

# Personalized VTE Risk Prediction in Cancer Patients using Clinical Informatics

---

Tuesday | October 10, 2023 | 12:00pm ET

# Presenters



**Ang Li, MD, MS**  
Assistant Professor  
*Baylor College of Medicine*



**Kristen Sanfilippo, MD**  
Associate Professor of Medicine,  
Staff Physician  
*Washington University of  
Medicine in St. Louis; St. Louis VA  
Medical Center*



**Marc Carrier, MD, MSc, FRCPC**  
Head, Division of Hematology,  
Department of Medicine  
Professor, Faculty of Medicine  
*The Ottawa Hospital; University of  
Ottawa*



**Jean Connors, MD**  
Medical Director, Hemostatic  
Antithrombotic Stewardship  
Medical Director,  
Anticoagulation Management  
Services  
Hematology Division  
*Brigham and Women's Hospital  
/ Dana-Farber Cancer Institute*  
Professor of Medicine  
*Harvard Medical School*

# **Personalized VTE Risk Prediction in Cancer Patients using Clinical Informatics**

Ang Li, MD, MS

Assistant Professor

Section of Hematology-Oncology

Baylor College of Medicine

October 10, 2023

# Disclosure

- None

# Objectives

- Overview the advantages and limitations of modern curated electronic health record (EHR) research in cancer and thrombosis
- Provide case study on natural language processing (NLP) algorithms in classifying unstructured text for venous thromboembolism (VTE)
- Provide case study on derivation and validation of risk prediction models of VTE among cancer patients
- Provide case study on implementing patient centered clinical decision support (PC-CDS) tools for pharmacologic thromboprophylaxis

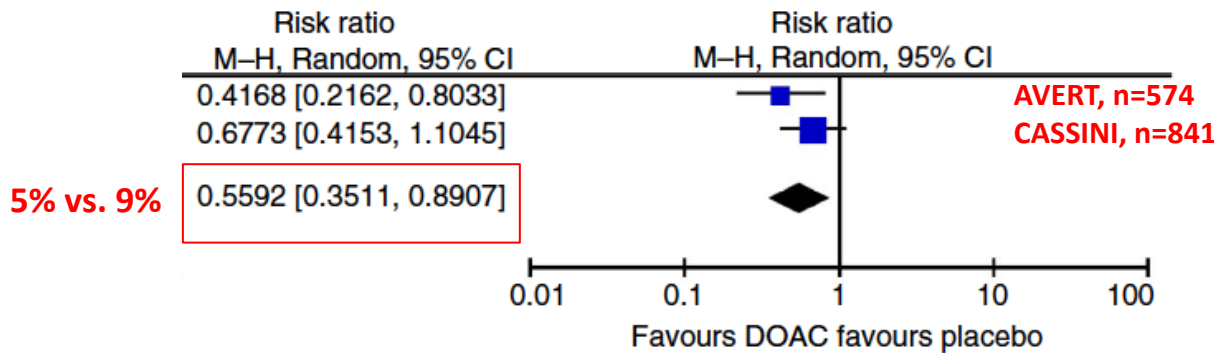
# I. Introduction of Cancer Associated Thrombosis (CAT)

# Importance of VTE Prediction in Cancer Patients

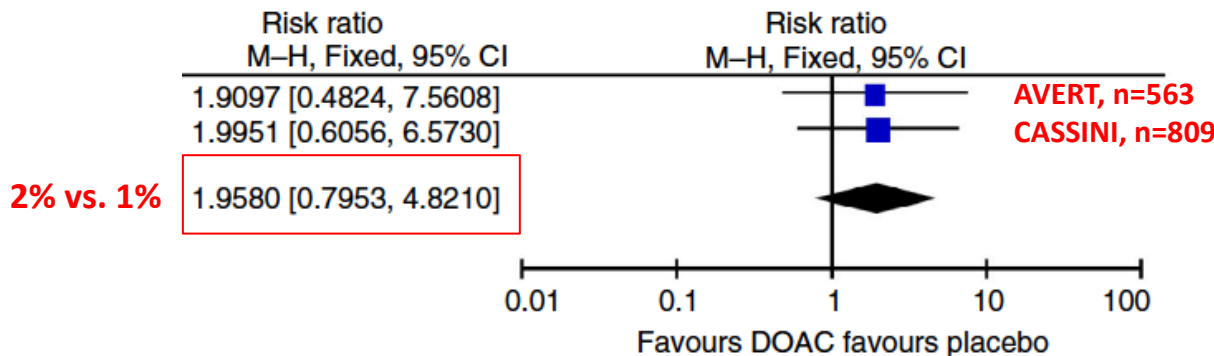
- VTE occurs 7-9 times more in cancer vs. non-cancer patients
- Incidence of VTE varies significantly by cancer type
- Thrombosis (venous + arterial) is 2nd leading cause of death in ambulatory patients with cancer along with infection (9%)
- Patients with active cancer have a one-year mortality of 65% after VTE diagnosis

# Data Supporting VTE Prevention in Selective High-risk Cancer Patients

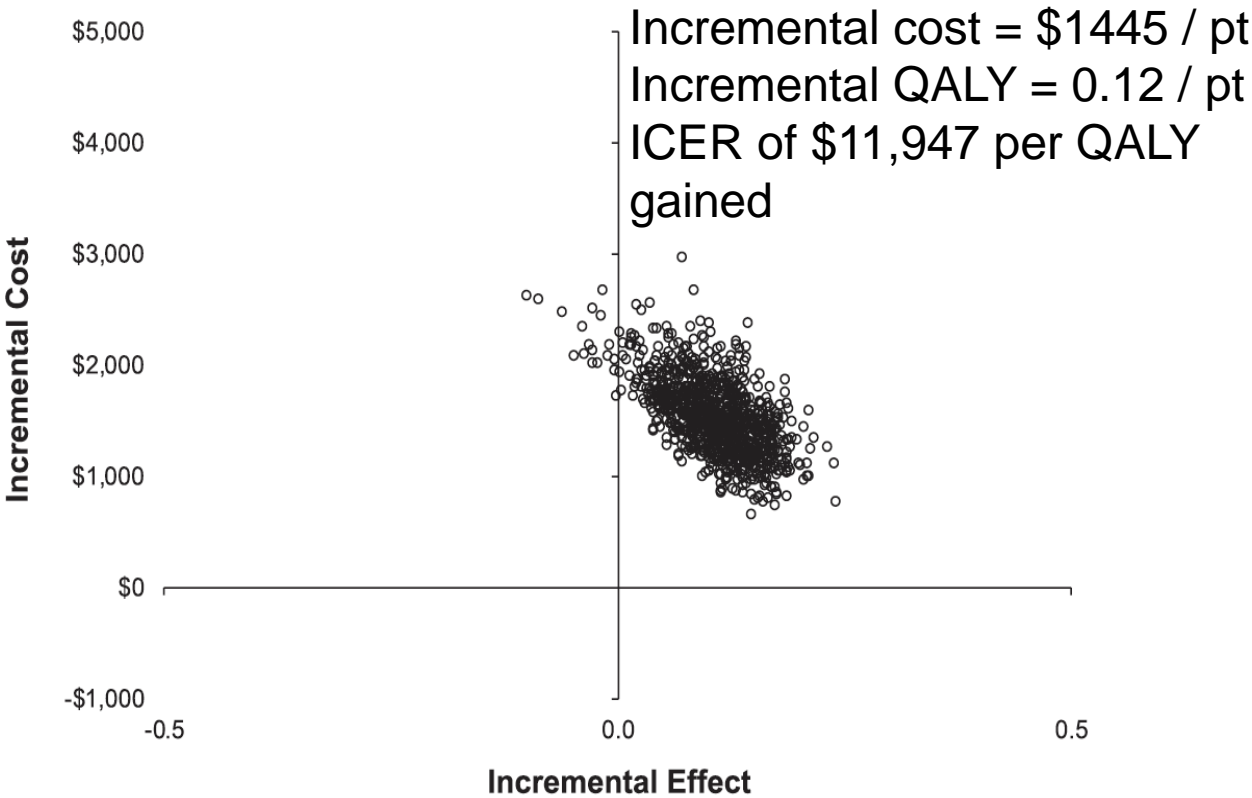
## Incident VTE at 6 months for Low-Dose DOAC vs. Placebo



## Major Bleeding at 6 months for Low-Dose DOAC vs. Placebo



Meta-analysis of randomized controlled trials enrolling patients with Khorana Score 2+



Cost Effectiveness Analysis



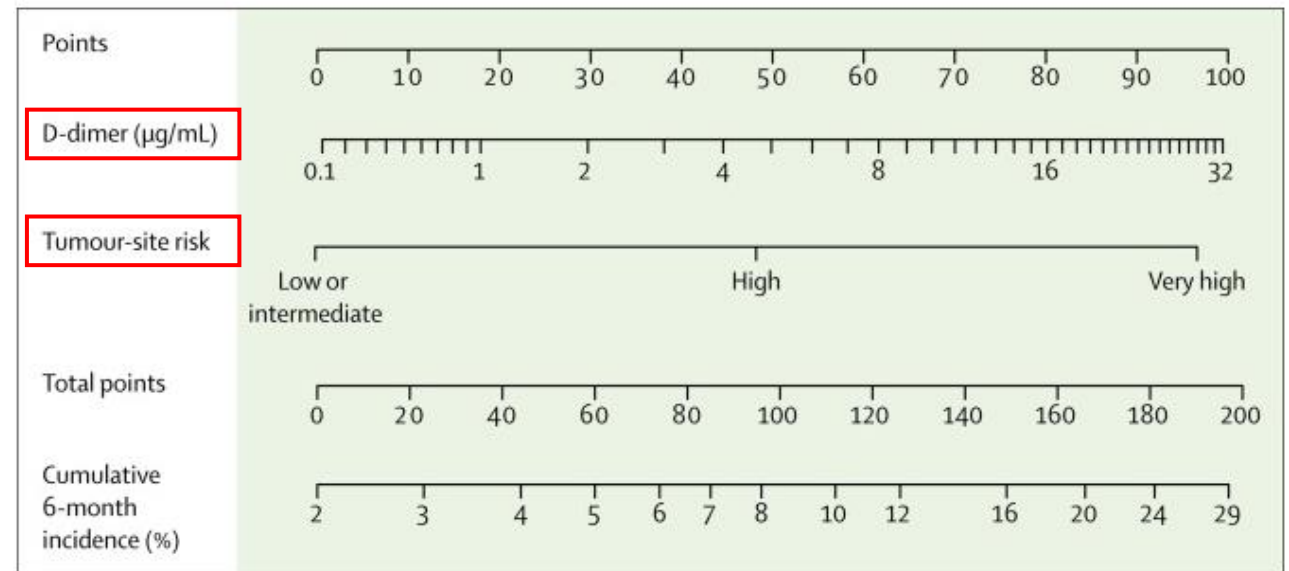
# Existing Risk Models in Cancer Associated Thrombosis

Khorana Score, Blood 2008

Variable	Score
Very high-risk tumor (stomach, pancreas)	2
High-risk tumor (lung, gynecologic, genitourinary excluding prostate)	1
Hemoglobin level <100 g/L or use of red cell growth factors	1
Prechemotherapy leukocyte count $>11 \times 10^9/\text{L}$	1
Prechemotherapy platelet count $350 \times 10^9/\text{L}$ or greater	1
Body mass index $35 \text{ kg/m}^2$ or greater	1

A score of 0 = low-risk category. A score of 1–2 = intermediate-risk category. A score of  $>2$  = very high-risk category.

Pabinger nomogram, Lancet Haematology 2018



Only ~50% of VTE is classified as high-risk

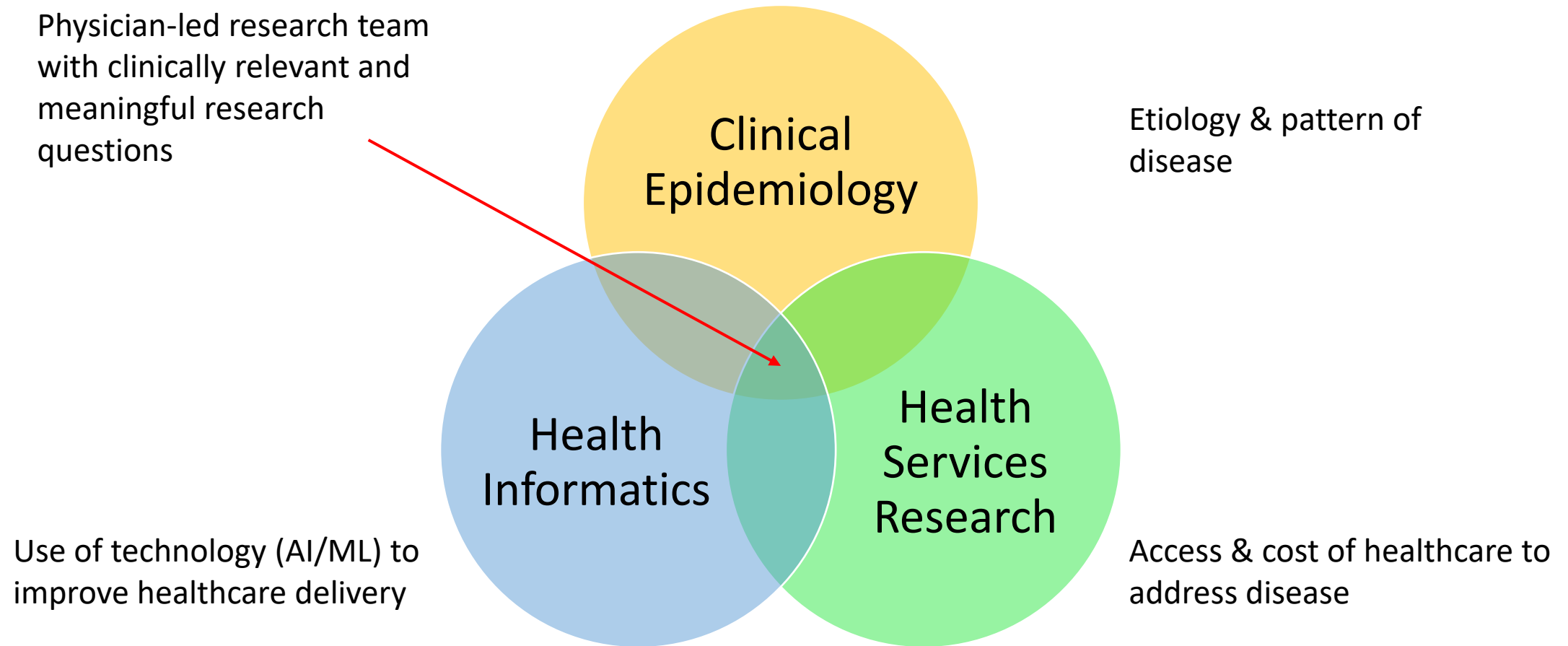
Difficult to incorporate non-standard biomarker

**Key: Khorana Score is the most commonly used clinical risk model. D-dimer is the most commonly used biomarker**

# Ambulatory Pharmacologic Prophylaxis is Rarely Implemented

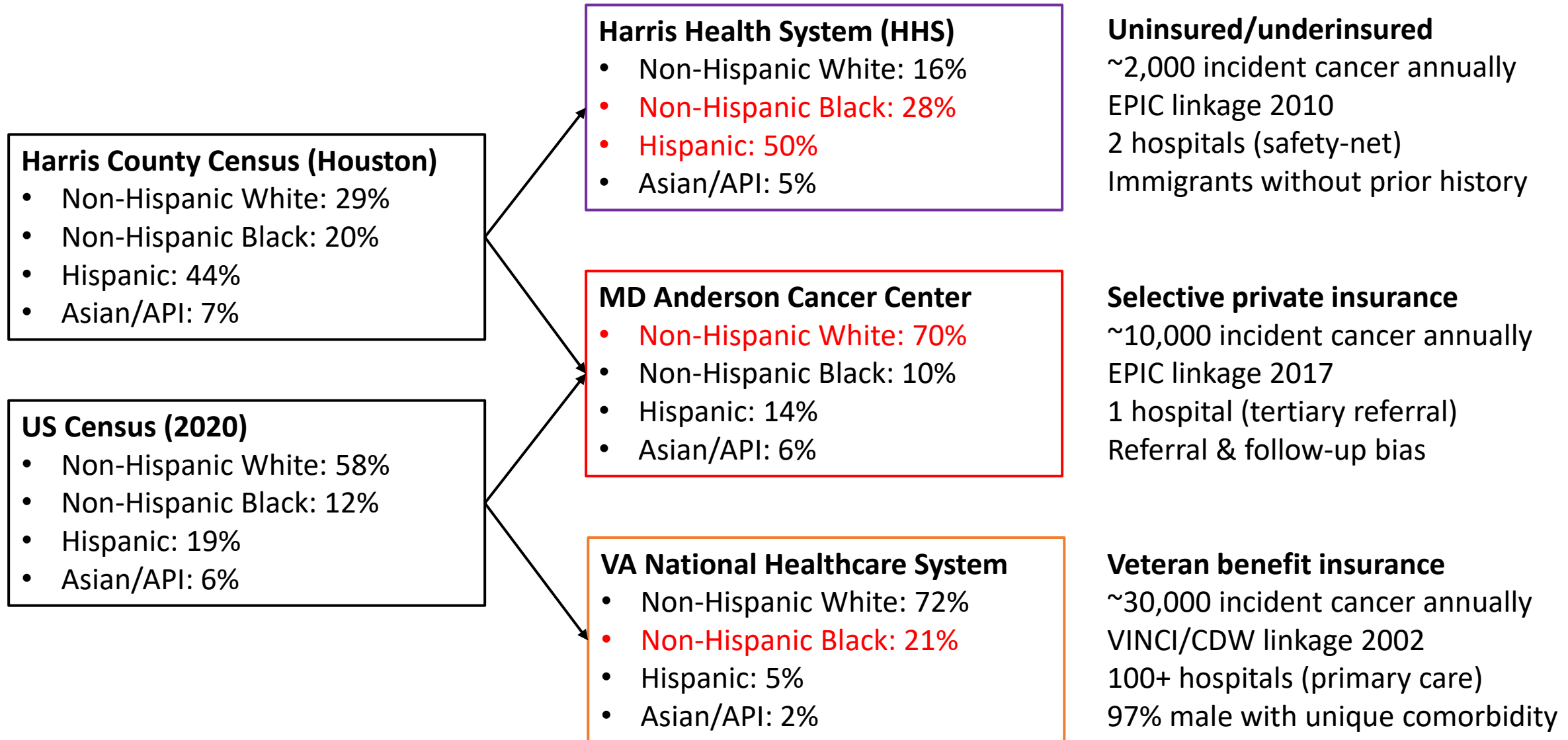
- Lack of precision: **Improved VTE prediction model**
  - “Khorana score complemented by clinical judgment and experience”
- Fear of bleeding: **Automated exclusion for bleeding risk**
  - “Used with caution in those with a high risk of bleeding”
- Lack of time: **Clinical decision support**
  - High volume clinic, not integrated into EHR
- Lack of awareness: **Simpler access to evidence**
  - Hematologist vs. oncologist; not comfortable to discuss

# Intersection of Medicine, Research and Technology

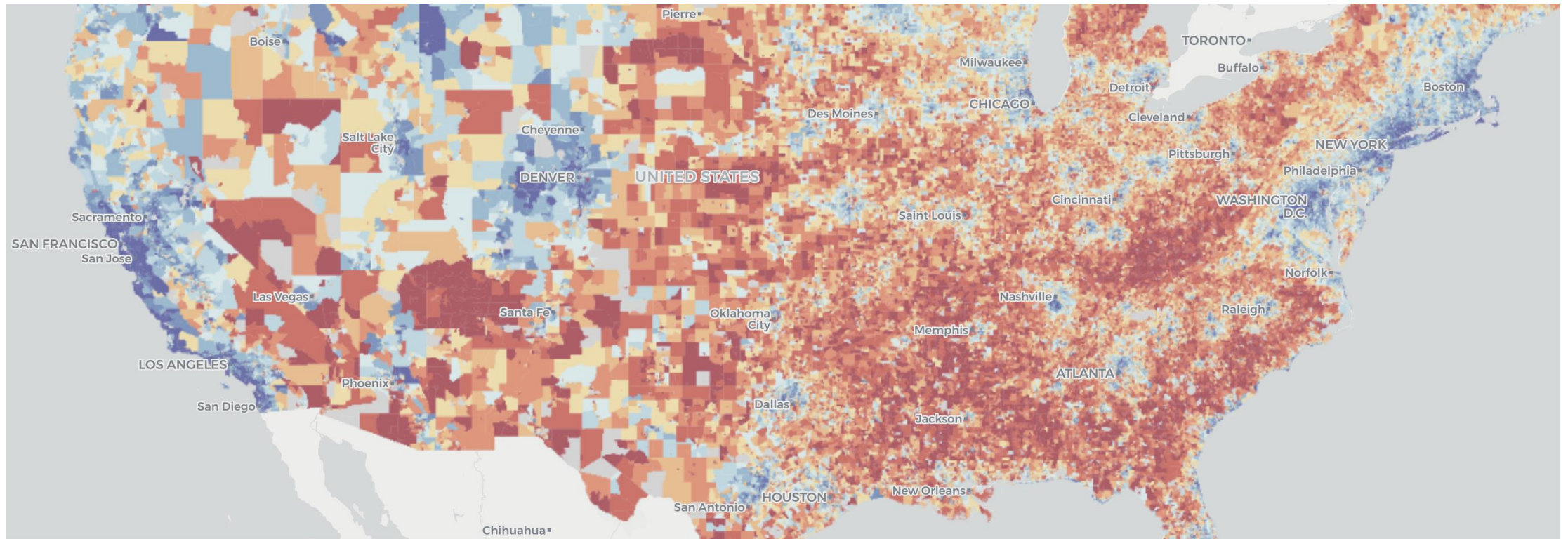


## II. EHR Database Overview

# Demographics from Different EHR Databases



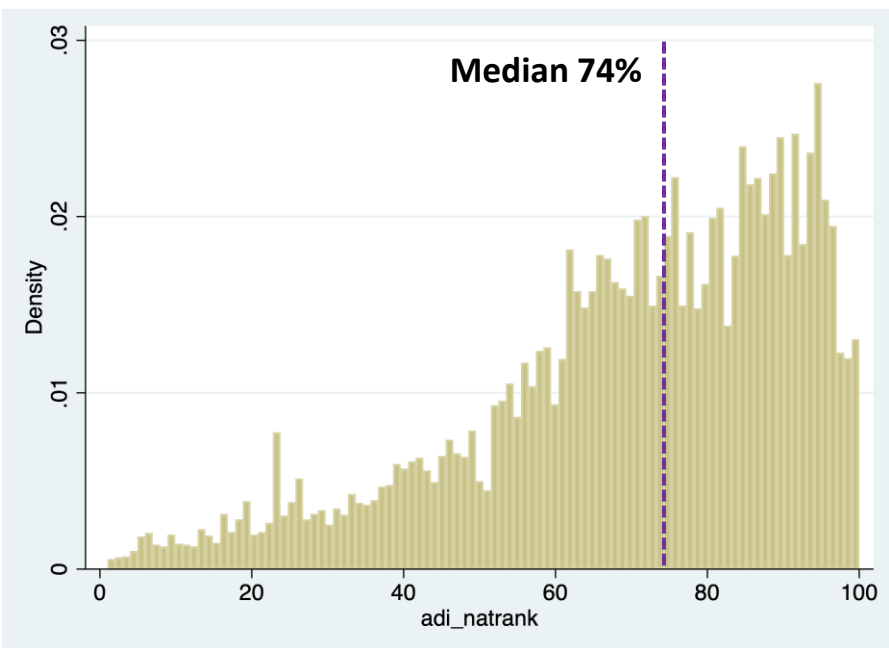
# Area of Deprivation Index



Small neighborhood social determinants of health estimated from  
Block Group level data from American Community Surveys  
4 domains: poverty, housing, employment, education

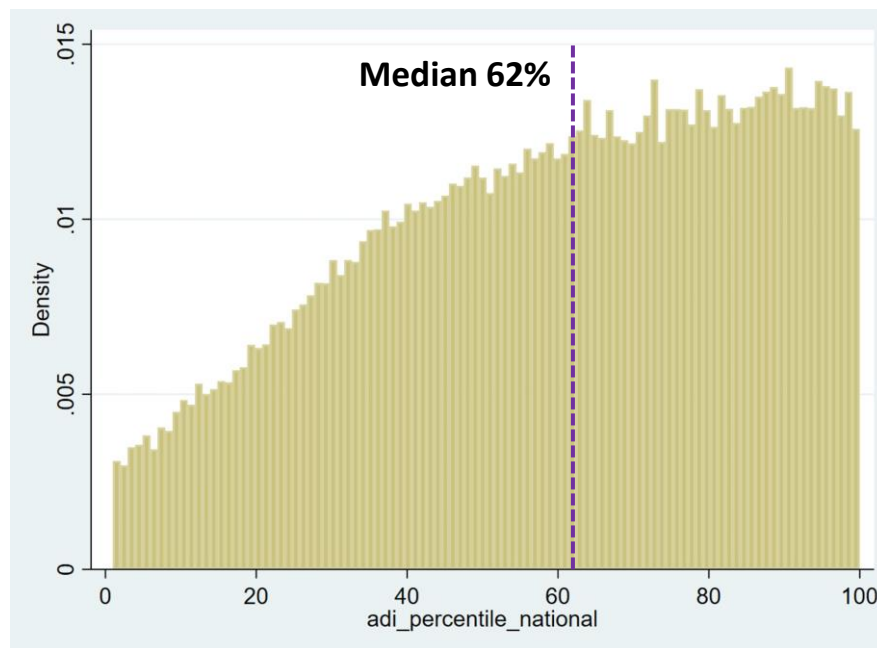
# National ADI Distribution in Cancer Databases

**HHS 2011-2021 (N=19,667)**



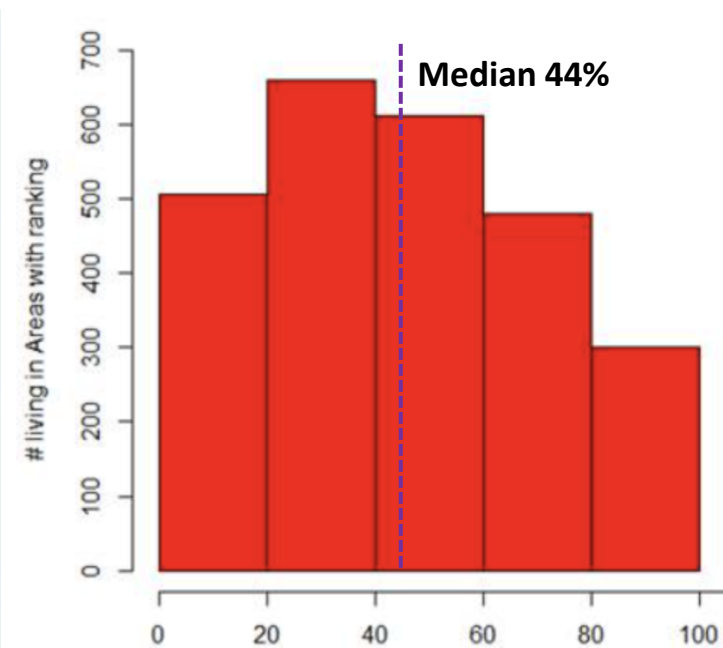
**Uninsured/underinsured**

**VA 2006-2021 (N=434,203)**



**Veteran benefit insurance**

**MDACC 2017-2021 (N=36,542)**



**Selective private insurance**

Low number = least deprived; high number = most deprived

# Data Abstraction & Linkage

## **Hospital system cancer registry (Cancer Registry)**

- Cancer registry data
  - Sequence
  - Diagnosis
  - Histology
  - Stage
  - Demographics
  - Mortality
  - Annual update with 1 year delay

+

## **Electronic Health Record (EPIC Caboodle/VINCI CDW)**

- Claims-level data
  - ICD/CPT/HCPCS codes
- Encounter-level data
  - Encounter appointments/codes
  - Medical/surgical history
  - Medications prescribed/administered
  - Laboratory/transfusion/micro
  - Imaging/procedures
  - Hospital/clinic notes
  - Daily update

Extensive data validation, cleaning, filtering, and linkage



# Integrated Cancer Data Warehouse (n=20,000 at HHS)

- Diagnosis, histology, staging
- Annually updated mortality
- Demographics at diagnosis
- Address => **geo-coded ADI**
- Comorbidities => **CCI / NCI**
- Encounter/appointment
- Scheduled/performed surgeries
- Prescribed/administered medication  
=> **lines of therapy**
- Vitals: weight/height
- Laboratory: lab, micro, transfusion

- ICD diagnosis codes (facility)
- ICD diagnosis codes (encounter)
- ICD diagnosis codes (problem list, medical history, surgical history)
- ICD procedure codes (facility coded)
- CPT/HCPCS procedure codes (facility transaction)

- Radiology impression
- Discharge summary
- Clinic progress notes
- Procedure: TTE, PFT, EGD

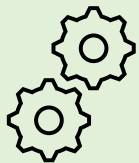
**NLP**

**Key: clinician validated & cleaned data from electronic health record is paramount for ANY methodology!**

# Health Informatics in Cancer Care Delivery

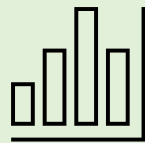
## • Research Methods

- Develop machine learning methods to address health disparity research
- Examples:
  - Computable phenotype of VTE via NLP
  - Goal = accurate/precise phenotyping of disease



## • Research Application

- Apply traditional & novel prediction models in different healthcare systems
- Examples:
  - Risk prediction model for VTE and bleeding
  - Goal = reproducible & generalizable model

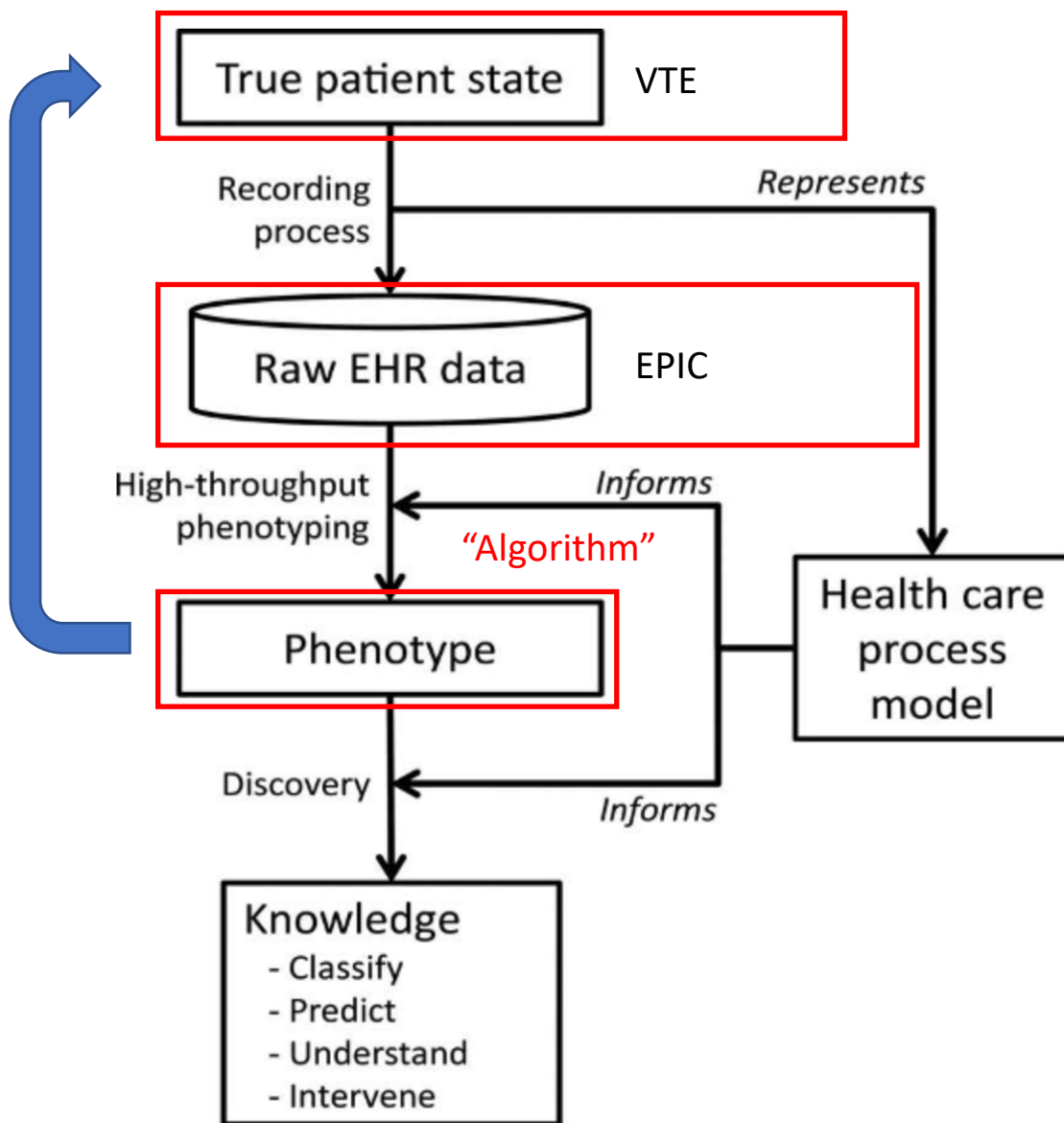


## • Clinical Application

- Integrate risk prediction models at point of care decision making in EHR databases
- Example:
  - PC-CDS for VTE prophylaxis
  - Goal = user friendly unintrusive decision aid



### III. Phenotyping VTE & Epidemiology of Cancer Associated Thrombosis (CAT)



Source: [Hripcsak and Albers 2013](#). (Used under a Creative Commons license.)

# How to determine the VTE phenotype

- Structured data

- ICD codes

- Billing: inpatient vs. outpatient
    - Encounter
    - ~~Problem list~~

- CPT codes

- Radiology studies
    - IVC filter

- Medications

- Anticoagulant

- Unstructured data (NLP)

- Sequence in repeated notes

- Region of interest

- Radiology report: impression
    - Discharge note: hospital course
    - Office progress note: A/P

- Rule-based vs. ML-based

- VTE keyword
    - Assertion negation
    - Deep learning model

Key: EHR database (billing + charting) provides much more granularity than claims database (billing)

# Defining VTE Computable Phenotype – Validation

HHS: Predicted vs. observed VTE at 12 months (selective review)

	Predicted No.	Reviewed No.	True+ VTE	True- VTE	PPV
ICD- NLP-	8,957 (92%)	300	1	299	0.33%
NLP+ only	115 (1.2%)	115	88	27	76.5%
ICD+ only	127 (1.3%)	127	78	49	61.4%
ICD+ NLP+	570 (5.9%)	200	192	8	96.0%

HHS: Performance of prediction algorithms (weighted sample)

	True+ VTE		True- VTE		
ICD- NLP-	8,957 x 0.33%	30	8,957 x 99.7%	8,927	NPV 100%
NLP+ only	115 x 76.5%	710	115 x 23.5%	102	PPV 87%
ICD+ only	127 x 61.4%		127 x 38.6%		
ICD+ NLP+	570 x 96.0%		570 x 4.0%		
	Sensitivity 96%		Specificity 99%		

ICD/medication: PPV 90%, sensitivity 84%

NLP/radiology: PPV 92%, sensitivity 84%

ICD or NLP: PPV 87%, sensitivity 96%

VA: Predicted vs. observed VTE at 12 months (selective review)

	Predicted No.	Reviewed No.	True+ VTE	True- VTE	PPV
ICD- NLP-	74,145 (93%)	300	1	299	0.33%
NLP+ only	799 (1.0%)	200	159	41	79.5%
ICD+ only	1,758 (2.2%)	200	161	39	80.5%
ICD+ NLP+	2,813 (3.5%)	200	198	2	99.0%

VA: Performance of the prediction algorithms (weighted sample)

	True+ VTE		True- VTE		
ICD- NLP-	74,145 x 0.3%	222	74,145 x 99.7%	73,923	NPV 100%
NLP+ only	799 x 79.6%	4,836	799 x 20.4%	534	PPV 90%
ICD+ only	1,758 x 80.5%		1,758 x 19.5%		
ICD+ NLP+	2,813 x 99.0%		2,813 x 1.0%		
	Sensitivity 96%		Specificity 99%		

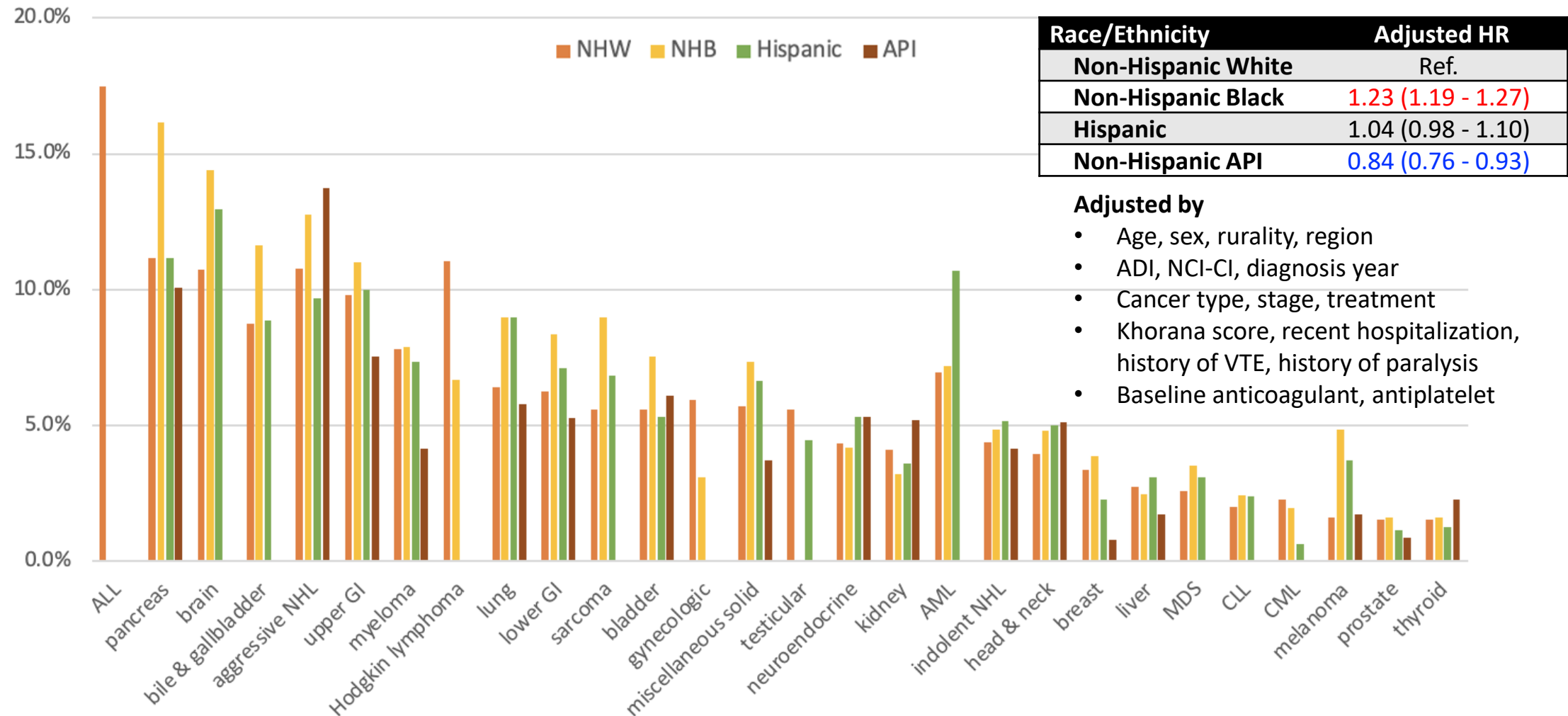
ICD/medication: PPV 89%, sensitivity 83%

NLP/radiology: PPV 95%, **sensitivity 68%**

ICD or NLP: PPV 90%, sensitivity 96%

**Key: NLP is system-specific but can greatly augment accuracy of structured phenotyping algorithm**

# Incidence of CAT by Cancer Type and Race/Ethnicity in 434,203 Veterans



**Key: CAT incidence is specific to patient (race, weight, comorbidity) and cancer (type, stage, treatment)**

## IV. VTE Risk Prediction Modeling



# Creating Validated, Optimized, and Inclusive Risk Prediction Model for CAT

- Population:
  - First cancer diagnosis receiving first-line systemic therapy within 1 year
  - Exclude if recent acute VTE last 6 months or on therapeutic AC
  - Assess VTE from index treatment until loss of follow-up (90+ day gap)
- Derivation cohort:
  - HHS (N=9,769, 2011-2020, VTE 6.2% at 6-month)
- Validation cohorts:
  - VA national (N=79,517, 2011-2020, VTE 5.1% at 6-month)
  - MD Anderson (N=21,142, 2017-2020, VTE 5.7% at 6-month)

# Clinical Knowledge is Important for Initial Variable Selection

Khorana score (KS) factors	Cancer site/histology subtype
	Pre-therapy body mass index $\geq 35$
	Pre-therapy white blood cell count $>11$
	Pre-therapy hemoglobin $<10$
	Pre-therapy platelet $\geq 350$
Cancer-specific risk factors	Cancer Stage
	Treatment initiation timing
	Treatment regimen
Patient-demographic factors	Age
	Sex
	Race/Ethnicity
Additional clinical factors	History of PE/LE-DVT
	Recent prolonged hospitalization $>3d$ last 3 months
	Anticoagulant prescription in last 3 months
	Antiplatelet prescription in the last 3 months
	Surgery within last 3 months
Additional lab factors	Creatine
	Total bilirubin
	Alanine transaminase

Patient Comorbidities	Congestive heart failure
	Myocardial infarction
	Cardiac arrhythmia
	Cardiac valvular disease
	Peripheral vascular disease
	Cerebral vascular disease
	Chronic obstructive pulmonary disease
	Paralysis or immobility
	Diabetes
	Hypertension
	Renal Disease
	Liver disease
	HIV/AIDS
	Rheumatologic disease
	Coagulopathy

**A priori selected risk predictors for VTE**

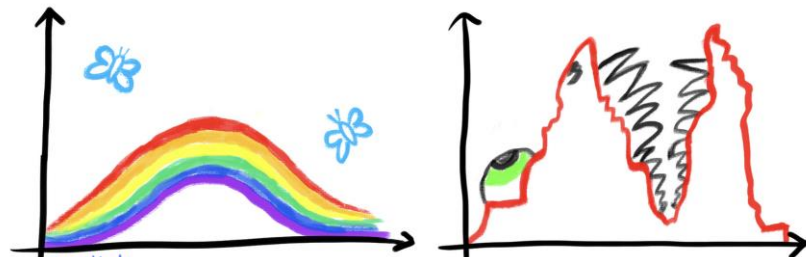
# Interpretable vs. Black Box Machine Learning models

- Linear regression, logistic regression, Cox regression
- Generalized linear models: non-Gaussian outcomes (family/link)
- Generalized additive models: non-linear outcomes (splines)

```
## [1] "SL.bartMachine"      "SL.bayesglm"        "SL.biglasso"  
## [4] "SL.caret"            "SL.caret.rpart"     "SL.cforest"  
## [7] "SL.earth"            "SL.extraTrees"      "SL.gam"  
## [10] "SL.gbm"              "SL.glm"             "SL.glm.interaction"  
## [13] "SL.glmnet"           "SL.ipredbag"         "SL.kernelKnn"  
## [16] "SL.knn"              "SL.ksvm"            "SL.lda"  
## [19] "SL.leekasso"         "SL.lm"              "SL.loess"  
## [22] "SL.logreg"           "SL.mean"            "SL.nnet"  
## [25] "SL.nnls"             "SL.polymars"        "SL.qda"  
## [28] "SL.randomForest"     "SL.ranger"          "SL.ridge"  
## [31] "SL.rpart"            "SL.rpartPrune"      "SL.speedglm"  
## [34] "SL.speedlm"          "SL.step"            "SL.step.forward"  
## [37] "SL.step.interaction" "SL.stepAIC"         "SL.svm"  
## [40] "SL.template"         "SL.xgboost"
```

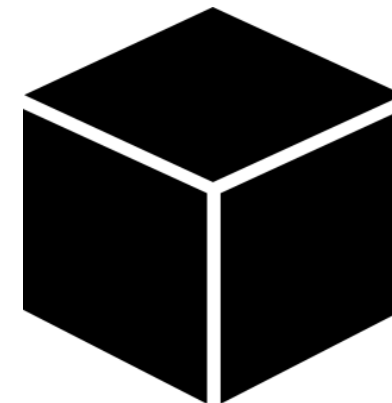
## UNDERLYING DISTRIBUTIONS:

PARAMETRIC  
ASSUMPTIONS VS. REALITY



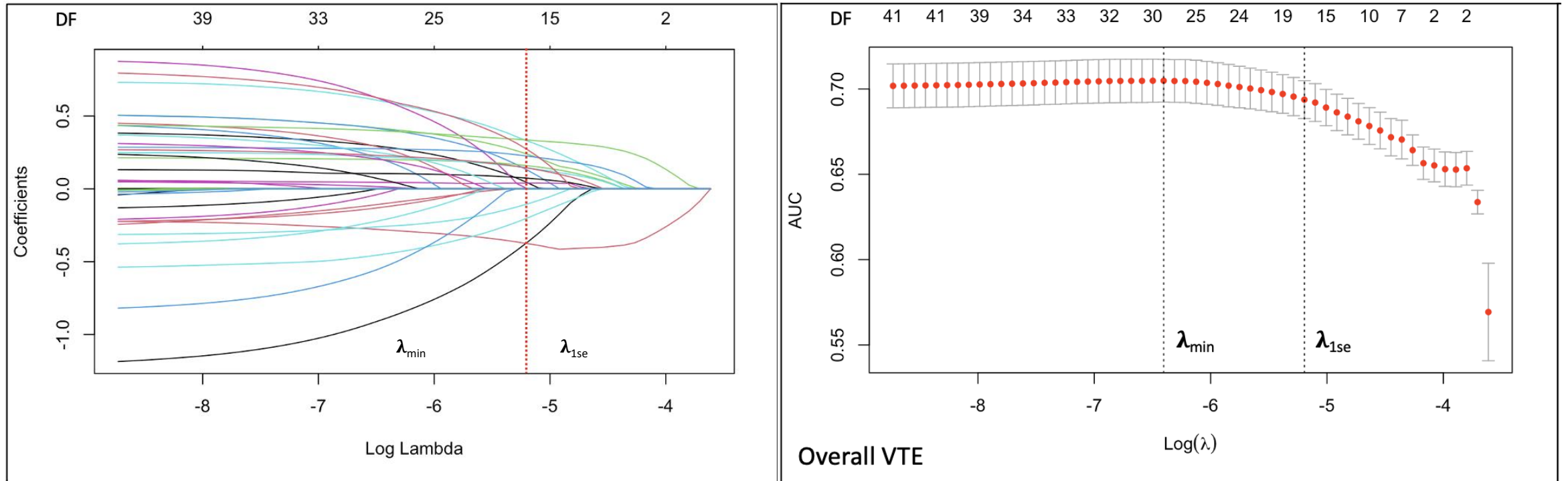
<https://www.khstats.com/blog/tmle/tutorial>

Features



Outcome

# Feature Selection via LASSO Penalized Shrinkage



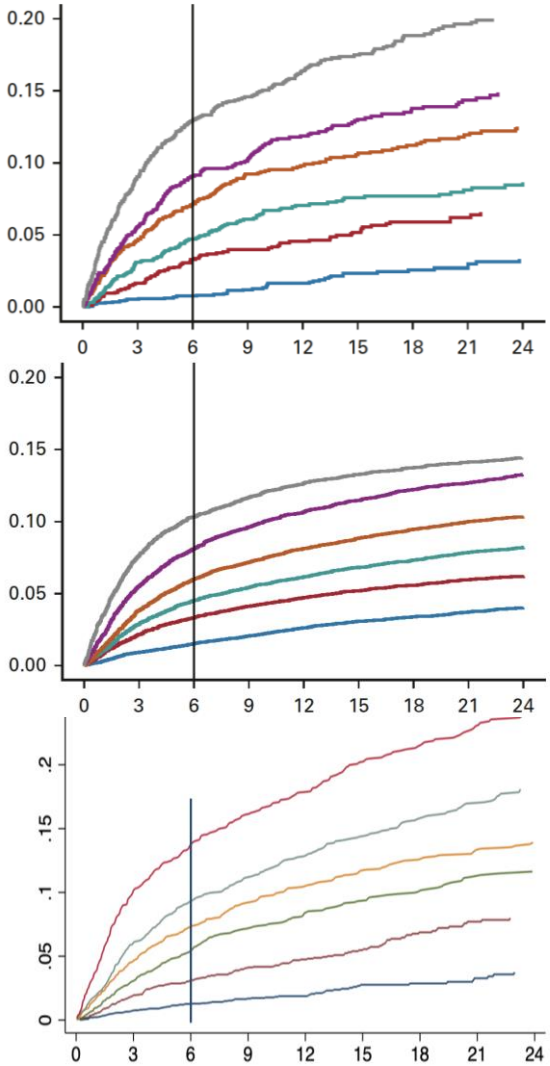
Goal is to optimize prediction with the most parsimonious model (trade-off between complexity & fit)

# Logistic Regression Model Fitted from LASSO Selection

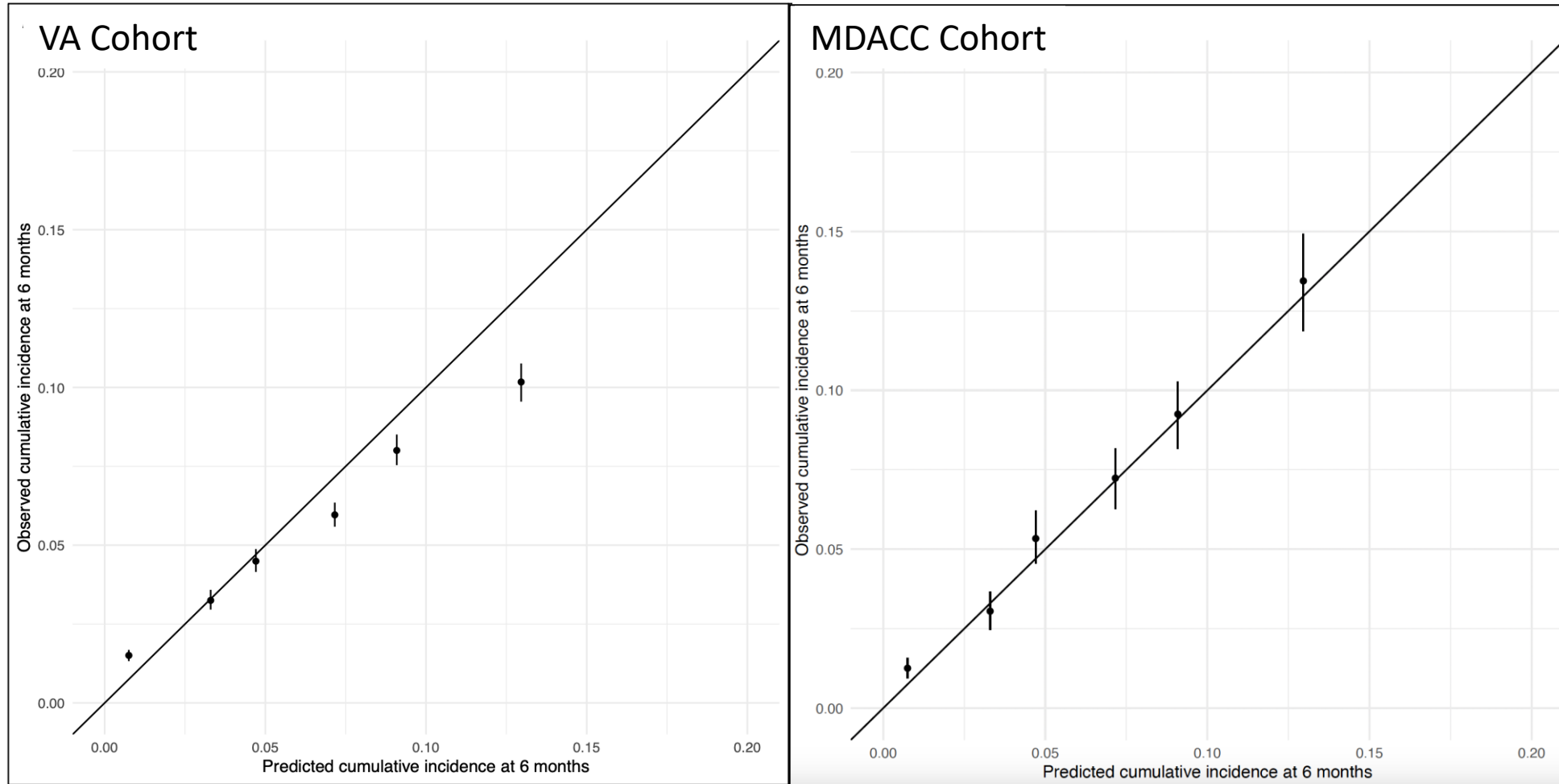
	Risk Predictors	Number (%)	OR for VTE (95% CI)	Point
Khorana score risk factors	<b>Modified cancer subtype risk</b>			
	- Other solid or heme cancer <sup>a</sup>	5,206 (53.3%)	Reference	0
	- <b>Colorectal cancer</b>	1,152 (11.8%)	1.36 (1.01-1.82)	1
	- Lung, ovarian, uterine, bladder, kidney, testicular, <b>aggressive NHL, myeloma, brain, soft tissue sarcoma</b>	2,644 (27.1%)	2.23 (1.81-2.74)	2
	- Pancreas, gastric, esophageal, <b>cholangiocarcinoma, gallbladder</b>	767 (7.9%)	2.26 (1.69-3.03)	3
	<b>Pre-therapy BMI ≥35</b>	1,318 (13.5%)	1.45 (1.14-1.83)	1
	<b>Pre-therapy WBC &gt;11</b>	1,652 (16.9%)	1.34 (1.09-1.65)	1
	<b>Pre-therapy hemoglobin &lt;10</b>	2,042 (20.9%)	1.49 (1.23-1.80)	1
	<b>Pre-therapy platelet ≥350</b>	2,700 (27.6%)	1.24 (1.03-1.49)	1

# Discrimination of Novel Risk Prediction Model

Dataset	Risk score	VTE % at 6 mo	Classification	VTE % at 6 mo	TD-C statistic (95% CI)
HHS Derivation Cohort	0- (1,938)	0.8% (14)	Low-risk 50.8% (4,958)	2.8% (131)	0.71 (0.69-0.72)
	1 (1,483)	3.3% (47)			
	2 (1,537)	4.7% (70)			
	3 (1,644)	7.2% (114)	High-risk 49.2% (4811)	9.8% (459)	
	4 (1,523)	9.1% (135)			
	5+ (1,644)	13.0% (210)			
VA Validation Cohort	0- (18,022)	1.5% (267)	Low-risk 54.2% (43,894)	3.0% (1,272)	0.68 (0.67-0.69)
	1 (12,551)	3.3% (411)			
	2 (13,321)	4.5% (594)			
	3 (14,969)	6.0% (888)	High-risk 44.8% (35,623)	7.8% (2,755)	
	4 (11,381)	8.1% (915)			
	5+ (9,273)	10.3% (952)			
MDACC Validation Cohort	0- (5,661)	1.3% (59)	Low-risk 60.0% (12,681)	2.6% (325)	0.71 (0.69-0.72)
	1 (3,558)	3.1% (99)			
	2 (3,462)	5.4% (167)			
	3 (3,489)	7.3% (232)	High-risk 40.0% (8,461)	8.8% (742)	
	4 (2,918)	9.3% (250)			
	5+ (2,054)	13.8% (260)			



# Calibration Curves in Validation Cohorts



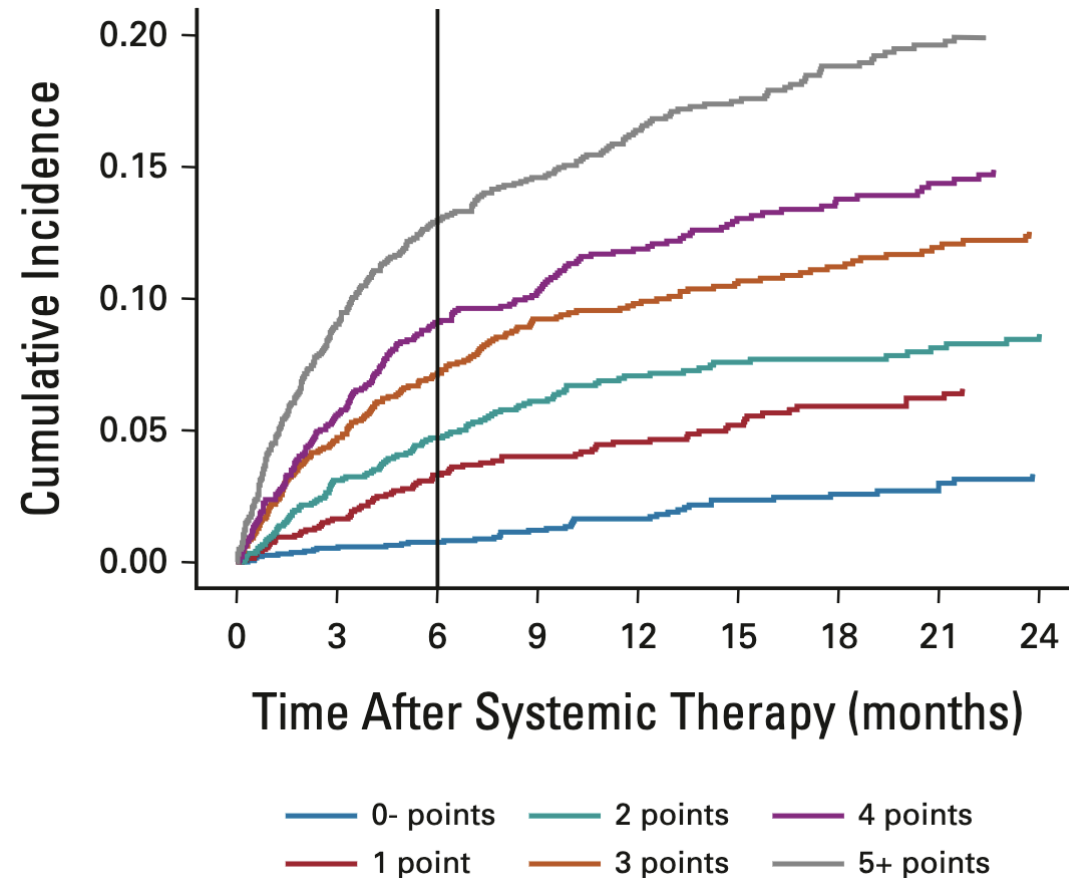
# Comparison with Khorana Score

Dataset	Category	Khorana Score	New RAM	Number	VTE % at 6 mo
<b>HHS Derivation Cohort</b>	Concordant (78%)	Low-risk	Low-risk	4,495	2.6% (112)
		High-risk	High-risk	3,107	10.5% (321)
	Reclassified (22%)	Low-risk	High-risk	1,704	8.4% (138)
		High-risk	Low-risk	463	4.3% (19)
<b>VA Validation Cohort</b>	Concordant (72%)	Low-risk	Low-risk	40,360	3.0% (1,184)
		High-risk	High-risk	17,242	8.2% (1,406)
	Reclassified (28%)	Low-risk	High-risk	18,381	7.4% (1,349)
		High-risk	Low-risk	3,534	2.5% (88)
<b>MDACC Validation Cohort</b>	Concordant (80%)	Low-risk	Low-risk	11,947	3.0% (303)
		High-risk	High-risk	4,931	10.0% (451)
	Reclassified (20%)	Low-risk	High-risk	3,530	9.0% (291)
		High-risk	Low-risk	734	3.4% (22)

**Key: New risk model increases VTE % in high-risk group by ~25% & improves overall C statistic ~0.07**



# Available Online Calculator



<https://dynamicapp.shinyapps.io/EHR-CAT/>

Cancer site/histology (choose one from the following)

Other cancers (0)

Cancer stage (AJCC)

- ☒ Early stage (I-II) (0)  
☐ Advanced stage (III-IV) (+1)

Type of systemic therapy

- ☒ Cytotoxic chemotherapy (chemo) and/or immune checkpoint inhibitor (ICI) (0)  
☐ Targeted and/or endocrine therapy without chemo/ICI (-1)

Patient race

- ☒ All other race (0)  
☐ East/South Asian, Pacific Islander, American Indian or Alaskan Native (-1)

Pretherapy body mass index  $\geq 35$

- ☒ No (0) ☐ Yes (+1)

Pretherapy white blood cell count  $> 11 \times 10^9/L$

- ☒ No (0) ☐ Yes (+1)

Pretherapy hemoglobin  $< 10 \text{ g/dL}$

- ☒ No (0) ☐ Yes (+1)

Pretherapy platelet count  $\geq 350 \times 10^9/L$

- ☒ No (0) ☐ Yes (+1)

History of VTE

- ☒ No (0) ☐ Yes (+1)

History of paralysis/immobility

- ☒ No (0) ☐ Yes (+1)

Recent/current hospitalization  $> 3$  days in the past 3 months

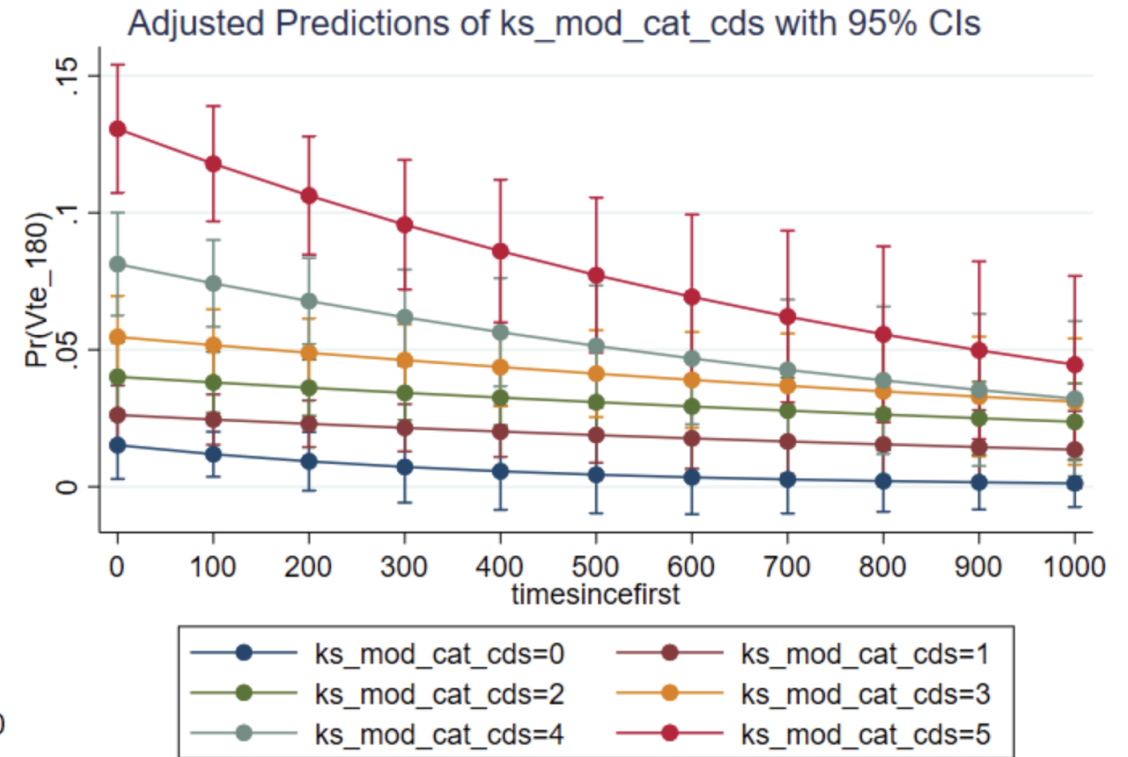
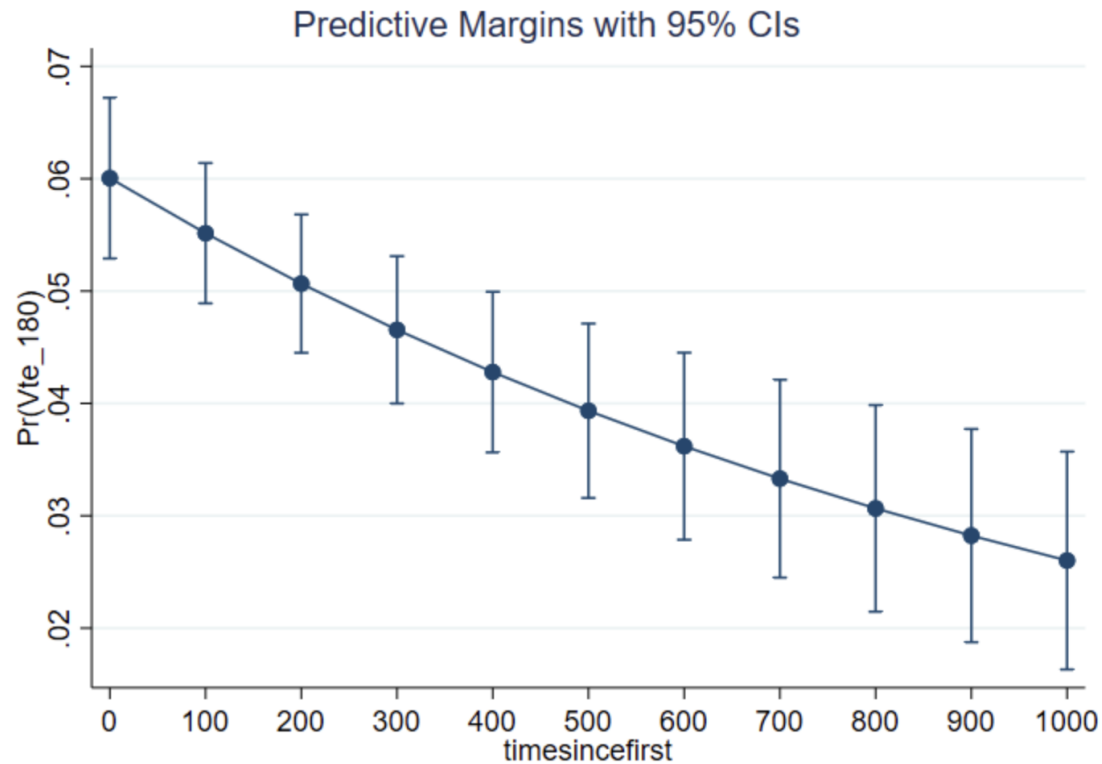
- ☒ No (0) ☐ Yes (+1)

# V. Dynamic Modeling & Implementation of PC-CDS

# Ambulatory Pharmacologic Prophylaxis is Rarely Implemented

- Lack of precision: **Improved VTE prediction model**
  - “Khorana score complemented by clinical judgment and experience”
- Fear of bleeding: **Automated exclusion for bleeding risk**
  - “used with caution in those with a high risk of bleeding”
- Lack of time: **Clinical decision support**
  - high volume clinic, not integrated into EHR
- Lack of awareness: **Simpler access to evidence**
  - hematologist vs. oncologist; not comfortable to discuss

# CAT Risk Decreases Over Time

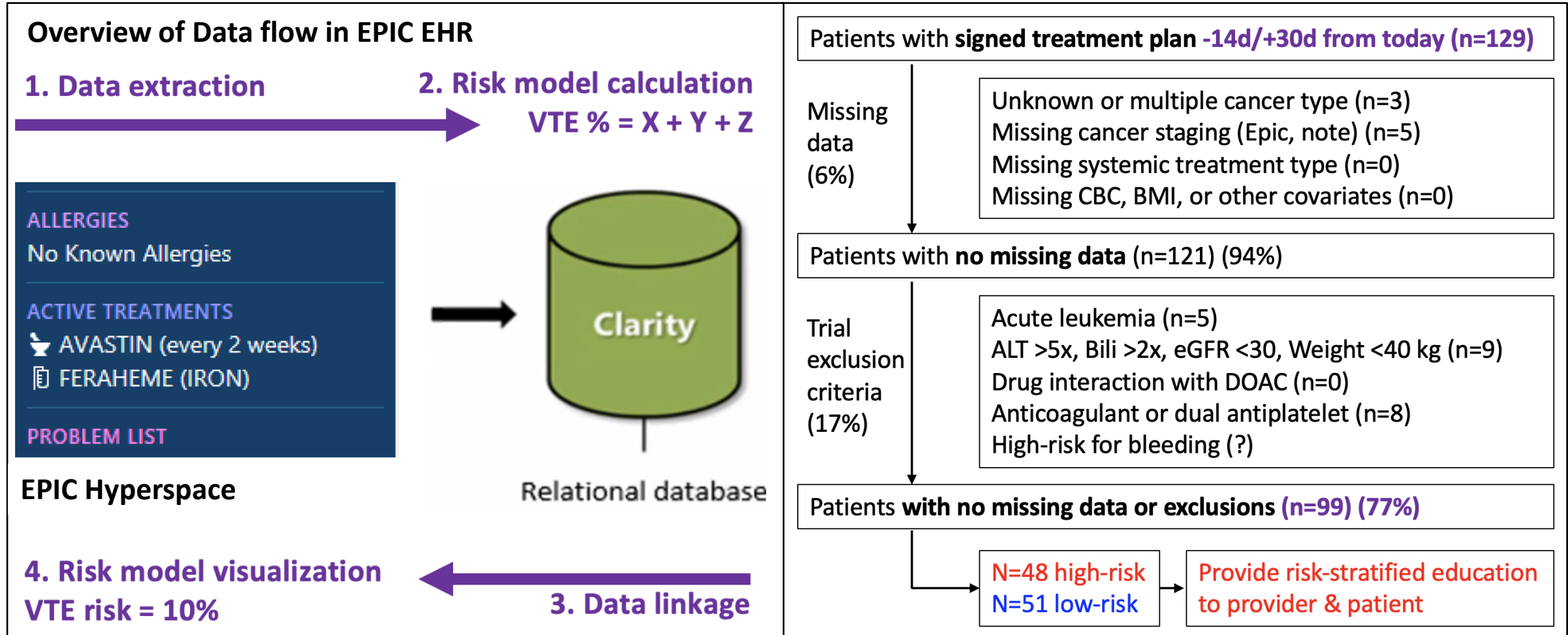


Key: a time adjustment factor is needed to apply a static model over time

# Patient Specific Risk Factors Change Over Time

id	plan_n	cancer_type	dx_date	treat_date	chemo_class_regimen_cds	risk_score	vte_180d	vte_date	vte_type
1	1	breast	5/26/20	07/23/20	chemo+/-others (no immuno)	2	0	.	.
1	2	breast	5/26/20	11/17/20	chemo+/-others (no immuno)	4	0	.	.
1	3	breast	5/26/20	03/09/21	chemo+/-others (no immuno)	2	0	.	.
2	1	breast	6/27/18	08/17/18	target+/-endo (no chemo/immuno)	1	0	.	.
2	2	breast	6/27/18	10/19/18	target+/-endo (no chemo/immuno)	2	0	.	.
2	3	breast	6/27/18	12/27/19	target+/-endo (no chemo/immuno)	1	0	.	.
24	1	lung	10/27/16	12/12/16	chemo+/-others (no immuno)	5	1	4/20/17	Acute PE
29	1	colorectal	6/23/17	08/04/17	chemo+/-others (no immuno)	5	0	7/6/18	Acute PE
29	2	colorectal	6/23/17	12/12/17	chemo+/-others (no immuno)	4	1	7/6/18	Acute PE
29	3	colorectal	6/23/17	03/19/18	chemo+/-others (no immuno)	2	1	7/6/18	Acute PE
105	1	lung	8/29/19	10/11/19	chemo/immuno+others	7	0	7/16/20	Acute PE
105	2	lung	8/29/19	01/22/20	immuno+/-others (no chemo)	6	1	7/16/20	Acute PE

# Automate Patient Selection & Exclusion in EHR Prospectively

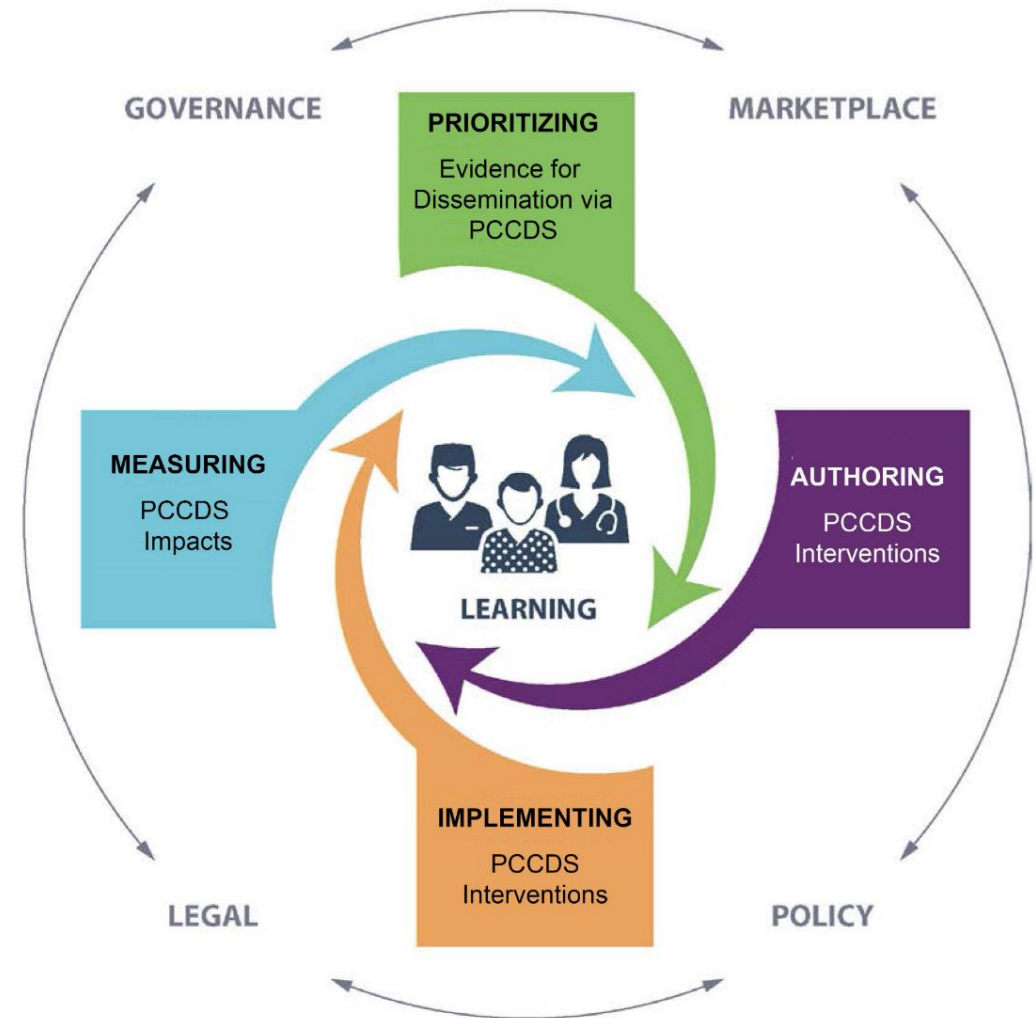


# Patient Centered Clinical Decision Support (PC-CDS)

- **Design/assess/optimize usage:**

- Time consuming process
- Design provider- & patient-specific surveys & education fliers
- Assess barriers to implementation (<25%): time, cost, difficulty, annoyance
- Assess outcomes after implementation
- Modify implementation strategy

- **BCM VCG QI project 2023**



# Thank You

- Research Team

- Danielle Guffey (statistician)
- Raka Bandyo (data analyst)
- Xiangjun Xiao (lead programmer)
- Shengling Ma (post-doc)
- Rockbum Kim (post-doc)
- Arash Maghsoudi (post-doc)
- Mahrukh Jamil (research coordinator)

- Collaborators

- Nathanael Fillmore
- Kelly Merriman
- Abiodun Oluyomi
- Bo Peng
- Cristhiam Rojas Hernandez
- Jordan Schaefer

- Mentors

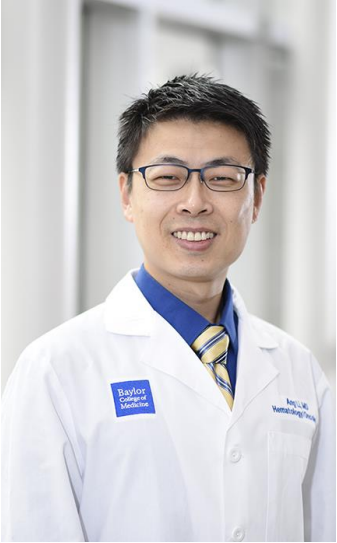
- Christopher Amos
- Christopher Flowers
- Stephanie Lee
- Marc Carrier
- Neil Zakai
- David Garica

- Funding Support

- CPRIT First-Time Tenure-Track
- NIH AIM-AHEAD
- NIH NHLBI K23



# Presenters



**Ang Li, MD, MS**  
Assistant Professor  
*Baylor College of Medicine*



**Kristen Sanfilippo, MD**  
Associate Professor of Medicine,  
Staff Physician  
*Washington University of  
Medicine in St. Louis; St. Louis VA  
Medical Center*



**Marc Carrier, MD, MSc, FRCPC**  
Head, Division of Hematology,  
Department of Medicine  
Professor, Faculty of Medicine  
*The Ottawa Hospital; University of  
Ottawa*

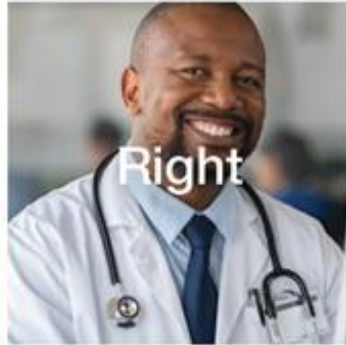


**Jean Connors, MD**  
Medical Director, Hemostatic  
Antithrombotic Stewardship  
Medical Director,  
Anticoagulation Management  
Services  
Hematology Division  
*Brigham and Women's Hospital  
/ Dana-Farber Cancer Institute*  
Associate Professor of  
Medicine  
*Harvard Medical School*

# Webinar Archive on YouTube

@AnticoagForum

| All Things Anticoagulation



**Live Broadcast  
Friday, October 13 &  
Saturday, October 14**



*Extended On-Demand Access for  
30 days Until November 14*



**Join us at this compact 2-day meeting!**

- ✓ **\$249** per person
- ✓ **22** presentations
- ✓ **Daily** chalk talks
- ✓ **15+** hours of CME for Physicians, Nurses, & Pharmacists
- ✓ **Virtual** exhibit hall



**<https://acforumbootcamp.org/2023/>**





# Webinar ▶

This webinar is brought to you, in part, by the support of the following companies:



[acforum.org](http://acforum.org)