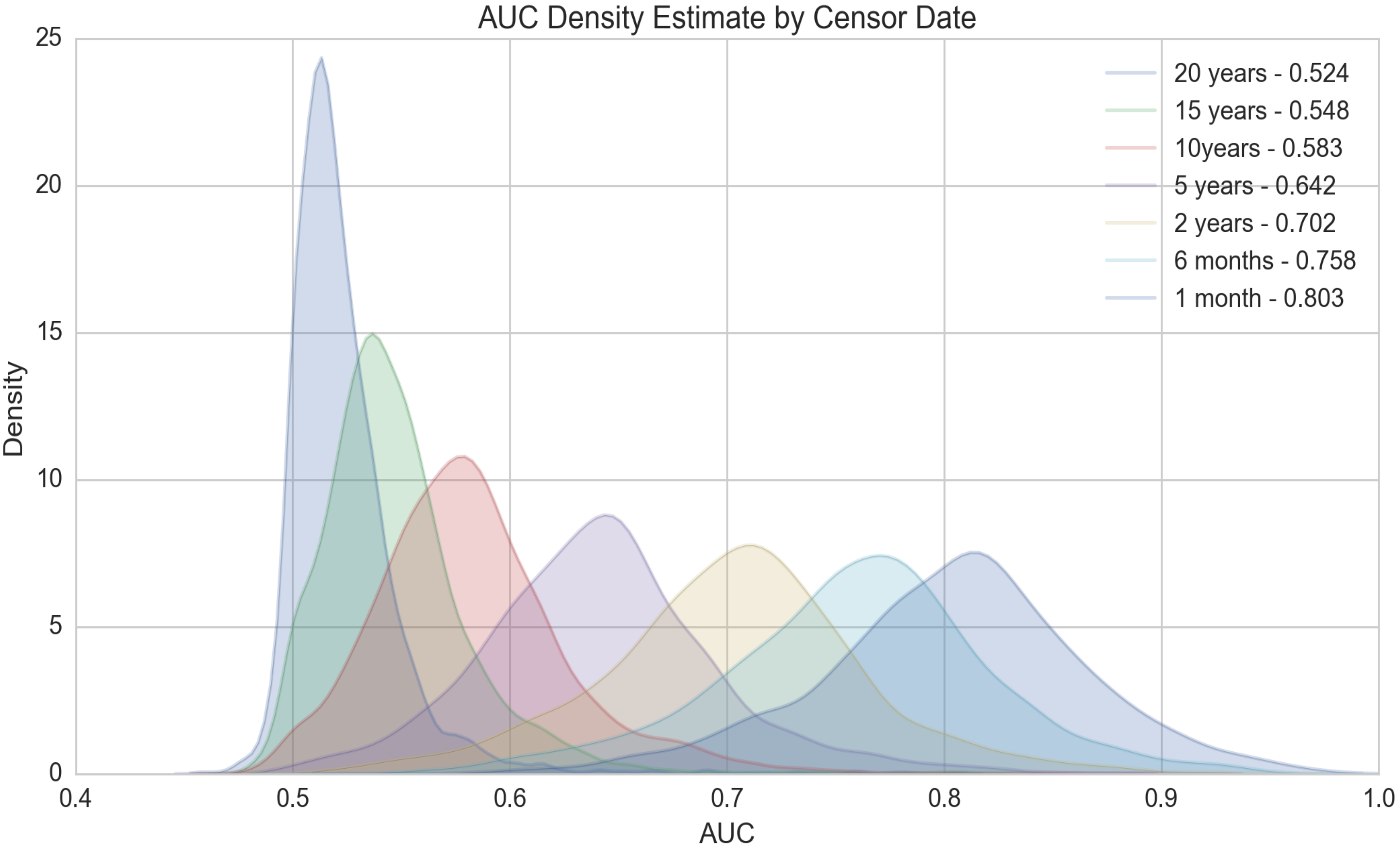


Predictive Accuracy of Models



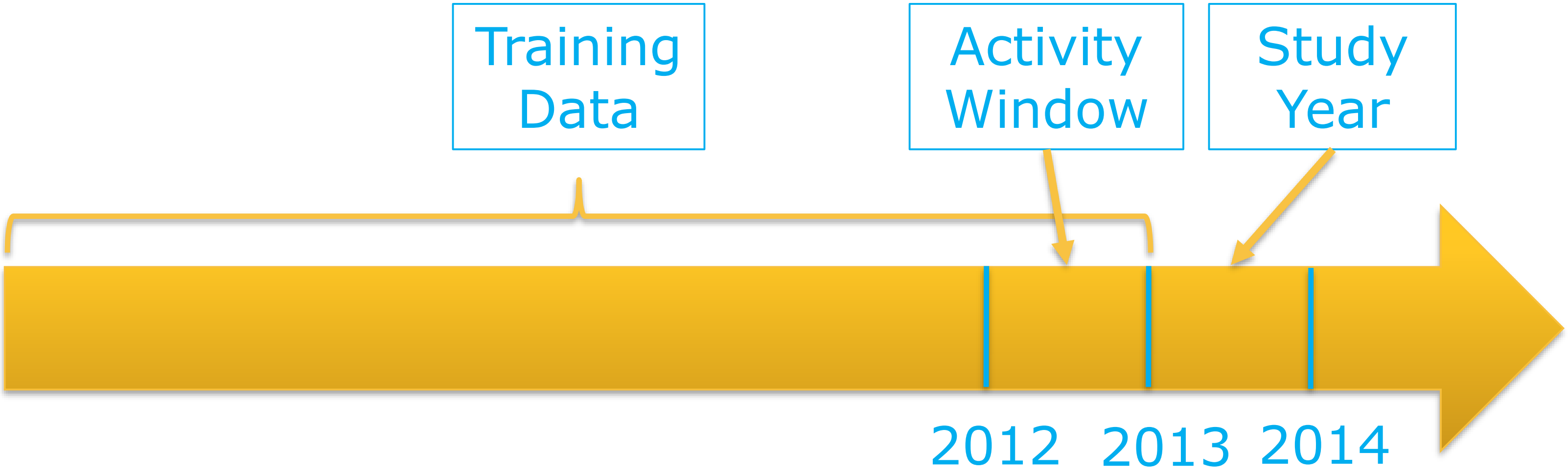
High-Throughput ML (Kleiman, Bennett, et al.)

Predicting Every ICD Diagnosis Code at the Press of a Button

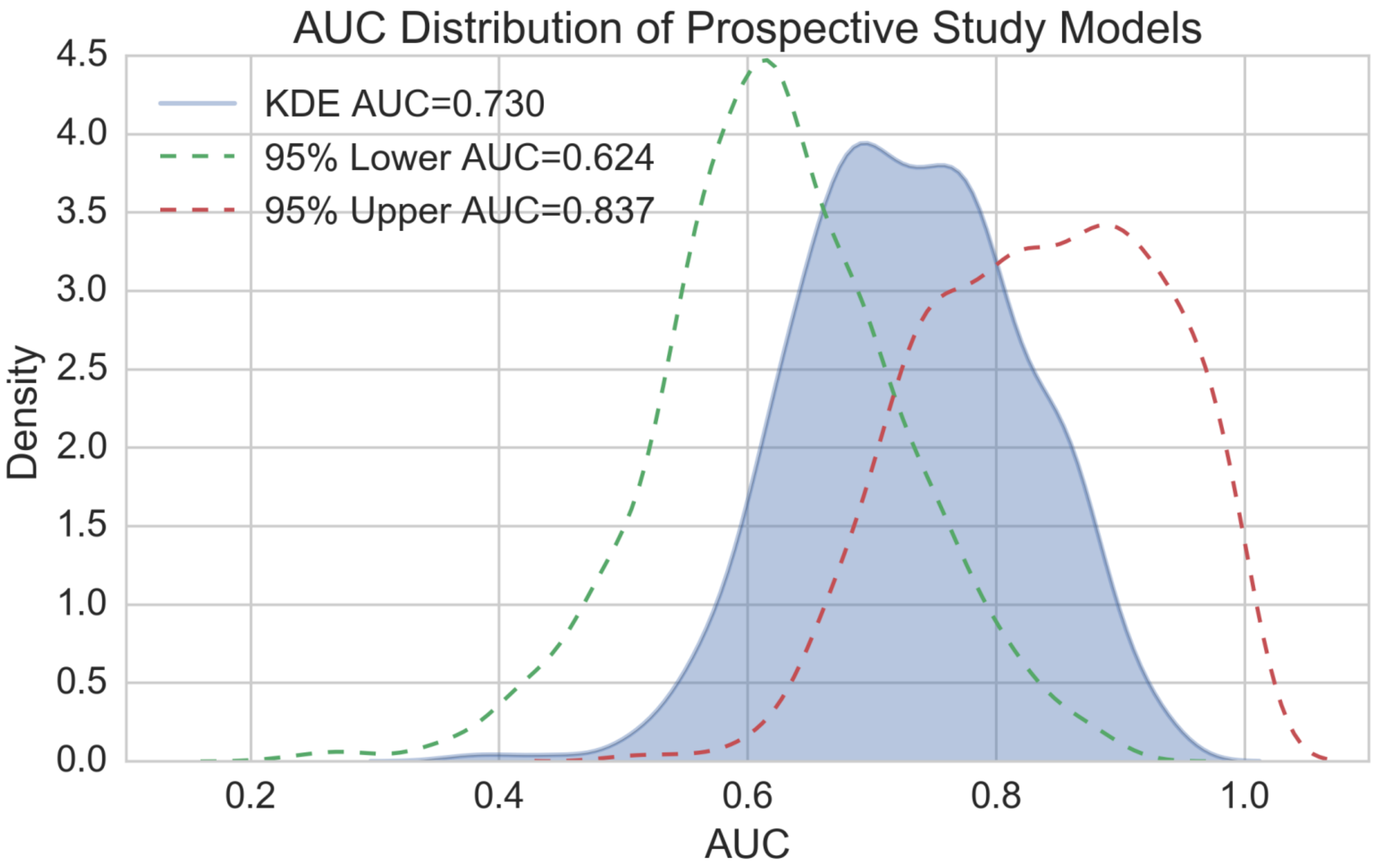


Simulated Prospective Study

- How well would these models perform in practice?
- Evaluate model accuracy on 10,000 test patients



Simulated Prospective Study Results



HTCondor Essential to this Work and Future Work

- Over 1M patients
- Over 4000 different diagnoses (models)
- 750 trees per model
- Producing slide 14 took 30K jobs and roughly 123 years of compute time
- In future, predict all drugs, procedures, and responses
- In future, predict on 100M or 1B patients
- In future, add genomics (3B bp per patient)
- In future, add tumor genomes (1000 genomes per tumor)
- High-throughput ML applicable to many other domains
- High-throughput computing applicable to many other tasks in NIH Big Data to Knowledge Program