Some Lessons Learned from Using Machine Learning Methods to Predict Adverse Health Outcomes

Mark Craven

Department of Biostatistics & Medical Informatics

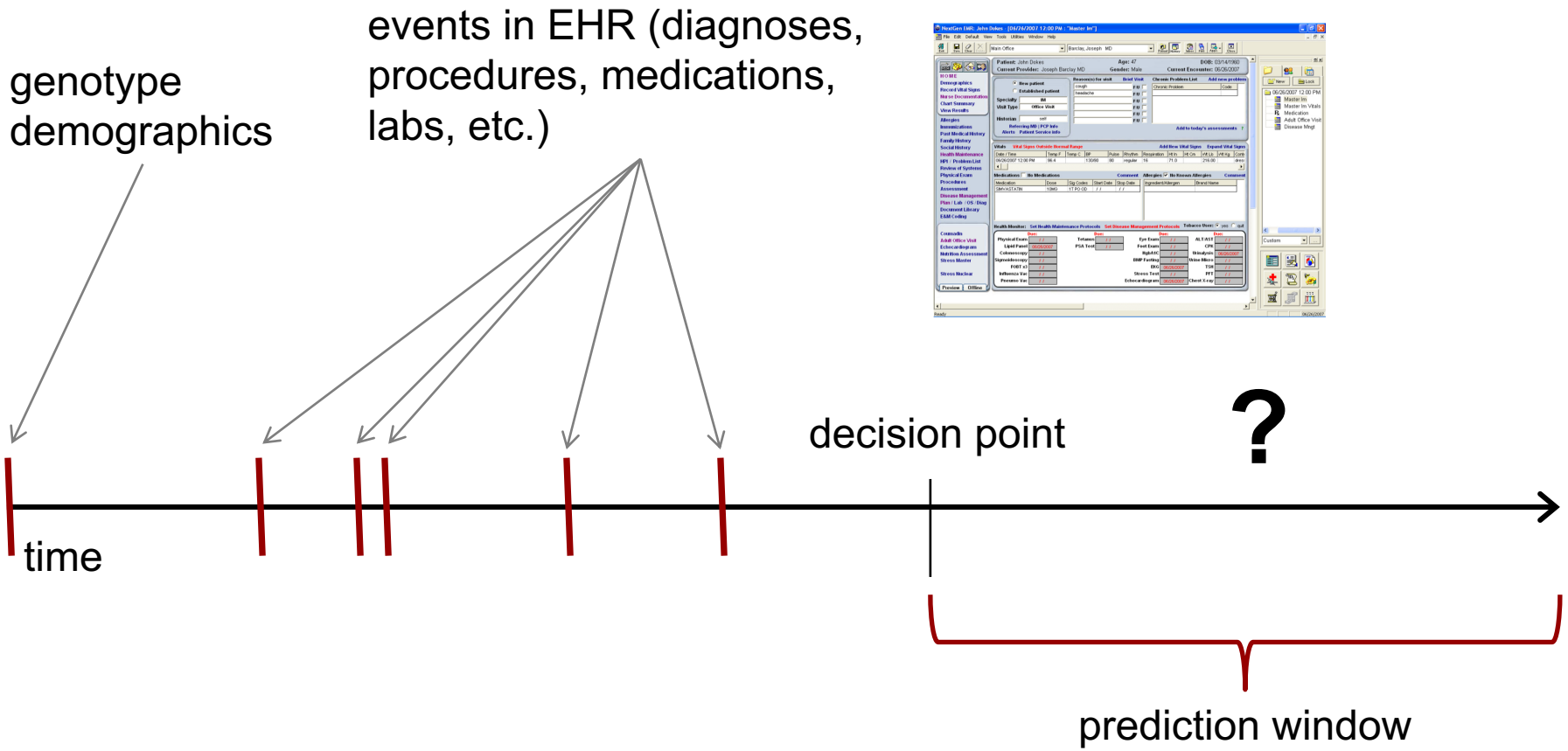
Department of Computer Sciences

University of Wisconsin

Center for Predictive Computational Phenotyping

THE UNIVERSITY of WISCONSIN MADISON

Predicting adverse health outcomes

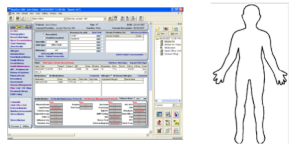


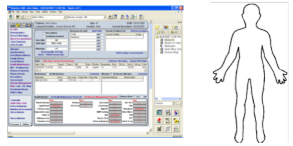
Given a patient's clinical history up to a given point in time

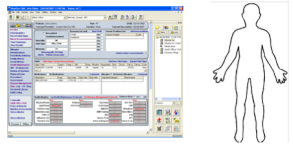
Predict whether the patient will have a specific type of adverse event in the future

Learning task

Can we find a model that accurately represents outcome y as a function of feature vector $\boldsymbol{x}$?



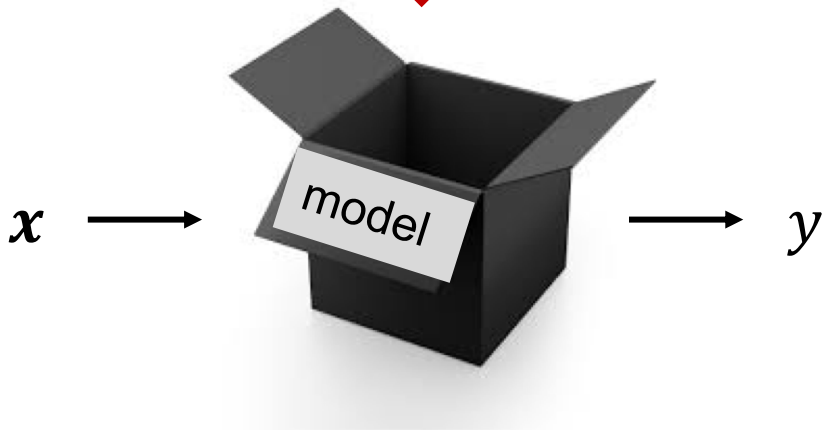




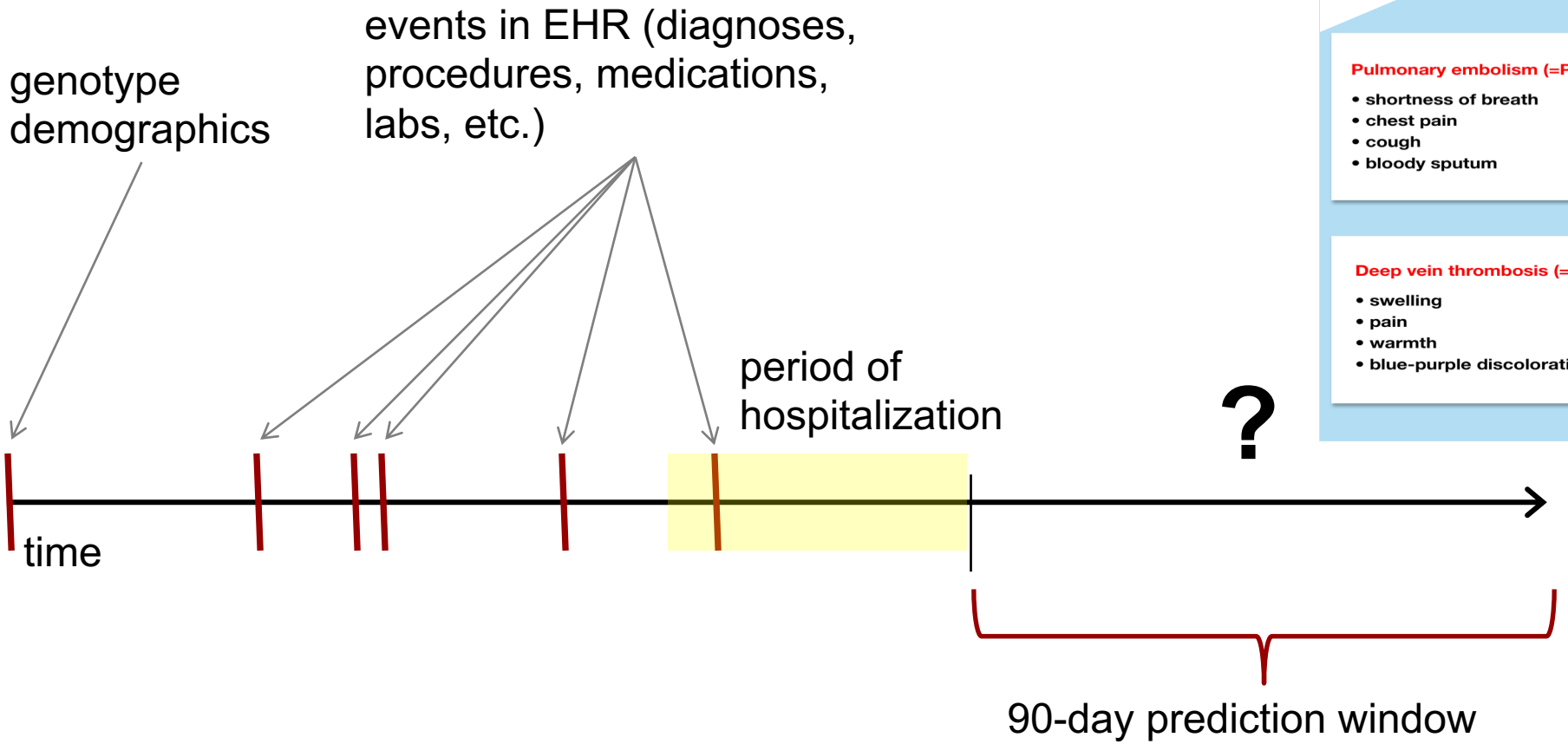
$\boldsymbol{x}^{(1)}$ $y^{(1)}$

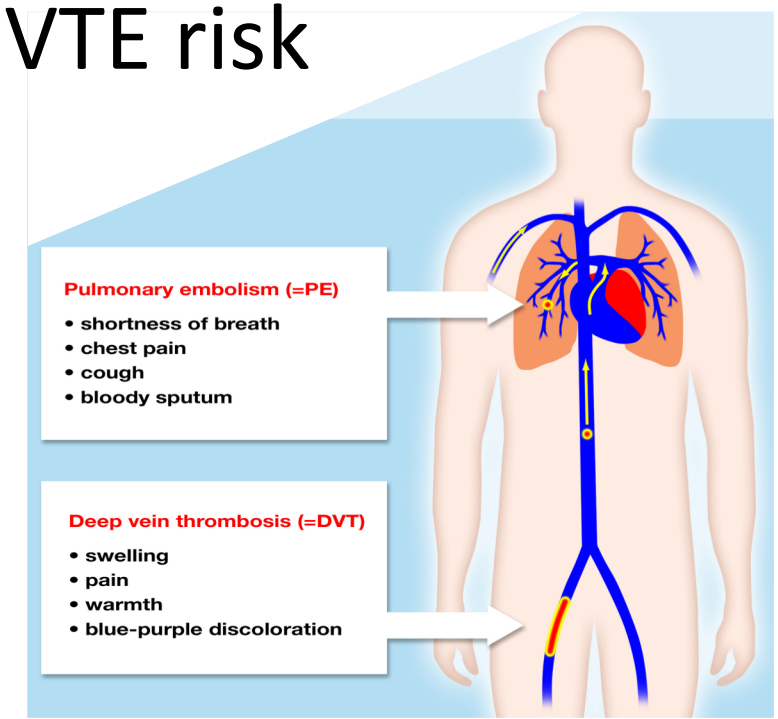
$\boldsymbol{x}^{(2)}$ $y^{(2)}$

$\boldsymbol{x}^{(n)}$ $y^{(n)}$



Example: predicting post-hospitalization VTE risk





Given a patient's clinical history up to EHR up to time of hospitalization

Predict whether the patient is high risk for VTE after release from hospital

Predicting post-hospitalization VTE risk

- 720 subjects
- cases/controls determined by expert review
- two feature sets
  - 119 features based on 78 risk factors for VTE or thrombophilia
  - 3330 features in “unabridged” representation

Emily Kawaler

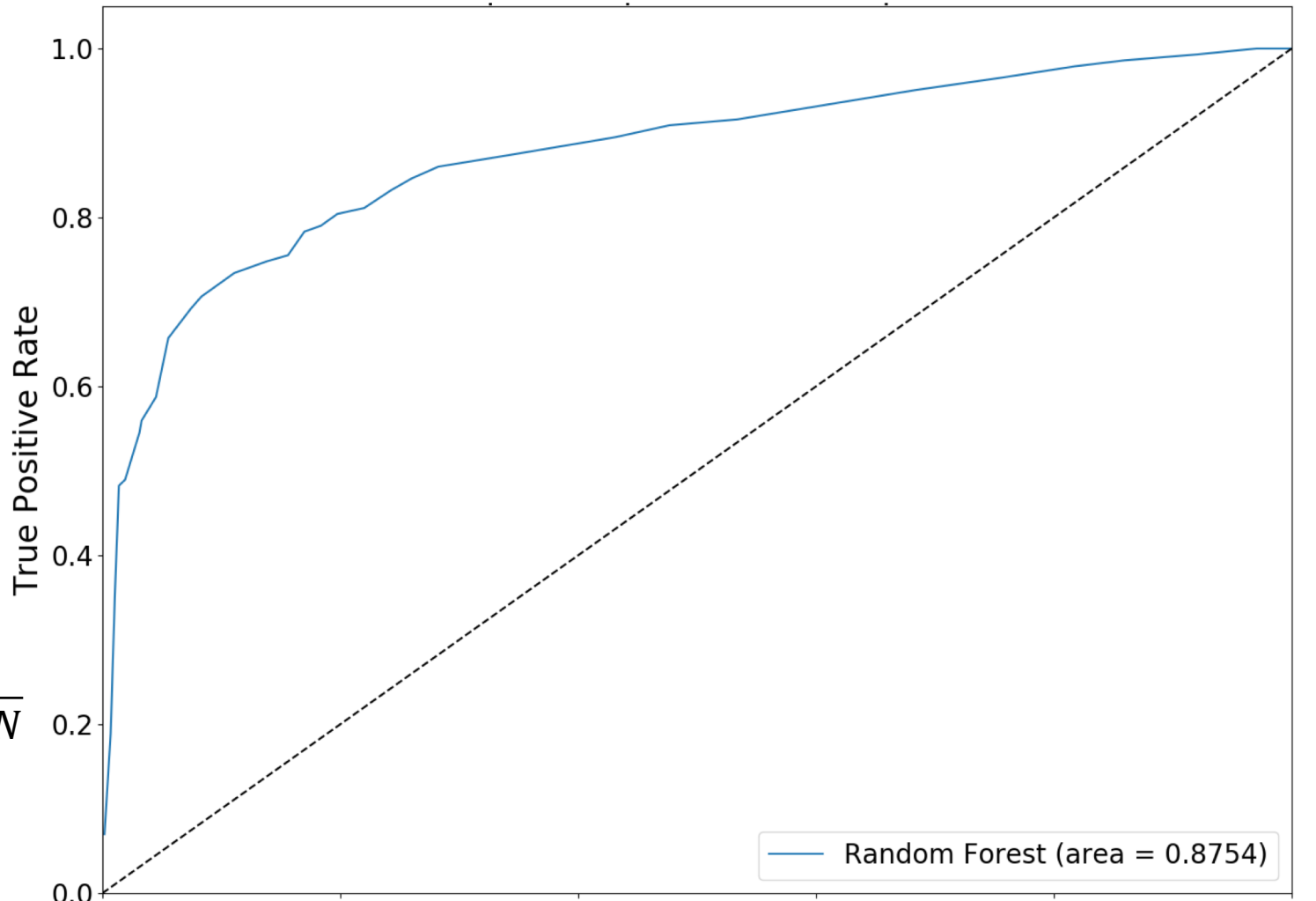
Alex Cobian

Peggy Peissig

Steven Yale

Assessing the accuracy of learned VTE models

[Kawaler et al. *AMIA* 2012]



$$\frac{TP}{TP + FN}$$

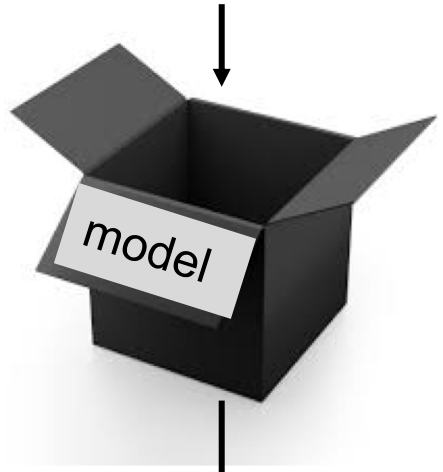
$$\frac{FP}{FP + TN}$$

1. We can learn models that are more accurate than conventional risk assessment tools.

2. Learned models can identify novel risk/protective factors.

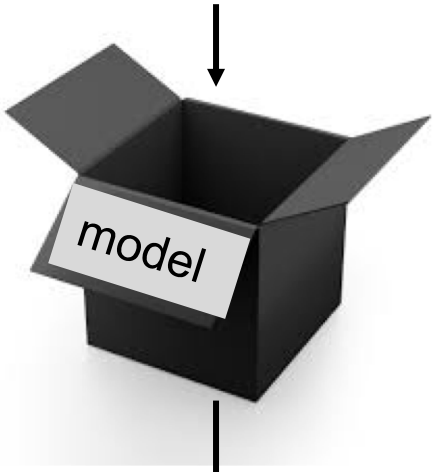
We compare three types of models for VTE task

curated representation
(known risk factors)



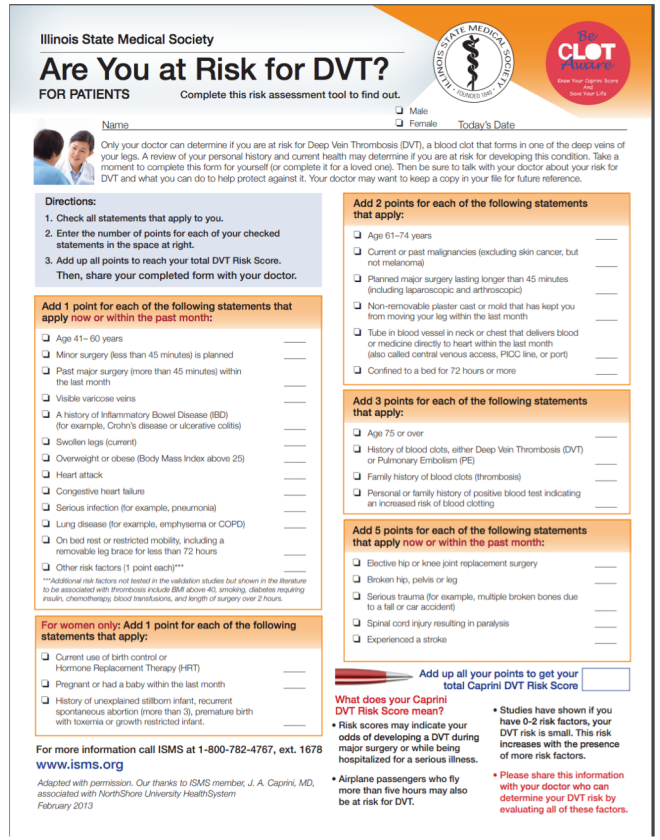
y

unabridged representation
(all features)



y

conventional risk
assessment tools



Illinois State Medical Society

Are You at Risk for DVT?

FOR PATIENTS Complete this risk assessment tool to find out.

❑ Male ❑ Female

Name Today's Date

Only your doctor can determine if you are at risk for Deep Vein Thrombosis (DVT), a blood clot that forms in one of the deep veins of your legs. A review of your personal history and current health may determine if you are at risk for developing this condition. Take a moment to complete this form for yourself (or complete it for a loved one). Then be sure to talk with your doctor about your risk for DVT and what you can do to help protect against it. Your doctor may want to keep a copy in your file for future reference.

Directions:

1. Check all statements that apply to you.
2. Enter the number of points for each of your checked statements in the space at right.
3. Add up all points to reach your total DVT Risk Score. Then, share your completed form with your doctor.

Add 1 point for each of the following statements that apply now or within the past month:

- ❑ Age 41–60 years
- ❑ Minor surgery (less than 45 minutes) is planned
- ❑ Past major surgery (more than 45 minutes) within the last month
- ❑ Visible varicose veins
- ❑ A history of Inflammatory Bowel Disease (IBD) (for example, Crohn's disease or ulcerative colitis)
- ❑ Swollen legs (current)
- ❑ Overweight or obese (Body Mass Index above 25)
- ❑ Heart attack
- ❑ Congestive heart failure
- ❑ Serious infection (for example, pneumonia)
- ❑ Lung disease (for example, emphysema or COPD)
- ❑ On bed rest or restricted mobility, including a removable leg brace for less than 72 hours
- ❑ Other risk factors (1 point each)***

***Additional risk factors not tested in the validation studies but shown in the literature to be associated with thrombosis include BMI above 40, smoking, diabetes requiring insulin, chemotherapy, blood transfusions, and length of surgery over 2 hours.

For women only: Add 1 point for each of the following statements that apply:

- ❑ Current use of birth control or Hormone Replacement Therapy (HRT)
- ❑ Pregnant or had a baby within the last month
- ❑ History of unexplained stillborn infant, recurrent spontaneous abortion (more than 3), premature birth with toxemia or growth restricted infant.

Add 2 points for each of the following statements that apply:

- ❑ Age 61–74 years
- ❑ Current or past malignancies (excluding skin cancer, but not melanoma)
- ❑ Planned major surgery lasting longer than 45 minutes (including laparoscopic and arthroscopic)
- ❑ Non-removable plaster cast or mold that has kept you from moving your leg within the last month
- ❑ Tube in blood vessel in neck or chest that delivers blood or medicine directly to heart within the last month (also called central venous access, PICC line, or port)
- ❑ Confined to a bed for 72 hours or more

Add 3 points for each of the following statements that apply:

- ❑ Age 75 or over
- ❑ History of blood clots, either Deep Vein Thrombosis (DVT) or Pulmonary Embolism (PE)
- ❑ Family history of blood clots (thrombosis)
- ❑ Personal or family history of positive blood test indicating an increased risk of blood clotting

Add 5 points for each of the following statements that apply now or within the past month:

- ❑ Elective hip or knee joint replacement surgery
- ❑ Broken hip, pelvis or leg
- ❑ Serious trauma (for example, multiple broken bones due to a fall or car accident)
- ❑ Spinal cord injury resulting in paralysis
- ❑ Experienced a stroke

Add up all your points to get your total Caprini DVT Risk Score

What does your Caprini DVT Risk Score mean?

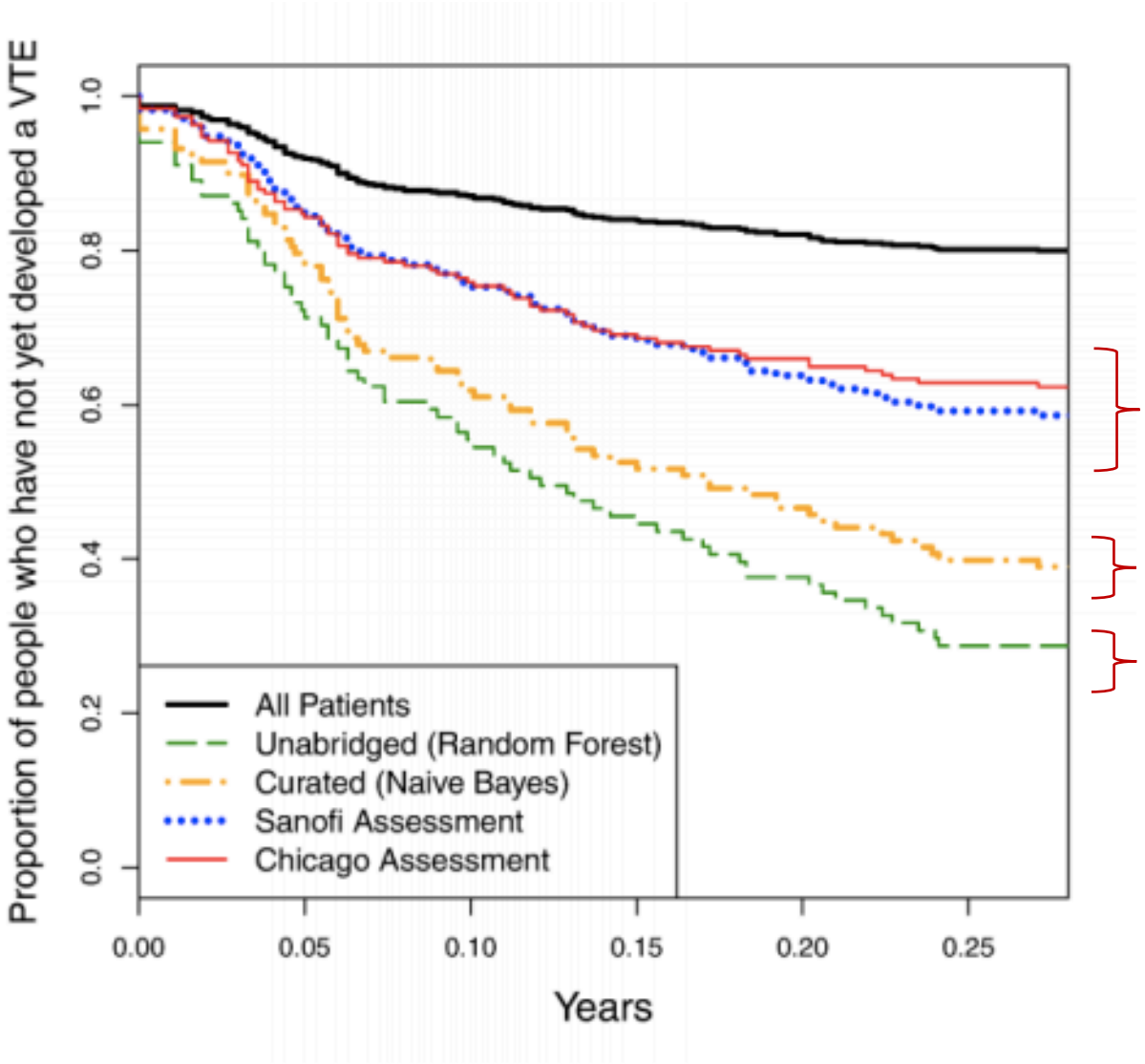
- Risk scores may indicate your odds of developing a DVT during major surgery or while being hospitalized for a serious illness.
- Airplane passengers who fly more than five hours may also be at risk for DVT.
- Studies have shown if you have 0-2 risk factors, your DVT risk is small. This risk increases with the presence of more risk factors.
- Please share this information with your doctor who can determine your DVT risk by evaluating all of these factors.

For more information call ISMS at 1-800-782-4767, ext. 1678
www.isms.org

Adapted with permission. Our thanks to ISMS member, J. A. Caprini, MD, associated with NorthShore University HealthSystem
February 2013

y

Survival curves for models that stratify patients for post-hospitalization VTE risk



conventional risk assessment tools

learned model (known risk factors)

learned model (all features)

Predicting asthma exacerbations

- 28,101 subjects with prior history of asthma
- exacerbations phenotyped by a rule: an urgent visit to a healthcare provider for asthma symptoms followed by treatment with oral corticosteroids
- features represent
  - demographics
  - diagnoses, problem-list diagnoses
  - medications
  - vitals
  - asthma control scores
  - prior exacerbations

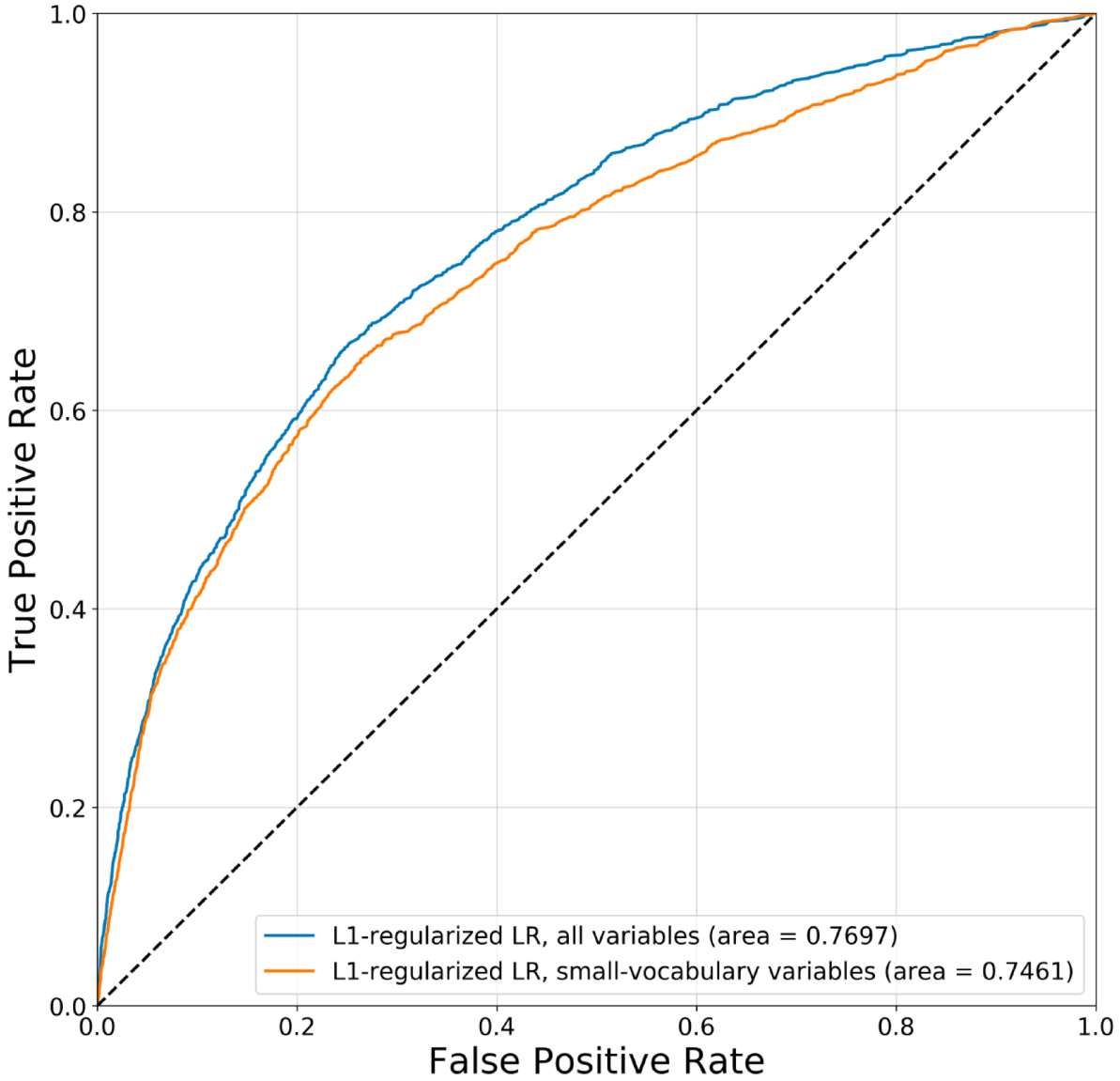
Akshay Sood

Alex Cobian

Theresa Guilbert

Lawrence Hanrahan

Assessing the accuracy of learned exacerbation prediction models



3. In some applications, we can learn more accurate models by using patient genetics in addition to clinical variables.

Predicting breast cancer risk

- 738 subjects for which both genetics and mammogram data was available
- cases/controls determined by cancer registry
- excluded cases with known BRCA1 or BRCA2 mutations
- features
  - 49 variables from BI-RADS lexicon
  - 77 high-frequency/low-penetrance SNPs identified in GWAS


Beth Burnside


Ming Yuan

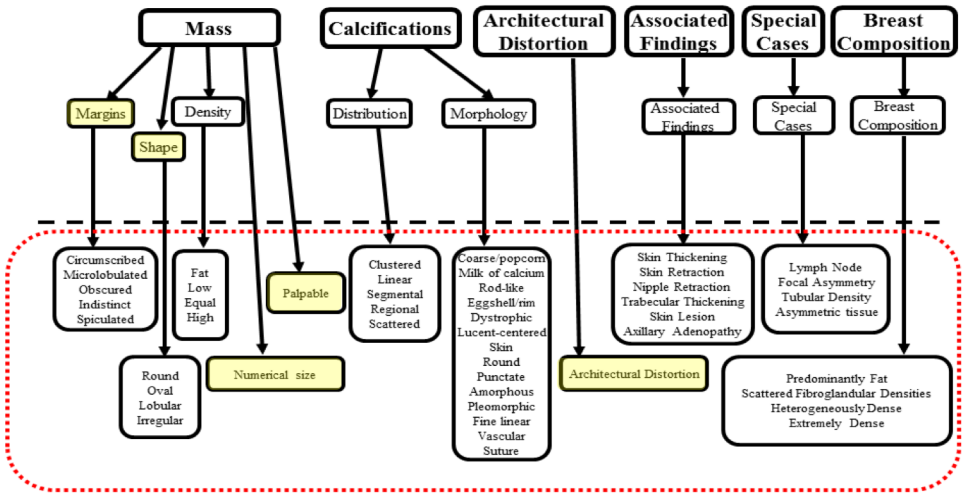

Jun Fan


Shara Feld

Center for Predictive Computational Phenotyping

Improved breast-cancer risk prediction with structure-leveraged methods [Fan et al. *JMLR* 2016]

49 mammography descriptors



learning method optimizes:

$$\min_{(\alpha,\beta)\in R^{d+1}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp[-y_i(\alpha+x_i^T\beta)]) + \lambda_1 \sum_{g=1}^{G} \sqrt{d_g}\left\|\beta_g\right\|_2 + \lambda_2 \sum_{g=1}^{G}\sum_{j=1}^{d_g} \left\|\beta_g^j - \beta_g^{j+1}\right\|_p^p \right\}$$

logistic regression

group lasso (grouping structure)

fused lasso (dependence structure)

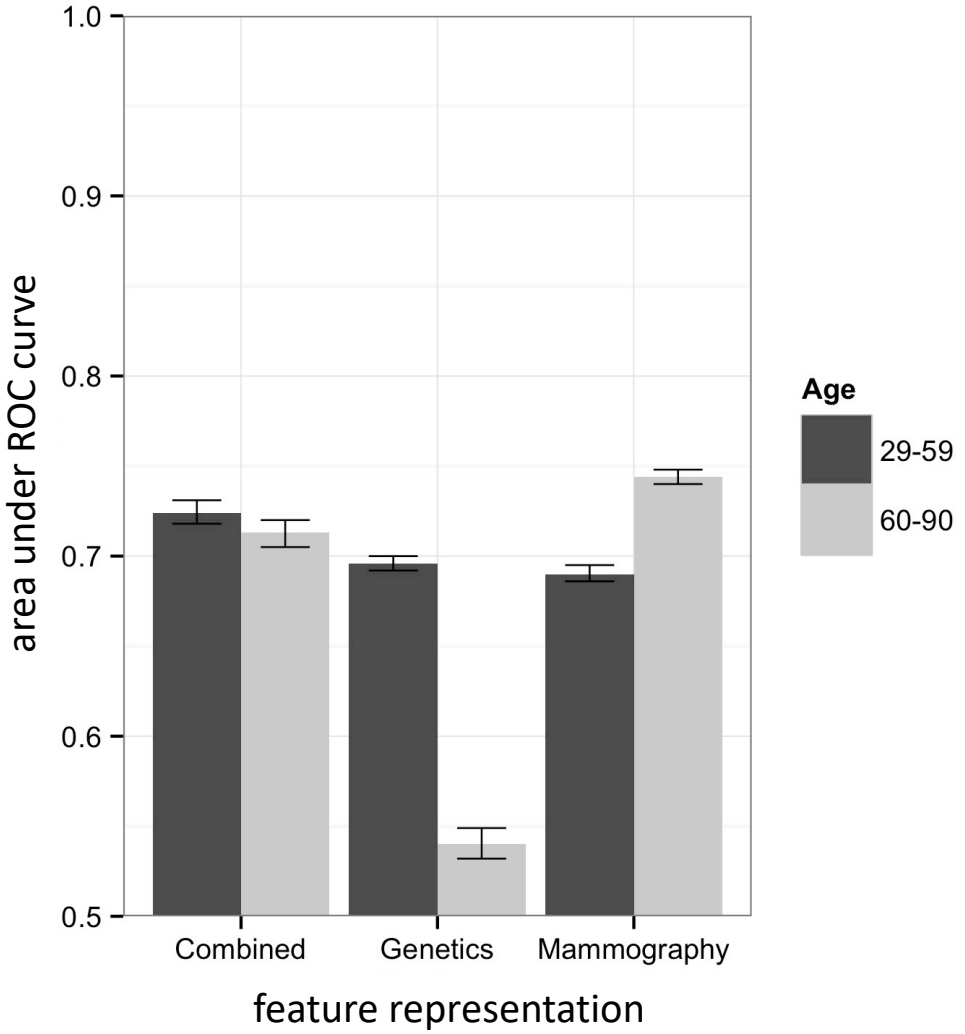
Assessing the accuracy of learned breast-cancer risk prediction models

area under ROC curve

	LR	Structure-leveraged model	*p*-value
49 Mammography features	0.700	0.717	<0.001
77 SNPs	0.590	0.679	<0.001
49 Mammography features and 77 SNPs	0.697	0.765	<0.001

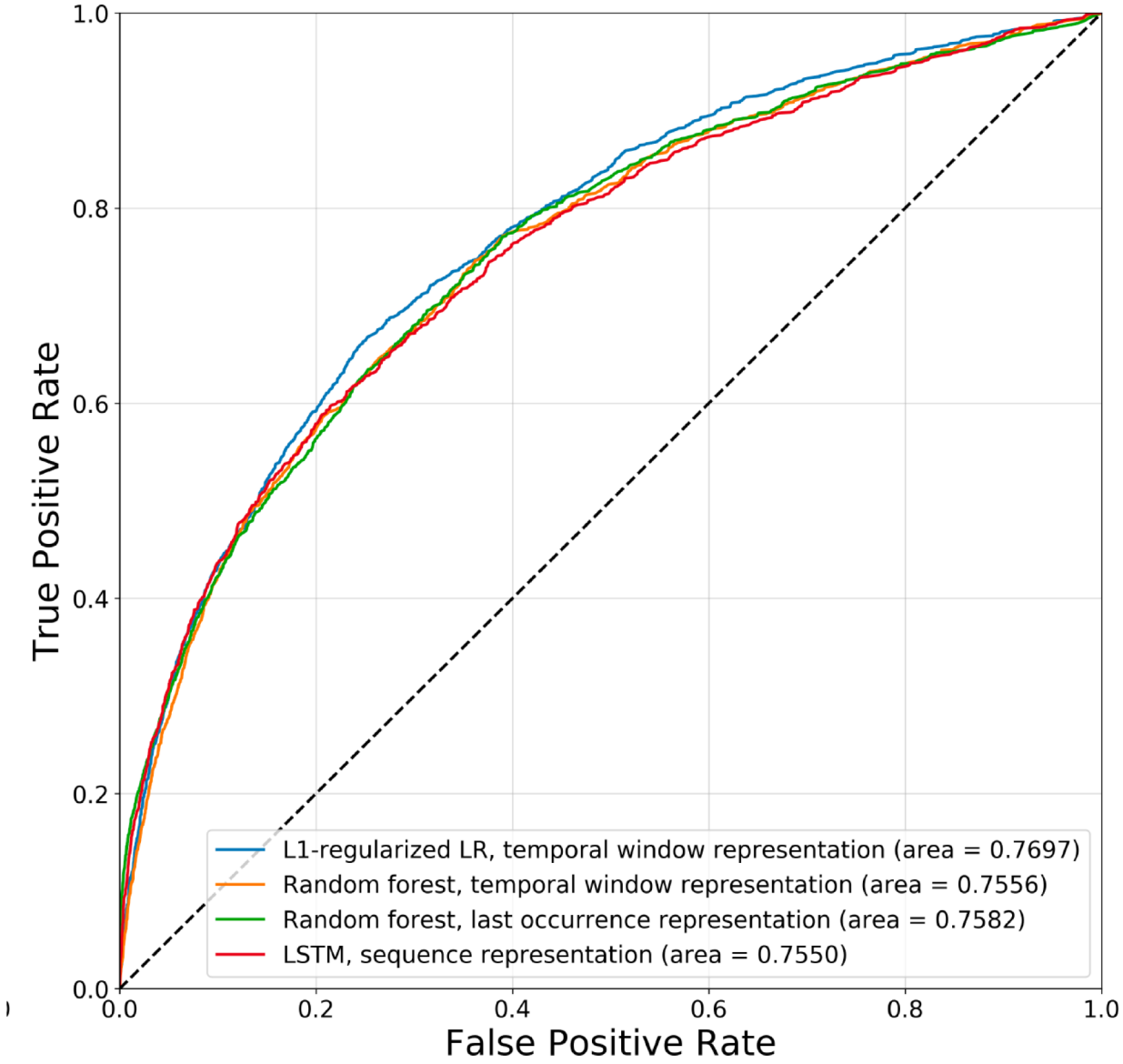
Assessing the accuracy of learned breast-cancer risk prediction models [Feld et al. *AMIA Informatics Summit* 2018]

the predictive value of genetics depends on the age of subjects



4. Simple, linear models are often competitive with fancier models.

Assessing the accuracy of learned exacerbation prediction models



npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar[1,2], Eyal Oren[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Michaela Hardt[1], Peter J. Liu[1], Xiaobing Liu[1], Jake Marcus[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang[1], Yi Zhang[1], Gerardo Flores[1], Gavin E. Duggan[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[3], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[4], Atul J. Butte[2], Michael D. Howell[1], Claire Cui[1], Greg S. Corrado[1] and Jeffrey Dean[1]

Predictive modeling with electronic health record (EHR) data is anticipated to drive personalized medicine and improve healthcare quality. Constructing predictive statistical models typically requires extraction of curated predictor variables from normalized EHR data, a labor-intensive process that discards the vast majority of information in each patient's record. We propose a representation of patients' entire raw EHR records based on the Fast Healthcare Interoperability Resources (FHIR) format. We demonstrate that deep learning methods using this representation are capable of accurately predicting multiple medical events from multiple centers without site-specific data harmonization. We validated our approach using de-identified EHR data from two US academic medical centers with 216,221 adult patients hospitalized for at least 24 h. In the sequential format we propose, this volume of EHR data unrolled into a total of 46,864,534,945 data points, including clinical notes. Deep learning models achieved high accuracy for tasks such as predicting: in-hospital mortality (area under the receiver operator curve [AUROC] across sites 0.93–0.94), 30-day unplanned readmission (AUROC 0.75–0.76), prolonged length of stay (AUROC 0.85–0.86), and all of a patient's final discharge diagnoses (frequency-weighted AUROC 0.90). These models outperformed traditional, clinically-used predictive models in all cases. We believe that this approach can be used to create accurate and scalable predictions for a variety of clinical scenarios. In a case study of a particular prediction, we demonstrate that neural networks can be used to identify relevant information from the patient's chart.

npj Digital Medicine (2018)1:18 ; doi:10.1038/s41746-018-0029-1

ARTICLE OPEN

Scalable and accurate deep learning with electronic health records

"baseline" models are regularized logistic regression; nearly as good as deep networks for all tasks

Supplemental Table 1: Prediction accuracy of each task of deep learning model compared to baselines

	Hospital A	Hospital B
Inpatient Mortality, AUROC[1](95% CI)		
Deep learning 24 hours after admission	**0.95**(0.94-0.96)	**0.93**(0.92-0.94)
Full feature enhanced baseline at 24 hours after admission	0.93 (0.92-0.95)	0.91 (0.89-0.92)
Full feature simple baseline at 24 hours after admission	0.93 (0.91-0.94)	0.90 (0.88-0.92)
Baseline (aEWS[2]) at 24 hours after admission	0.85 (0.81-0.89)	0.86 (0.83-0.88)
30-day Readmission, AUROC (95% CI)		
Deep learning at discharge	**0.77**(0.75-0.78)	**0.76**(0.75-0.77)
Full feature enhanced baseline at discharge	0.75 (0.73-0.76)	0.75 (0.74-0.76)
Full feature simple baseline at discharge	0.74 (0.73-0.76)	0.73 (0.72-0.74)
Baseline (mHOSPITAL[3]) at discharge	0.70 (0.68-0.72)	0.68 (0.67-0.69)
Length of Stay at least 7 days AUROC (95% CI)		
Deep learning 24 hours after admission	**0.86**(0.86-0.87)	**0.85**(0.85-0.86)
Full feature enhanced baseline at 24 hours after admission	0.85 (0.84-0.85)	0.83 (0.83-0.84)
Full feature simple baseline at 24 hours after admission	0.83 (0.82-0.84)	0.81 (0.80-0.82)
Baseline (mLiu[4]) at 24 hours after admission	0.76 (0.75-0.77)	0.74 (0.73-0.75)

5. We can we gain some understanding from complicated learned models (e.g. random forests, neural networks).

Interpreting black-box models

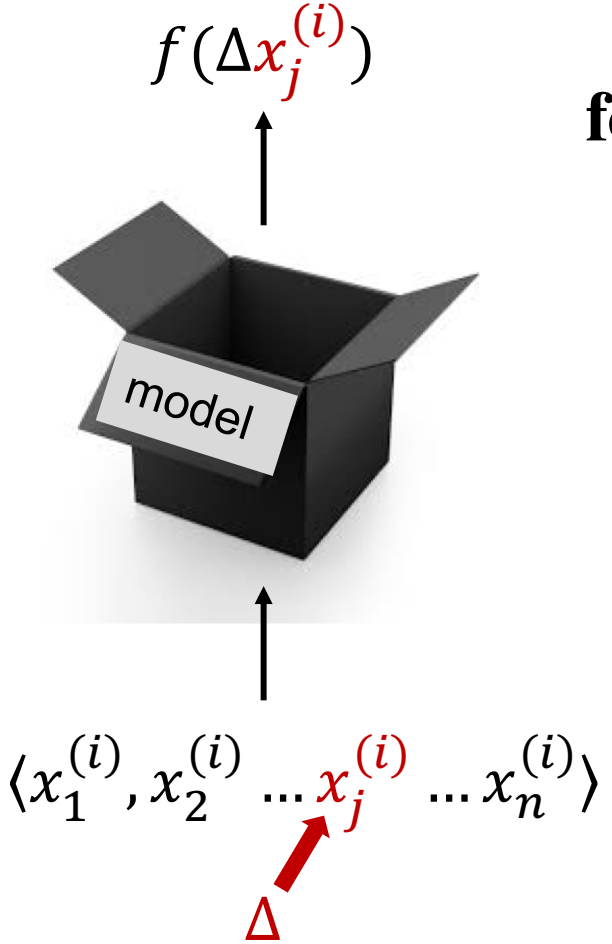
[Craven & Shavlik, *NeurIPS* 1996; Lee et al., *AAAI* 2019]

Kyubin Lee

Akshay Sood

Identifying important features via perturbations

- Feature importance can be ascertained by **perturbing** a feature and measuring its effect on the model loss



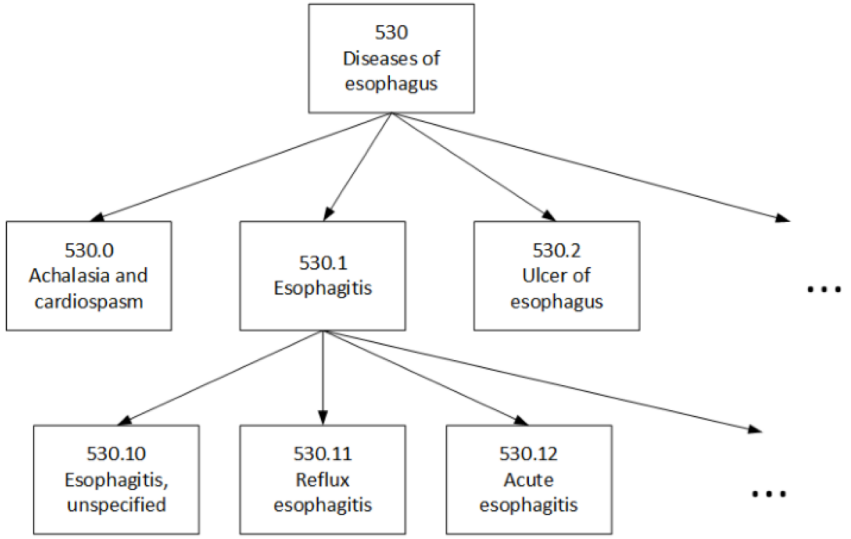
foreach test instance $\left(\mathbf{x}^{(i)}, y^{(i)}\right)$ **do**

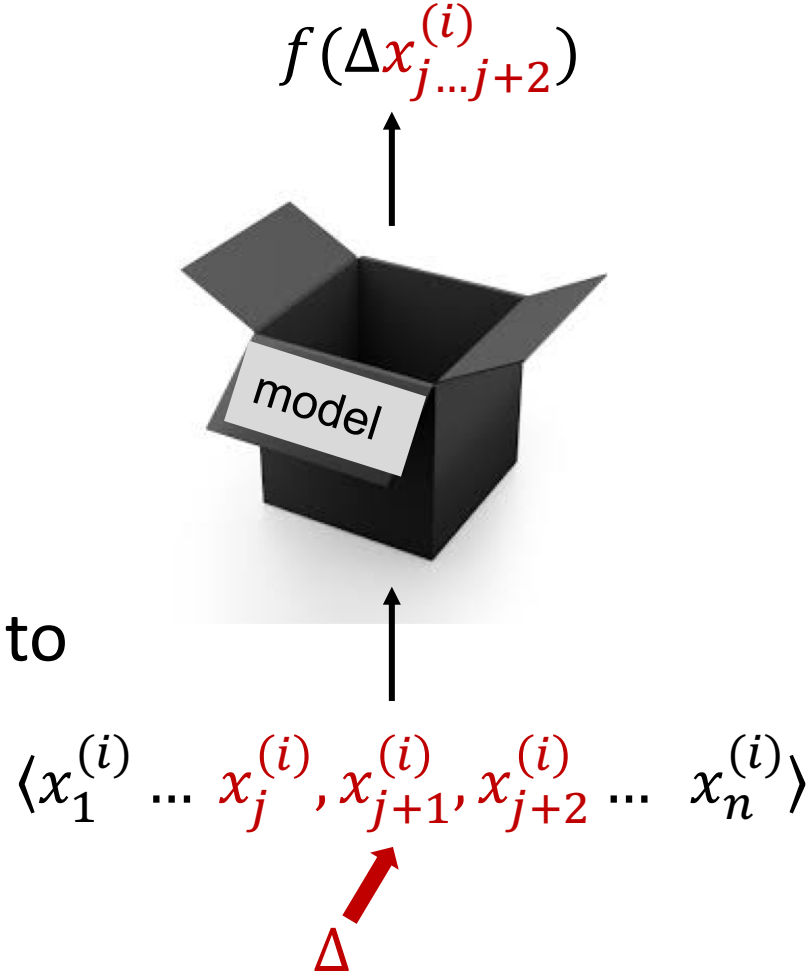
let $\Delta\mathbf{x}_j^{(i)}$ represent $\mathbf{x}^{(i)}$ with feature j perturbed in some way

compare loss $L\left[y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right]$ to $L\left[y^{(i)}, f\left(\Delta\mathbf{x}_j^{(i)}\right)\right]$

From features to hierarchies over features

- In many domains, there is hierarchical structure over features





- We can extend the idea of perturbing base features to perturbing **feature groups**

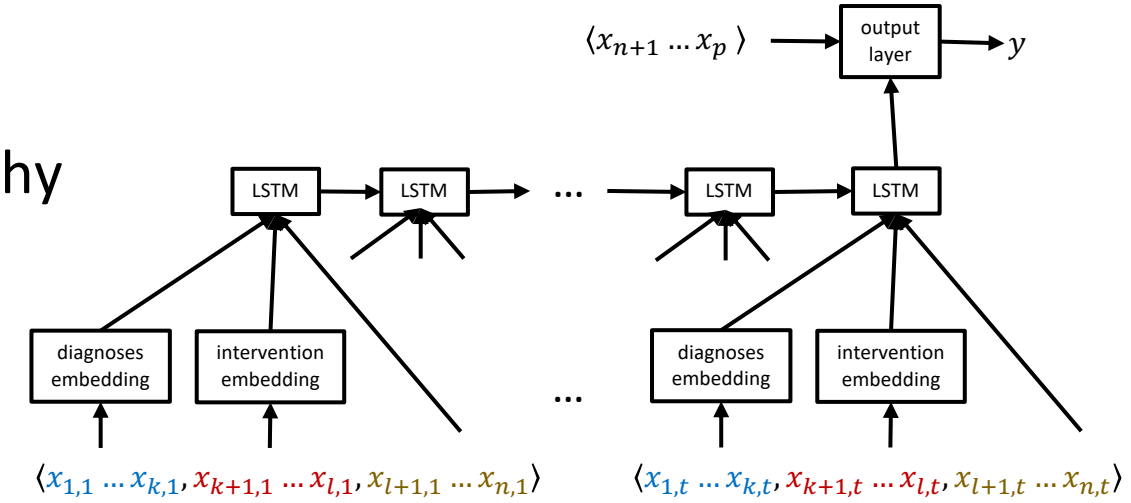
False Discovery Rate Control

- There is a **multiple comparisons problem** due to the large number of base features and feature groups under test

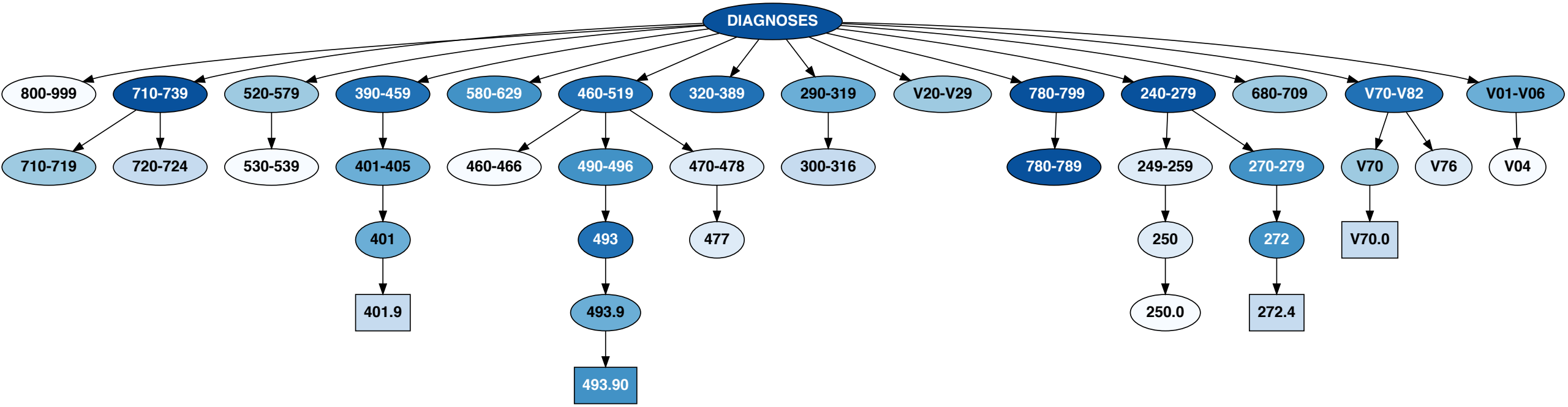
- We use **hierarchical FDR control** [Yekutieli *JASA* 2008] to identify the **finest level of resolution** at which we can find features that have a statistically significant effect on model loss

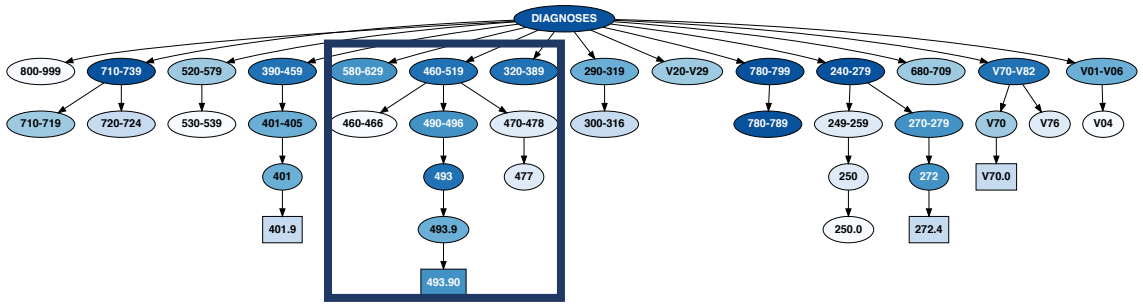
Analysis of learned asthma exacerbation model

1. use standard feature hierarchies (e.g. ICD-9)
2. do perturbations by erasure
3. calculate *p*-values for all nodes in the hierarchy
4. apply hierarchical FDR control



Identifying important features for asthma exacerbation prediction model





DIAGNOSES: ROOT
AUROC: 0.664

460-519: DISEASES OF THE RESPIRATORY SYSTEM
AUROC: 0.730

460-466: ACUTE RESPIRATORY INFECTIONS
AUROC: 0.744

490-496: CHRONIC OBSTRUCTIVE PULMONARY DISEASE AND ALLIED CONDITIONS
AUROC: 0.732

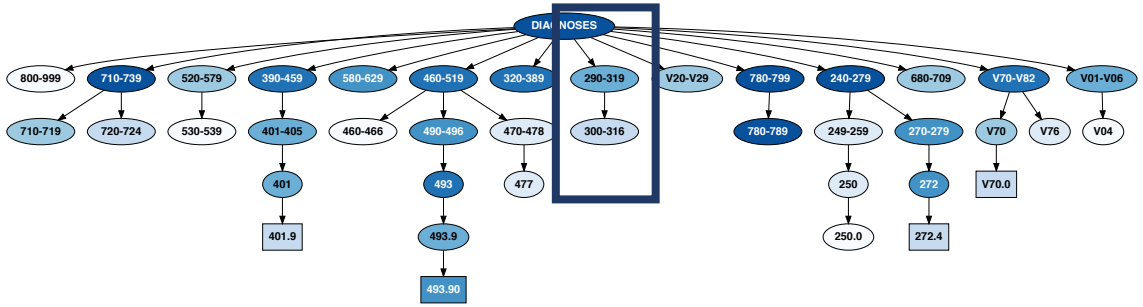
470-478: OTHER DISEASES OF THE UPPER RESPIRATORY TRACT
AUROC: 0.741

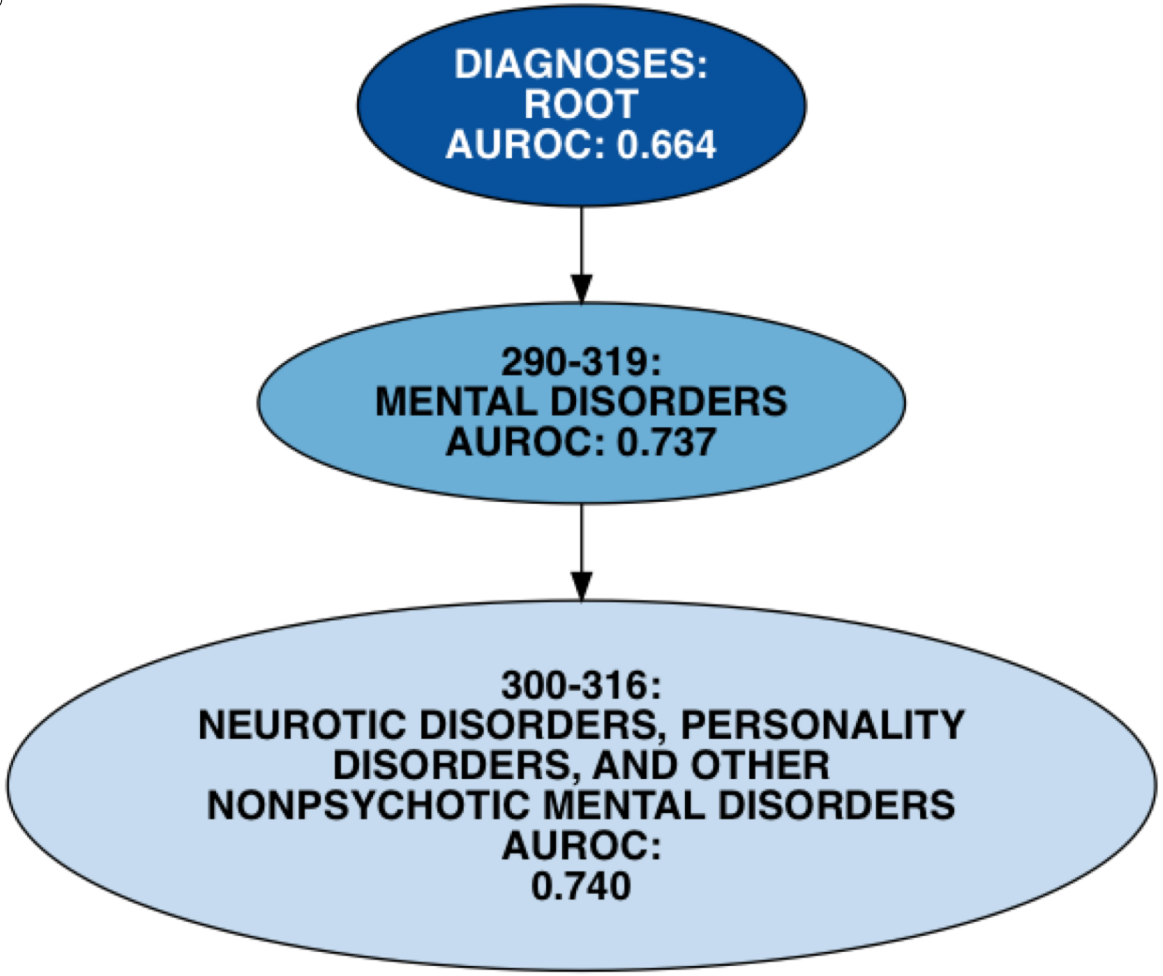
493: Asthma
AUROC: 0.730

477: Allergic rhinitis
AUROC: 0.743

493.9: Asthma, unspecified
AUROC: 0.735

493.90: Asthma, unspecified type, unspecified
AUROC: 0.735


DIAGNOSES
800-999
710-739
520-579
390-459
580-629
460-519
320-389
290-319
V20-V29
780-799
240-279
680-709
V70-V82
V01-V06
710-719
720-724
530-539
401-405
460-466
490-496
470-478
300-316
780-789
249-259
270-279
V70
V76
V04
401
493
477
250
272
V70.0
401.9
493.9
250.0
272.4
493.90

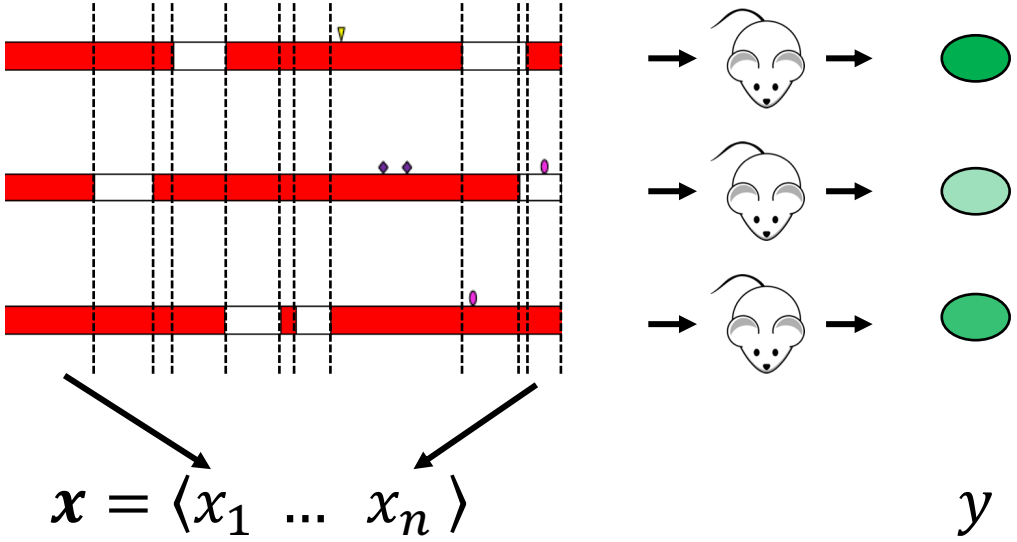

DIAGNOSES:
ROOT
AUROC: 0.664
290-319:
MENTAL DISORDERS
AUROC: 0.737
300-316:
NEUROTIC DISORDERS, PERSONALITY
DISORDERS, AND OTHER
NONPSYCHOTIC MENTAL DISORDERS
AUROC:
0.740

Conclusions

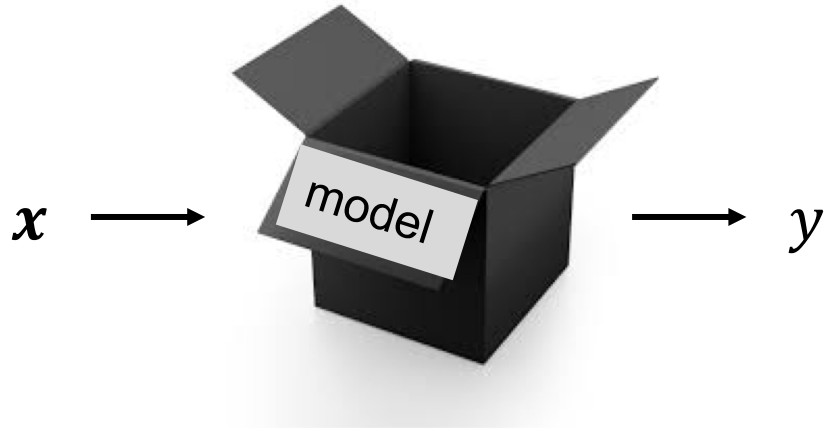
- learned models are sometimes more accurate than conventional risk assessment tools
- learned models can identify novel risk/protective factors
- genetics can augment predictive value of clinical data in some cases
- simple models often work well
- but we have tools to gain insight from complex black-box models

NIH U54 AI117924
NIH T32 HG002760
NIH T15 LM007359
NIH R01 EY023292
NSF IIS 1218880

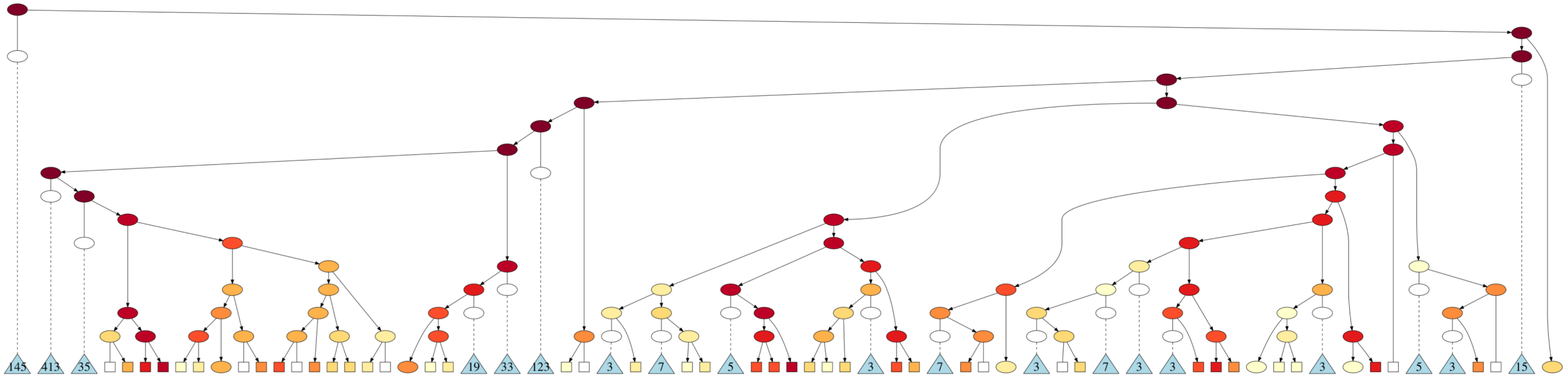
Learning associations between HSV-1 genotype and host phenotype [Lee et al. *J. Virology* 2016; Kolb et al. *PLoS Pathogens* 2016]



- 69 viral genomes
- features represent 547 haplotype blocks
- we have learned regression models using Lasso, ridge, and random forest approaches



Analysis of learned HSV-1 genotype-to-phenotype models



large parts of feature space are pruned from hypothesis testing

other important features can only be determined at group resolution

some ”base” features are important

important feature

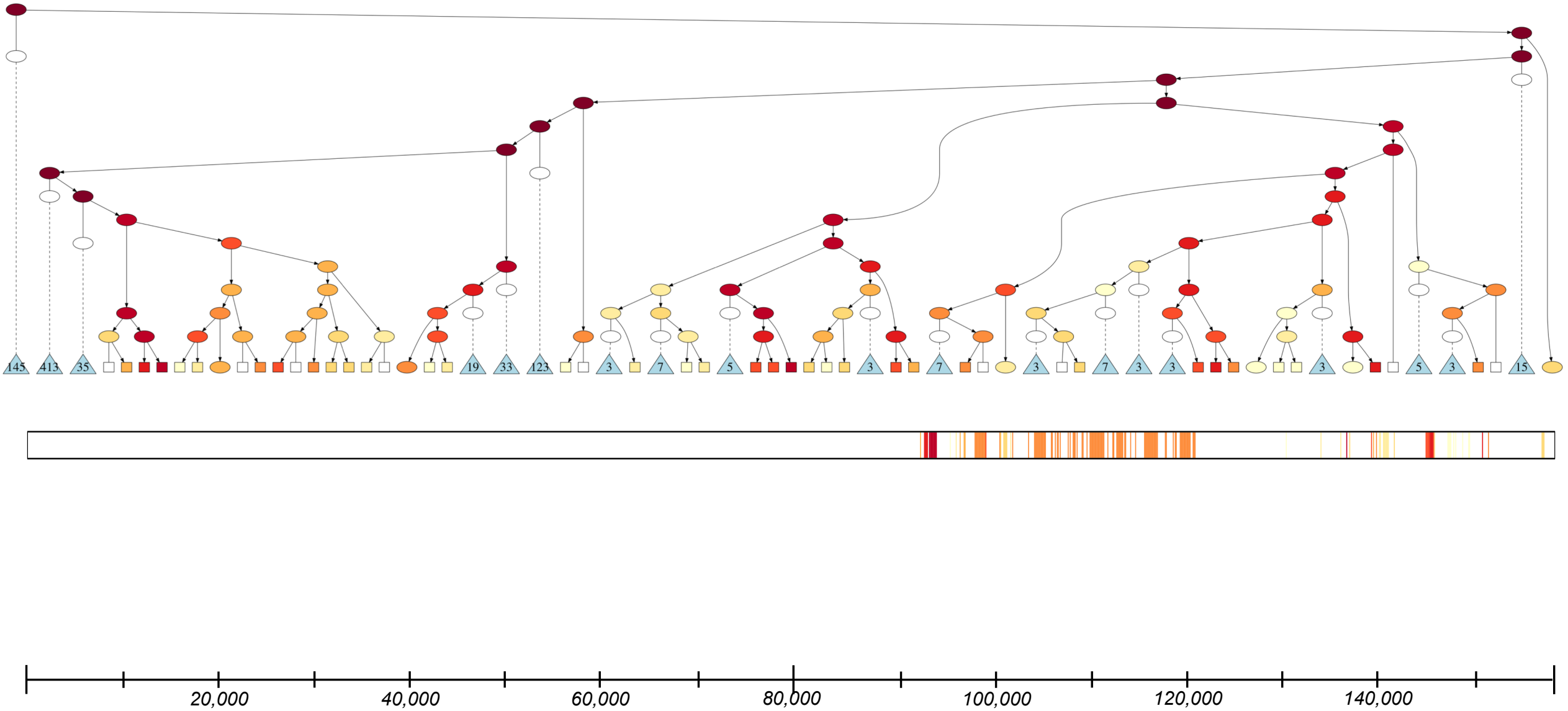
important feature group

pruned subtree

features/feature groups

- 40 “outer” features/feature groups
- 34 base features

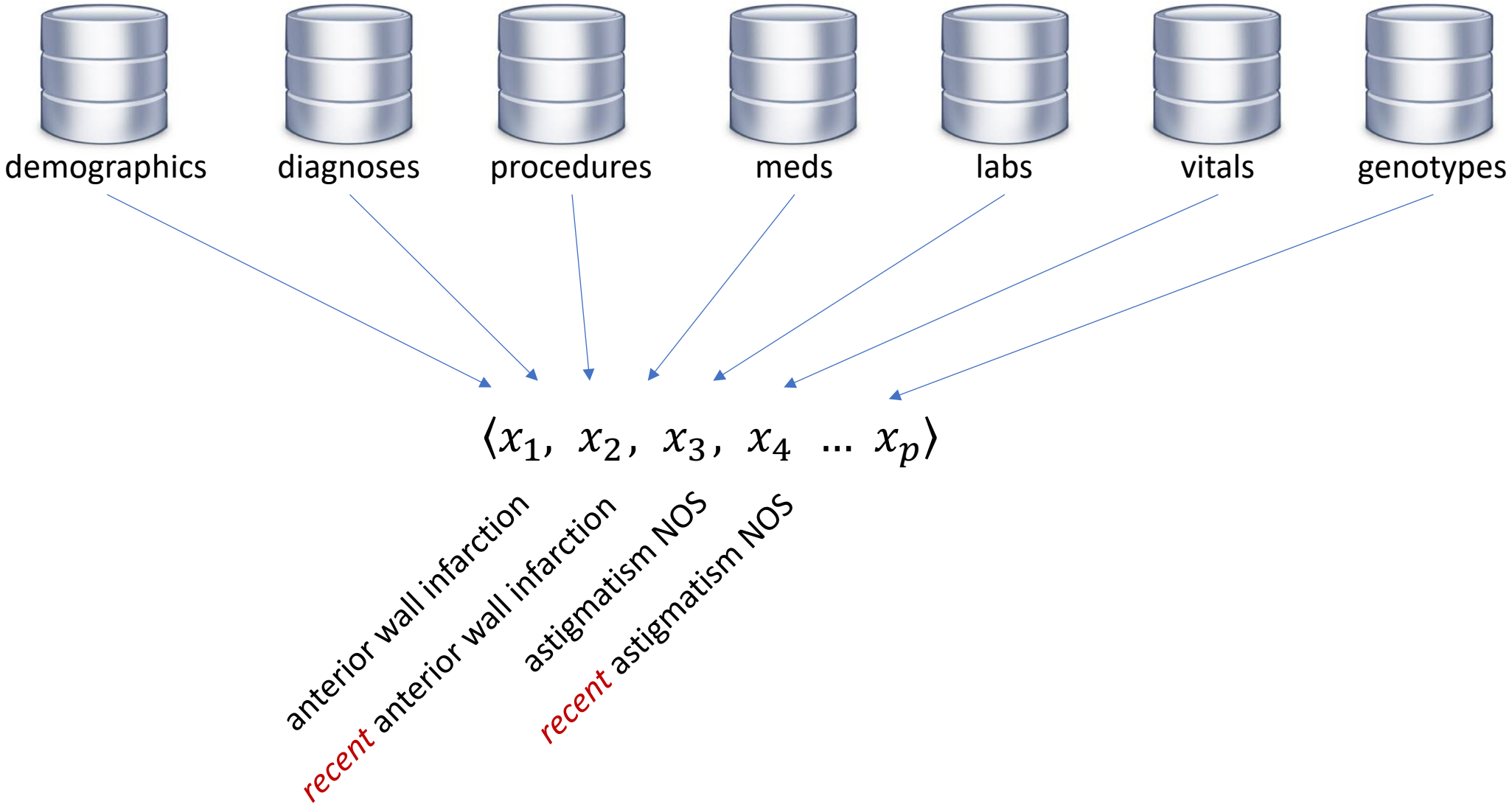
Important variables are localized in the viral genome



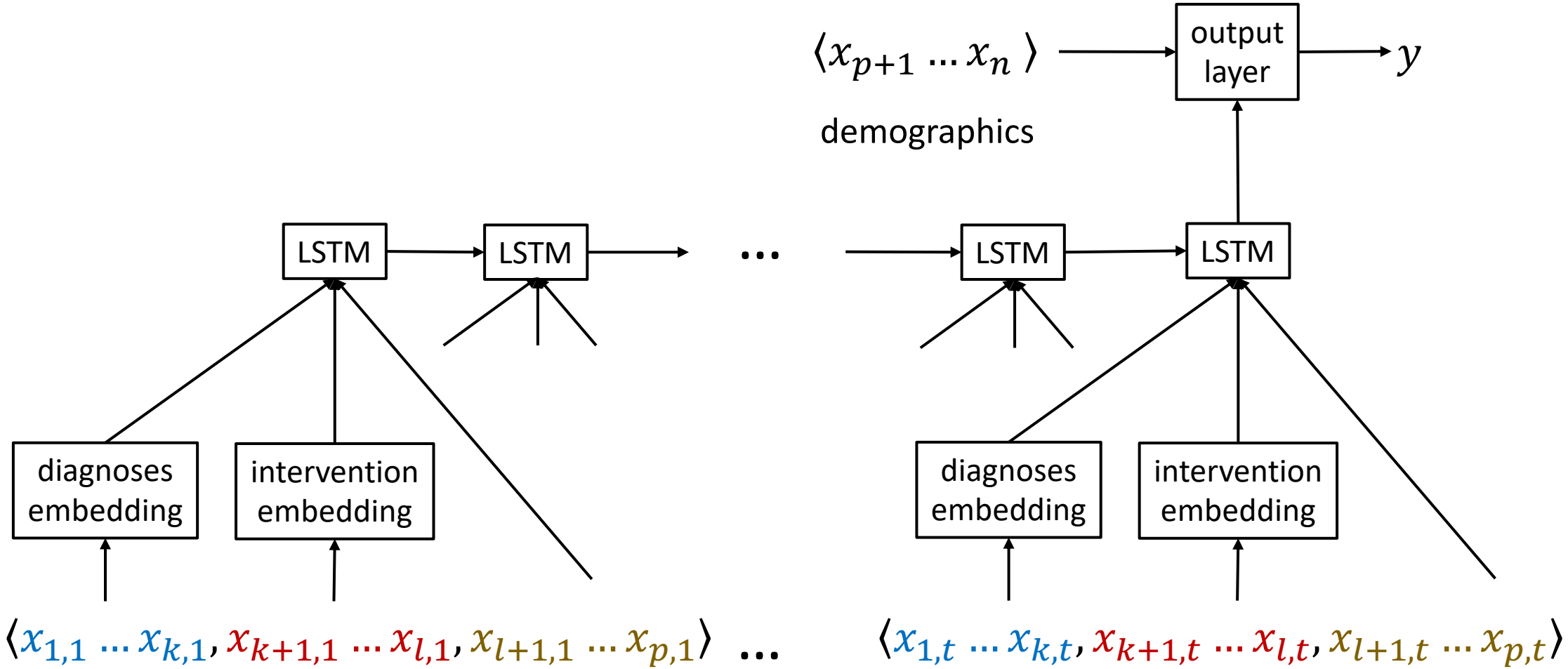
Using machine-learning to predict adverse outcomes

1. select cohort of subjects
2. retrospectively phenotype outcomes (e.g. determine who is a case/control)
3. define feature representation
4. learn models
5. evaluate models

Defining feature representations



An LSTM neural network for asthma exacerbations

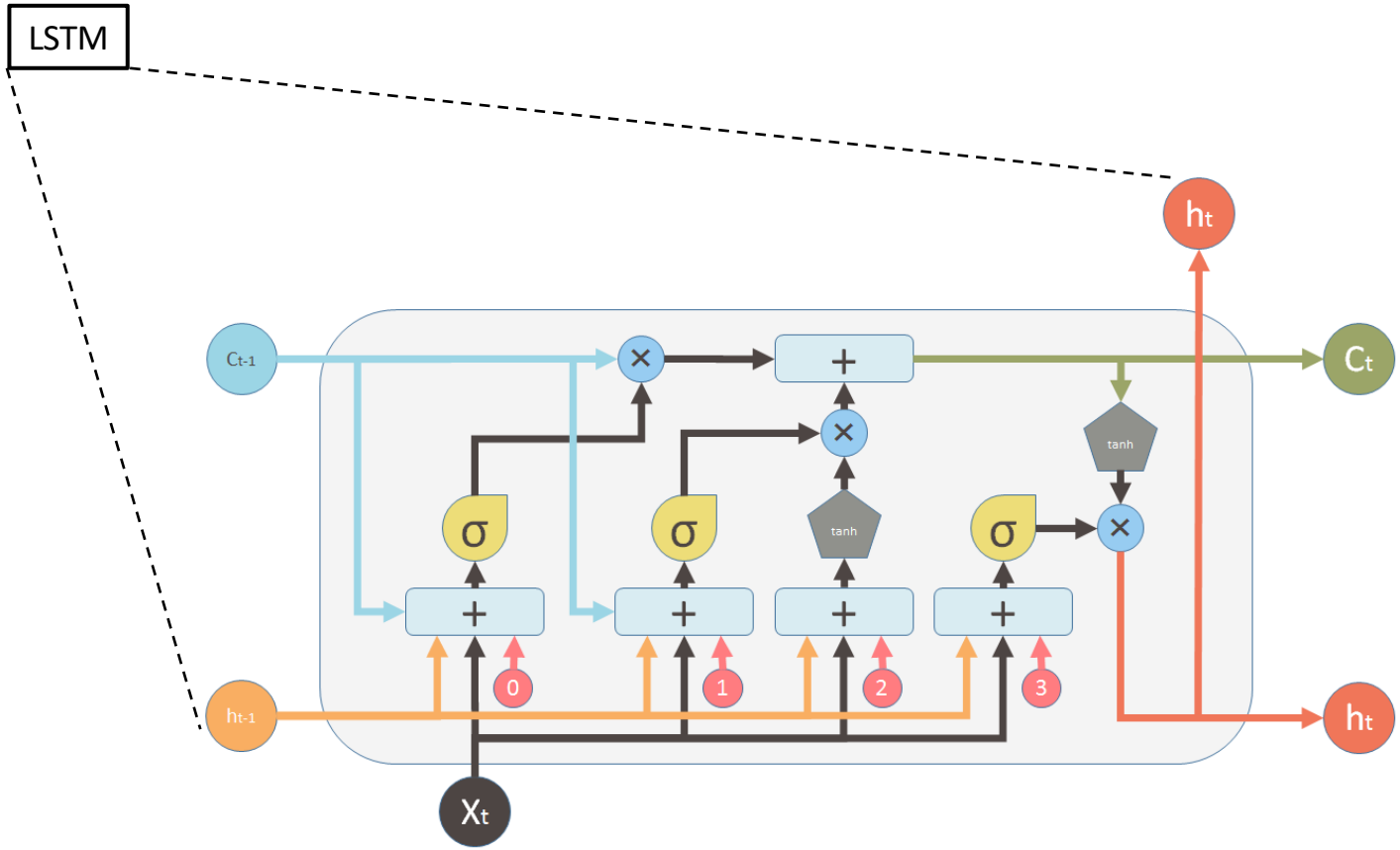


diagnoses

interventions (meds + procedures)

vitals, prior exacerbations, etc.

An LSTM neural network model



Identifying important features via perturbation

- We can use **hypothesis testing** to characterize the effect of perturbing a feature

Instance	Perturbed Instance	Loss	Perturbed Loss
$(\mathbf{x}^{(1)}, y^{(1)})$		$L\left[y^{(1)}, f\left(\mathbf{x}^{(1)}\right)\right]$	
$(\mathbf{x}^{(2)}, y^{(2)})$		$L\left[y^{(2)}, f\left(\mathbf{x}^{(2)}\right)\right]$	
...		...	...

Null hypothesis: Median change in loss when perturbing feature j is 0