Computational Phenotyping in Electronic Health Records

Slides from Pedro L. Teixeira, PhD

Evaluation of Data Sources and Algorithms for Hypertension Phenotyping in the Electronic Health Record

Hypertension has high prevalence and mortality

- Consistently high blood pressure readings:
  - Systolic > 139 mmHg OR Diastolic > 89 mmHg
- Hypertension is a risk factor for:
  - Heart attack
  - Heart failure
  - Stroke
  - Kidney disease
  - Aneurysm
- Affects approximately 1/3rd of Americans
- Contributes to 1/6th of deaths
- Can be well controlled with lifestyle changes and medications

Cutler, J. et al. (2008). Trends in hypertension prevalence, awareness, treatment, and control rates in United States adults between 1988-1994 and 1999-2004.

http://www.nhlbi.nih.gov/guidelines/hypertension/express.pdf

A seemingly simple question…

Which of the patients in the EHR have hypertension?

Phenotyping algorithms are usually sets of conditions applied across data types

- If the individual has three hypertensive blood pressure readings

OR

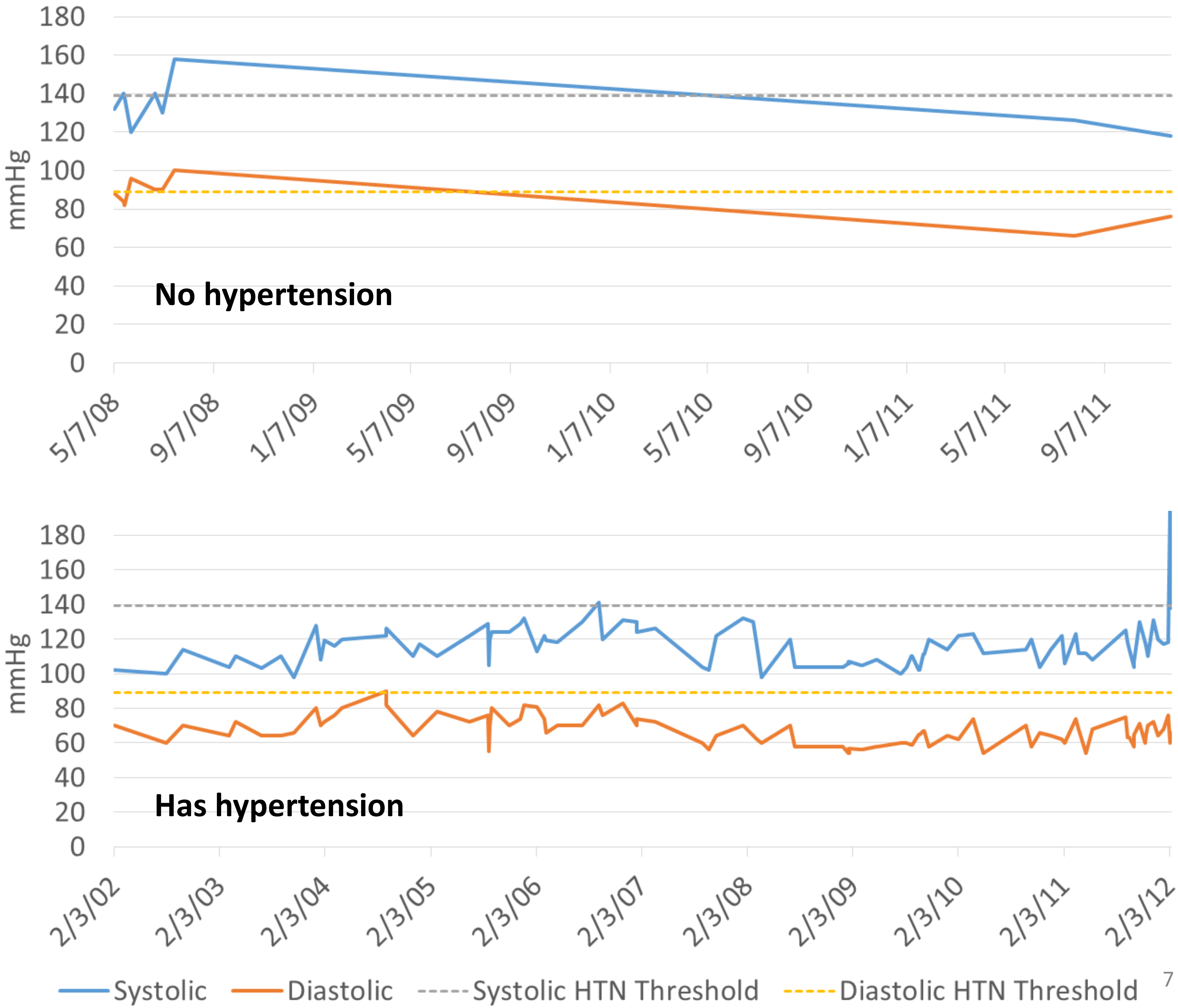
- **(**If the individual has a billing code for hypertension

AND

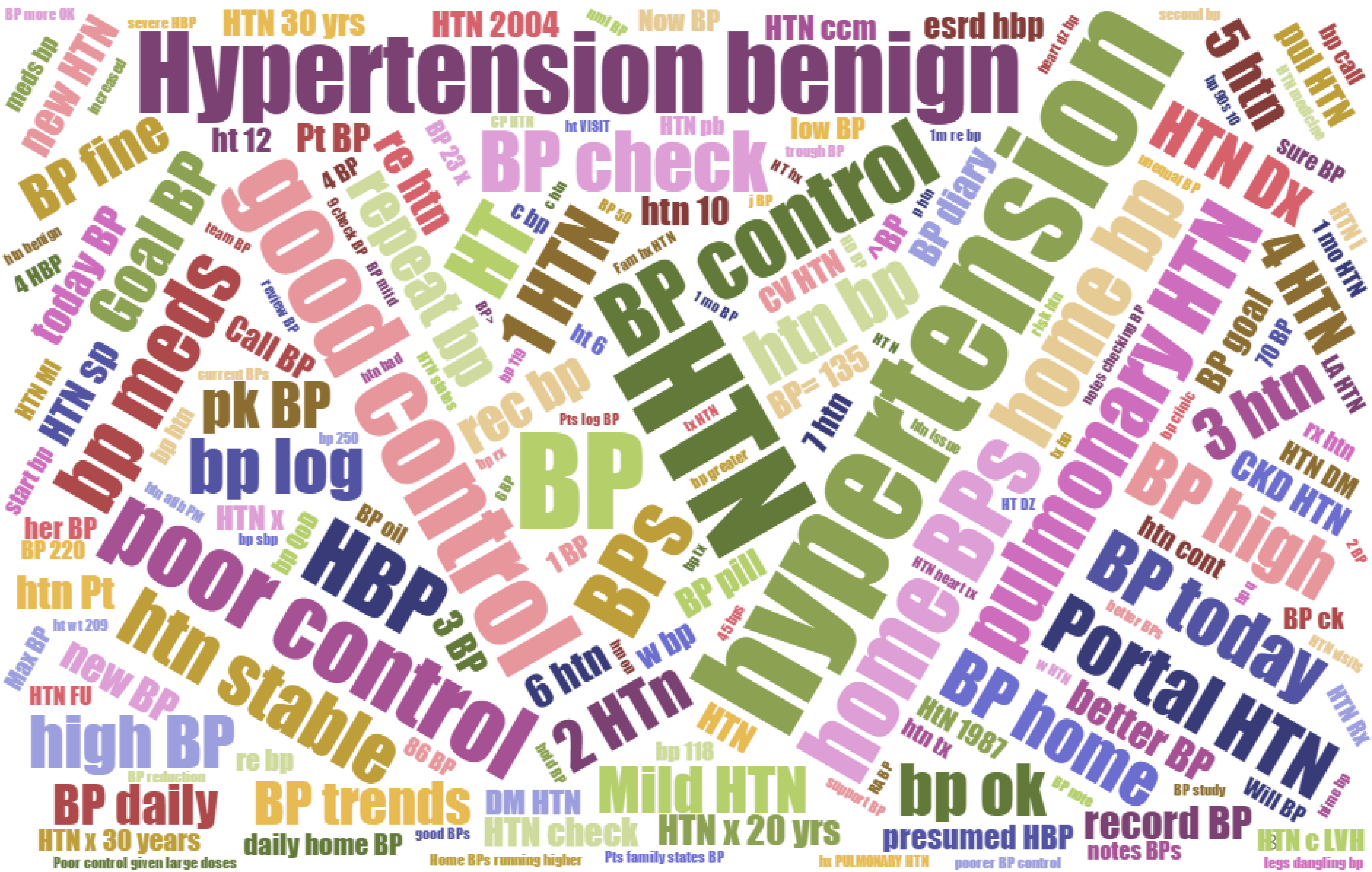
- One or more medications that are used to treat hypertension**)**

Phenotyping algorithms are important for research and clinical practice

- Large scale genotype-phenotype studies
- Outbreak surveillance
- Clinical decision support
- Quality improvement


No hypertension
mmHg
180
160
140
120
100
80
60
40
20
0
5/7/08
9/7/08
1/7/09
5/7/09
9/7/09
1/7/10
5/7/10
9/7/10
1/7/11
5/7/11
9/7/11
Has hypertension
mmHg
180
160
140
120
100
80
60
40
20
0
2/3/02
2/3/03
2/3/04
2/3/05
2/3/06
2/3/07
2/3/08
2/3/09
2/3/10
2/3/11
2/3/12
Systolic
Diastolic
Systolic HTN Threshold
Diastolic HTN Threshold

Many ways to indicate in text that a patient has hypertension



Regular expressions can match text with limited sets of rules

- Search for ‘hypertension’ OR ‘htn’ – ignore case
- Matches:
  - **<u>Hypertension</u>**
  - **<u>hypertension</u>**
  - **<u>HTN</u>**
  - **<u>HtN</u>**
  - **<u>htn</u>**
  - … pulmonary-**<u>hypertension</u>**
  - The patient does not have **<u>hypertension</u>**.
  - Mr. Smith’s father and mother have **<u>hypertension</u>**.

Advanced natural language processing tools can extract concepts from narrative text

Pre-processing	Mr. Smith is a 57yo man with HLD and HTN. He's here todya for f/u.
Sentence boundary detection	Mr. Smith is a 57yo man with HLD and HTN. | He's here today for f/u.
Tokenization	Mr . Smith is a 57yo man with HLD and HTN . He's here today for f / u .
Part-of-speech tagging	NNP NNP VBZ DT NN NN IN NNP CC NNP . | PRP ' VBZ RB NN IN NN .

Named entity recognition

Mr . Smith	man	HLD	HTN	f / u
Name title C01704446	Male gender C00024554	Hyperlipidemia C00020473	Hypertension, Malignant [Disease/Finding] C00020540	Follow-up C01522577

Denny, J., *et al* (2003). "Understanding" medical school curriculum content using KnowledgeMap
Friedman C, Rindflesch TC, Corn M. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. 2013

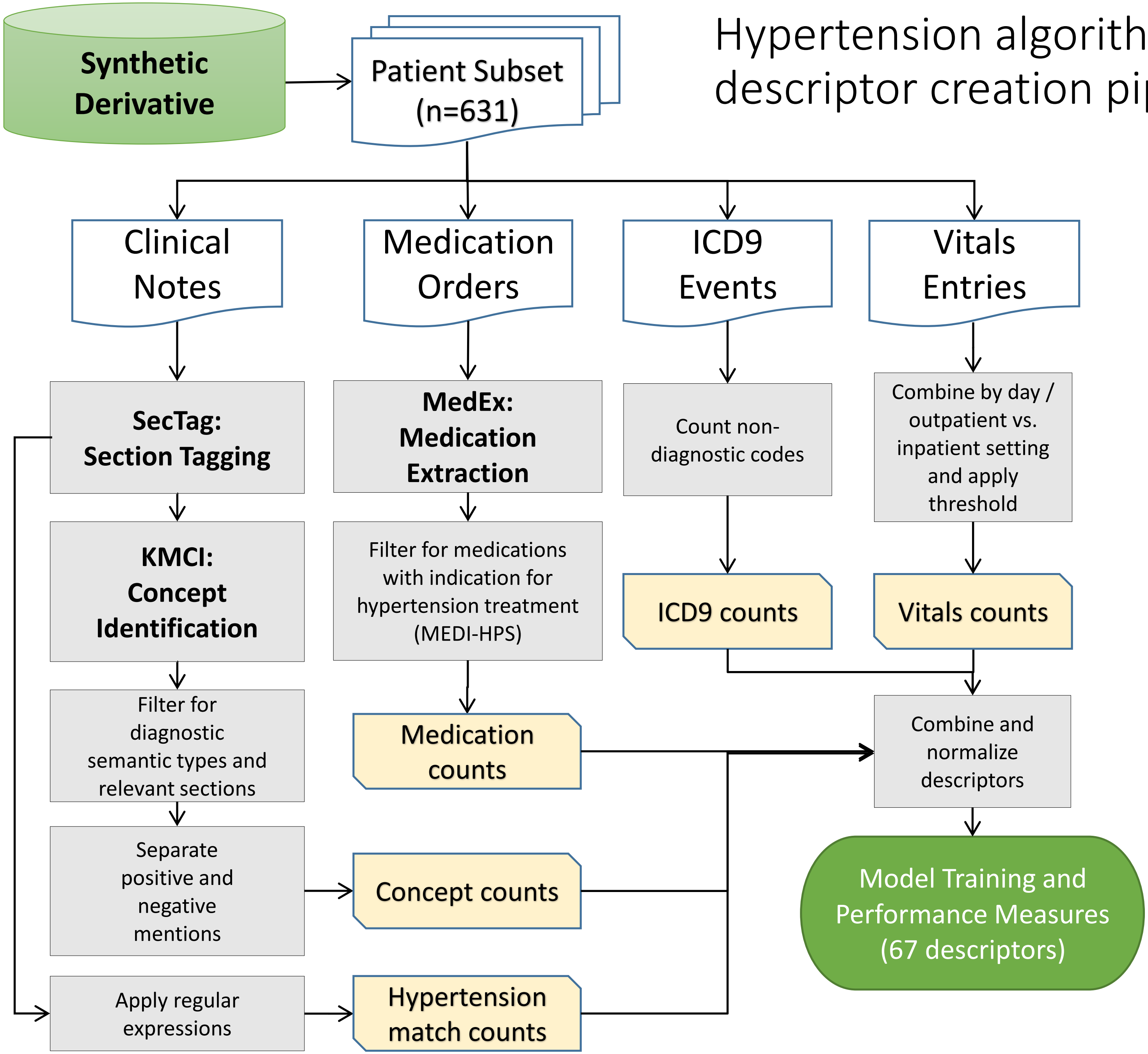
Goal: Evaluate various EHR data sources for hypertension phenotyping

- Examine four categories of data:
  - Vitals
  - Billing (ICD9) codes
  - Medication orders
  - Clinical notes
- Develop and test various algorithms on these data sources
  - Individually
  - Categorically
  - In combination
  - Using machine learning

Hypertension Algorithm Development Dataset (n=643)

- 643 Randomly selected adults with:
  - 2 or more outpatient visits between 1/1/07 -1/1/09
  - 2 or more vitals readings between 1/1/07 -1/1/09
- 631 labeled as case or control
- Manually reviewed by three physicians and one medical student
  - 20% overlap (on initial 303 of 643)
  - kappa = 0.93
- ~2/3rd Cases

Hypertension algorithm descriptor creation pipeline



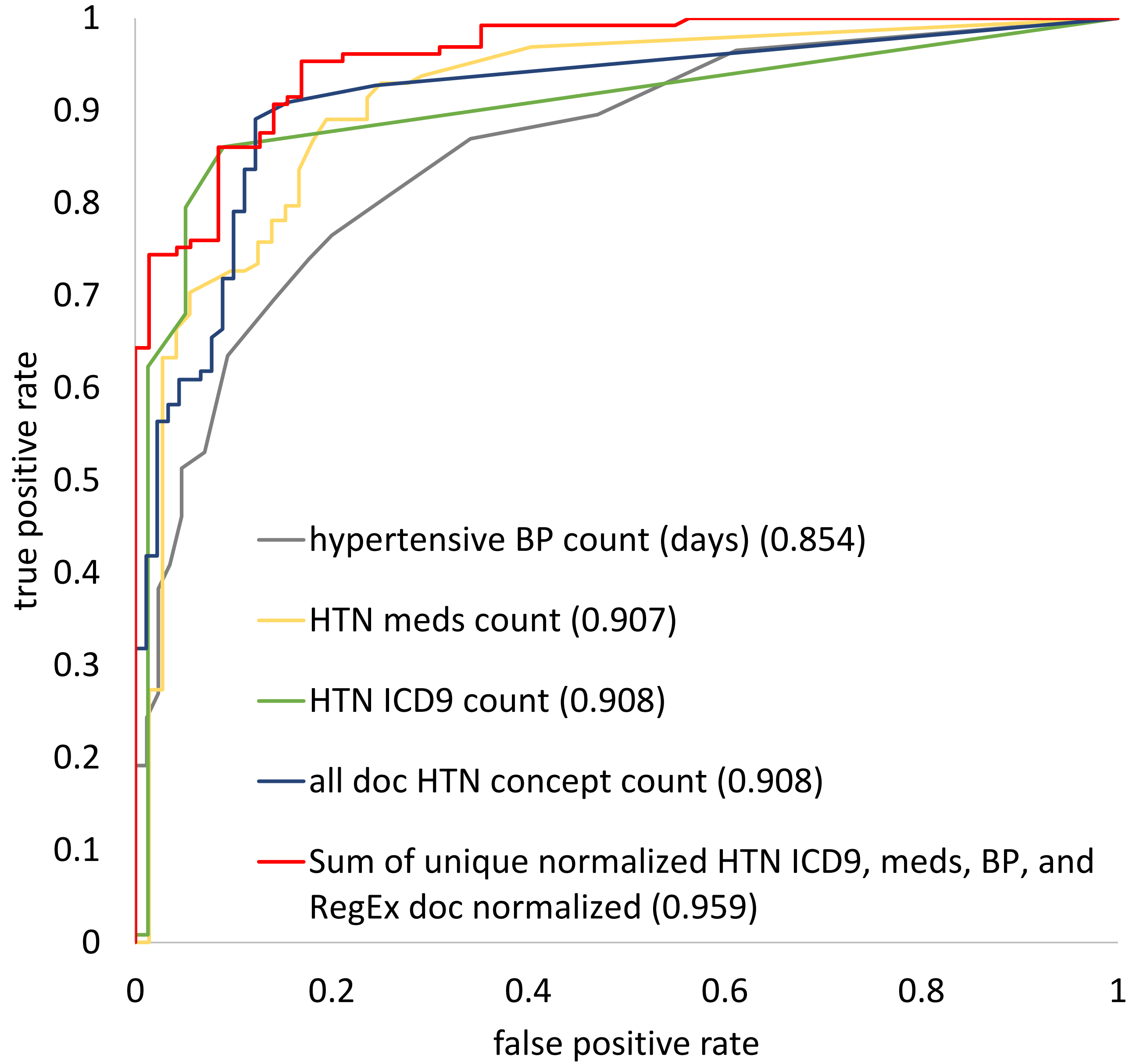
All 67 descriptors extracted

Category	Column description
ICD9, medications, all BP	date-ICD9 count
	unique ICD9 count
	HTN ICD9 count
	unique HTN ICD9 count
	HTN ICD9 count normalized
	HTN ICD9 count unique normalized
	unique HTN ICD9 count normalized
	unique HTN ICD9 count unique normalized
	meds count
	unique meds count
	HTN meds count
	unique HTN meds count
	HTN meds count normalized
	HTN meds count unique normalized
	unique HTN meds count normalized
	unique HTN meds count unique normalized
	max age
	vital reading time span (days)
	visits with vitals count (days)
	hypertensive BP count (days)
	median systolic (all)
	median diasystolic (all)
	hypertensive BP count normalized
	vitals density

Category	Column description
All vitals with separate outpatient and inpatient blood pressures	outpatient visits with vitals count (days)
	outpatient hypertensive BP count (days)
	median systolic (outpatient)
	median diasystolic (outpatient)
	outpatient hypertensive BP count normalized
	outpatient vitals density
	inpatient visits with vitals count (days)
	inpatient hypertensive BP count (days)
	median systolic (inpatient)
	median diasystolic (inpatient)
	inpatient hypertensive BP count normalized
	inpatient vitals density
	median pulse (all)
	median pulse (outpatient)
	median pulse (inpatient)
Document counts	PL count
	DS count
	HPC count
	All doc count
Regular expression counts and normalized descriptors	PL HTN RegEx doc matches
	DS HTN RegEx doc matches
	HPC HTN RegEx doc matches
	All doc HTN RegEx doc matches
	PL HTN RegEx doc matches doc normalized
	DS HTN RegEx doc matches doc normalized
	HPC HTN RegEx doc matches doc normalized
	All doc HTN RegEx doc matches doc normalized

Category	Column description
NLP-based concepts and normalized descriptors	PL concept count
	DS concept count
	HPC concept count
	All doc concept count
	PL HTN concept count
	DS HTN concept count
	HPC HTN concept count
	All doc HTN concept count
	PL HTN concept count doc normalized
	DS HTN concept count doc normalized
	HPC HTN concept count doc normalized
	All doc HTN concept count doc normalized
	PL HTN concept count concept normalized
	DS HTN concept count concept normalized
	HPC HTN concept count concept normalized
	All doc HTN concept count concept normalized

Combining categories improves performance



Random Forests are a robust and easy-to-use machine learning method

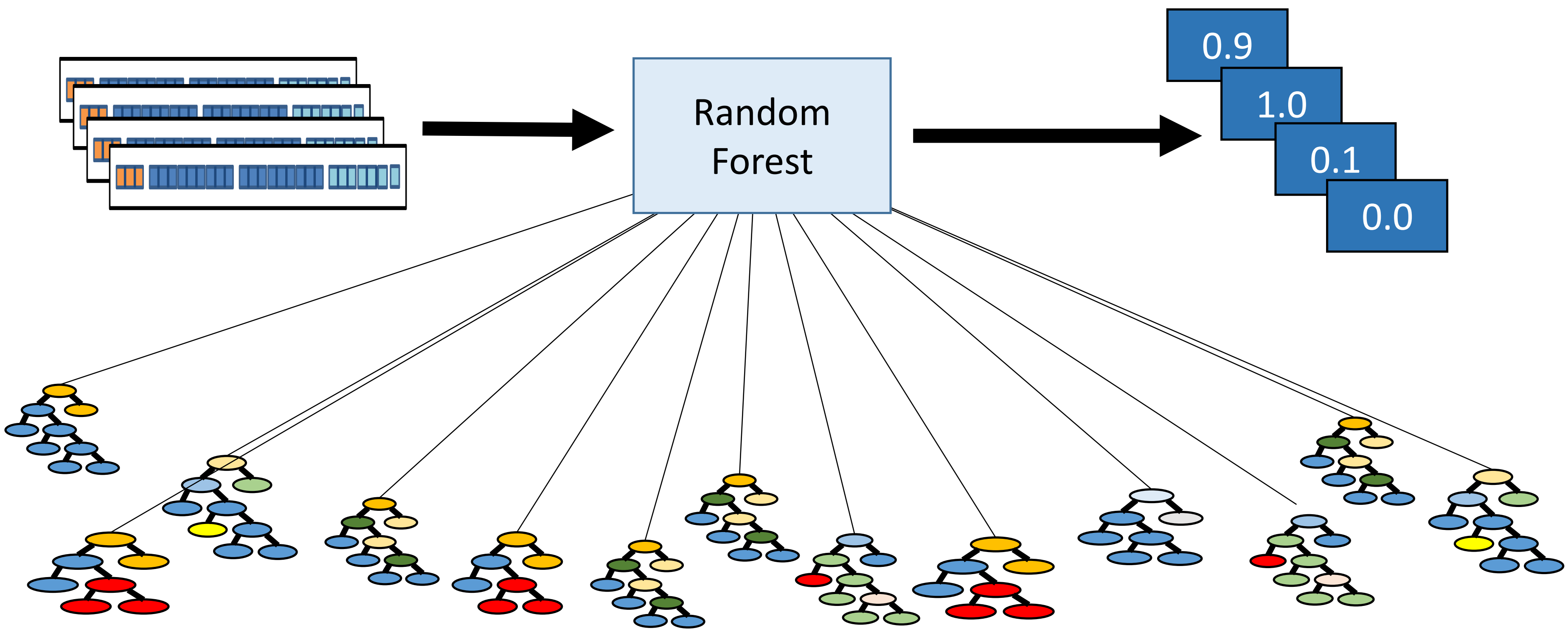
Pros

- More stable than decision trees
- Good on “messy” data
- Relatively few parameters to optimize
- Suggests descriptor “importance”

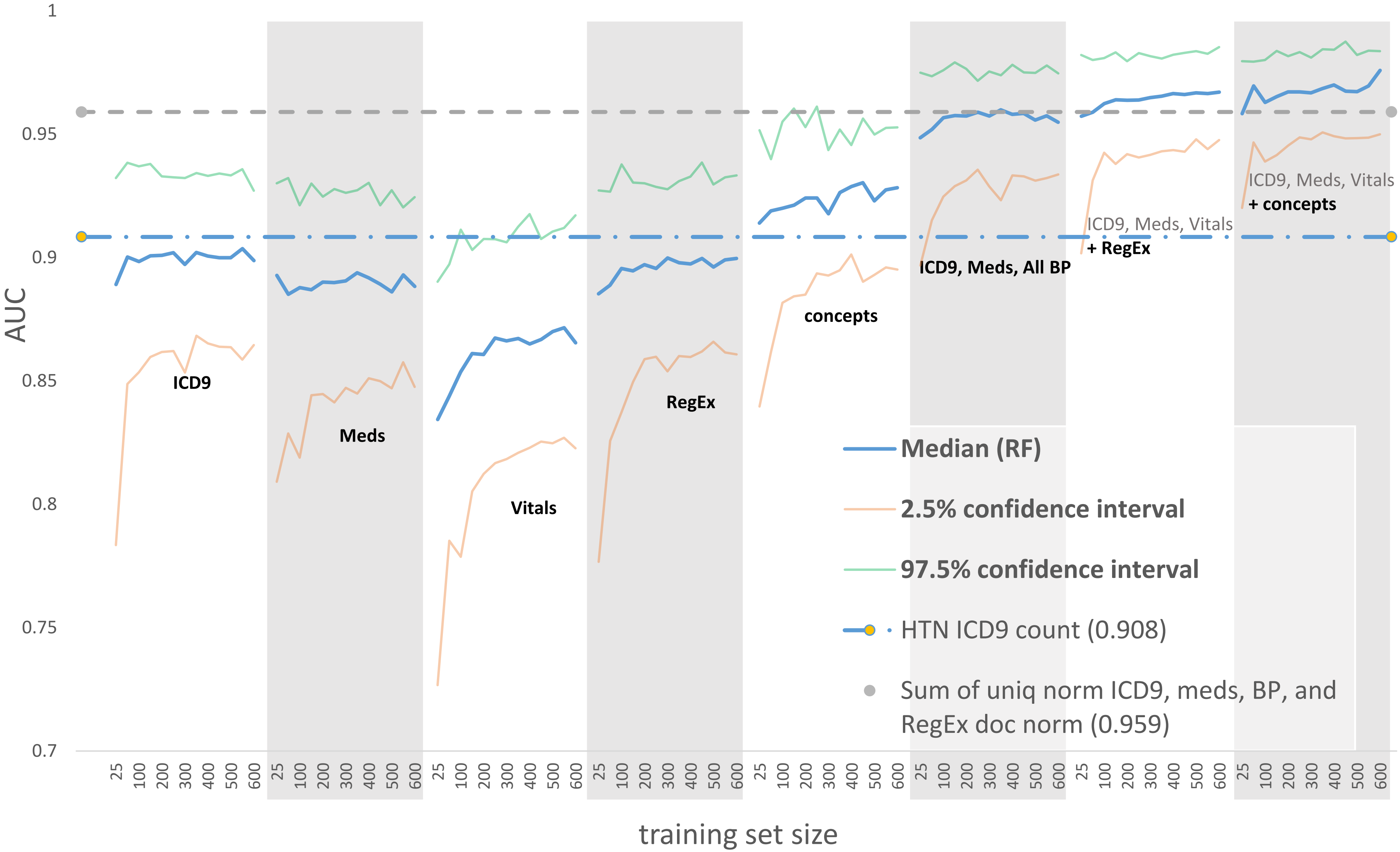
Cons

- More time intensive to train
- “Black box”
- May not capture complex relationships
- Random method

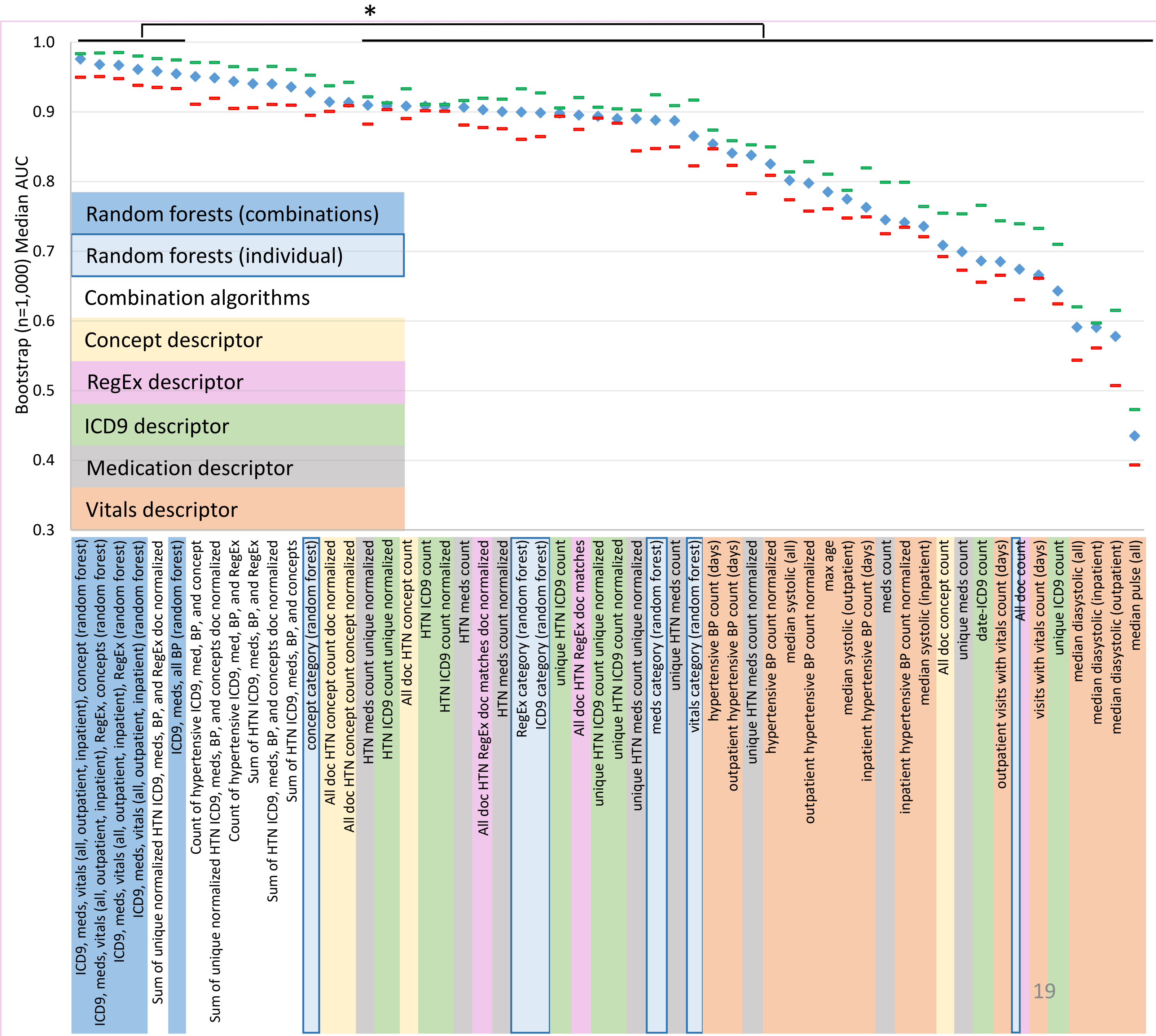
Random forest result is the mode of the trees



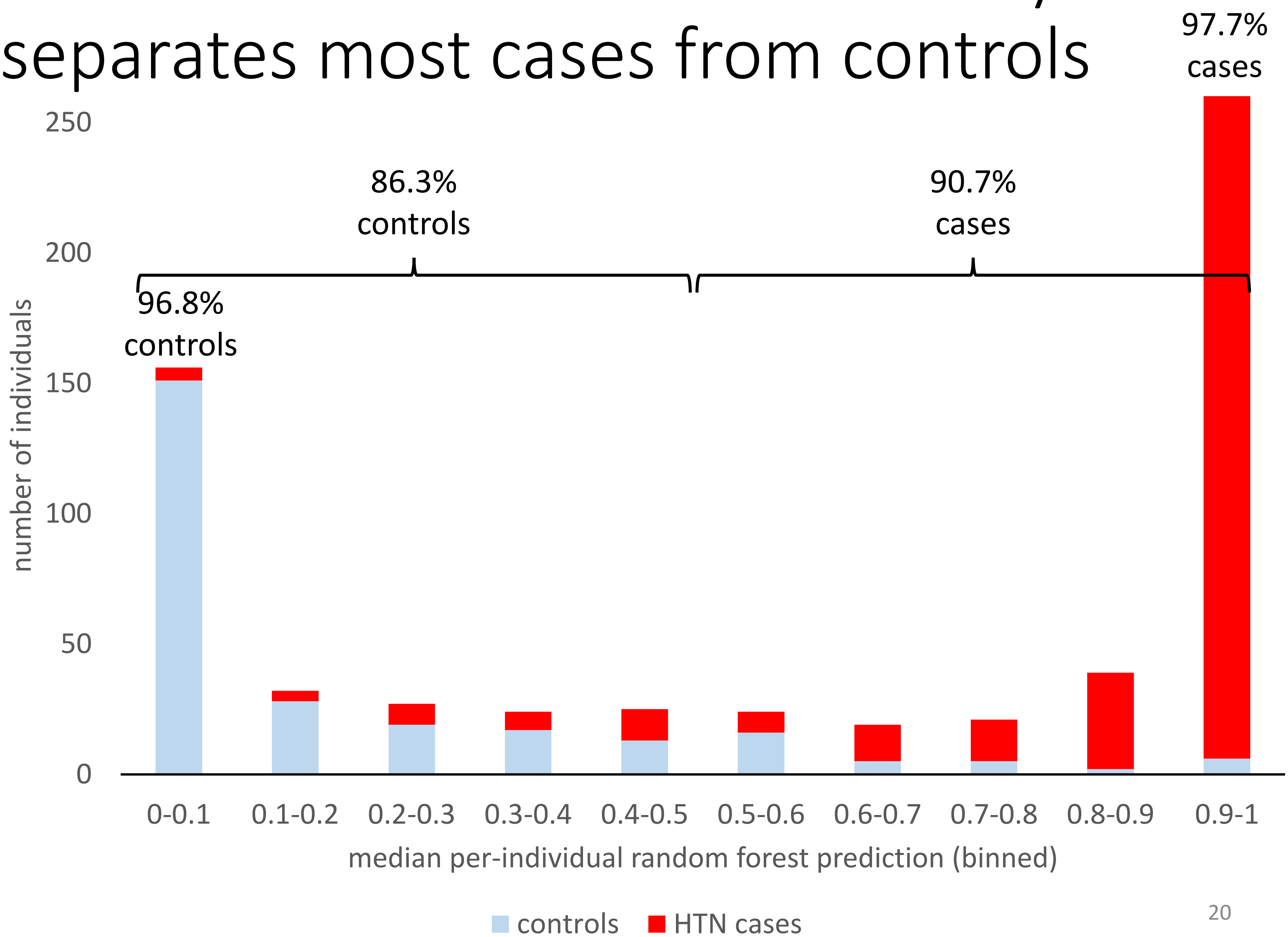
Random forests trained on combinations of data achieve the highest AUC



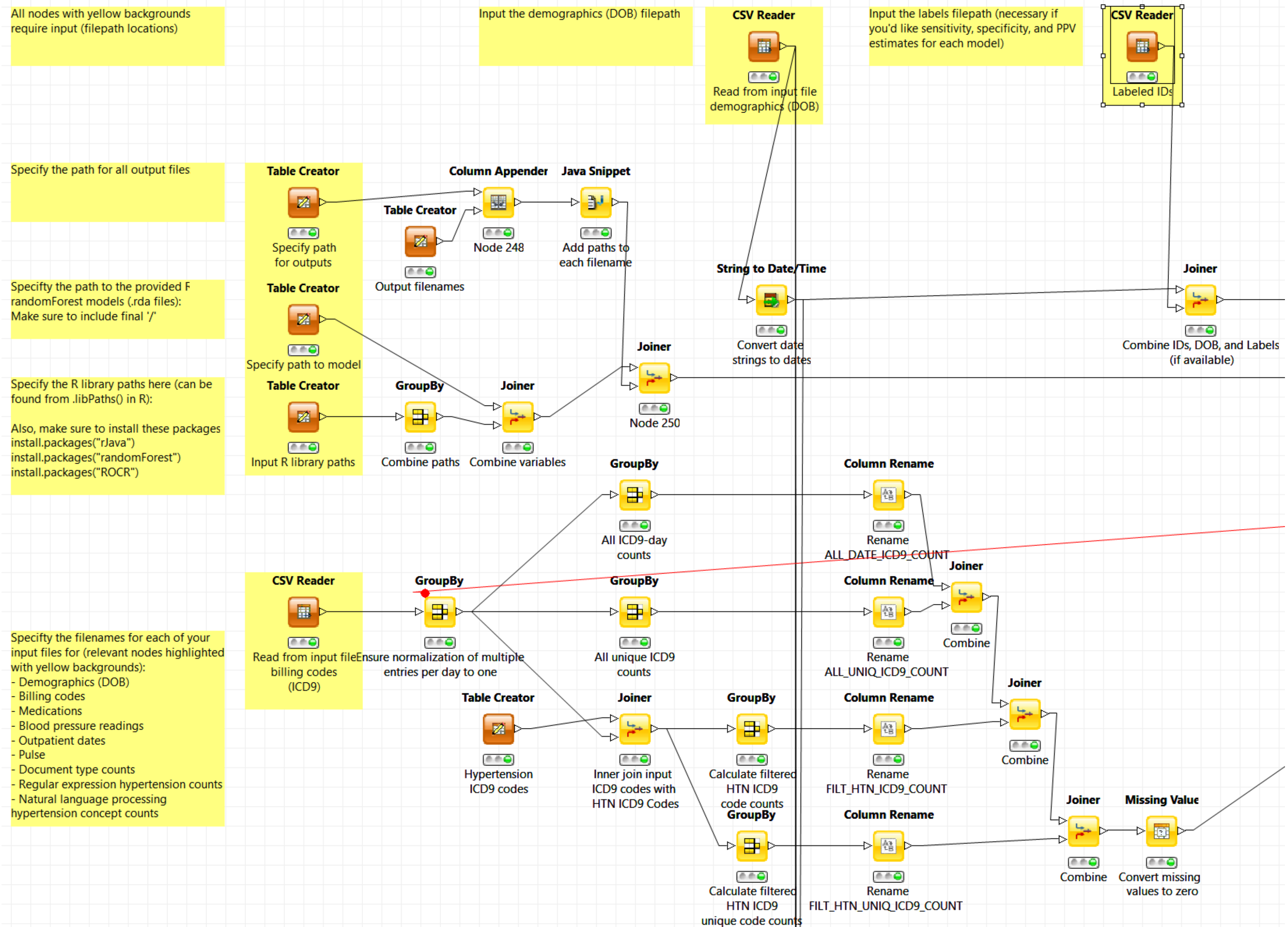
Vitals – Includes all blood pressure readings (inpatient, outpatient, and combined) and pulse
BP - Includes all blood pressure readings (combined only)


*
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Bootstrap (n=1,000) Median AUC
Random forests (combinations)
Random forests (individual)
Combination algorithms
Concept descriptor
RegEx descriptor
ICD9 descriptor
Medication descriptor
Vitals descriptor
ICD9, meds, vitals (all, outpatient, inpatient), concept (random forest)
ICD9, meds, vitals (all, outpatient, inpatient), RegEx, concepts (random forest)
ICD9, meds, vitals (all, outpatient, inpatient), RegEx (random forest)
ICD9, meds, vitals (all, outpatient, inpatient) (random forest)
Sum of unique normalized HTN ICD9, meds, BP, and RegEx doc normalized
ICD9, meds, all BP (random forest)
Count of hypertensive ICD9, med, BP, and concept
Sum of unique normalized HTN ICD9, meds, BP, and concepts doc normalized
Count of hypertensive ICD9, med, BP, and RegEx
Sum of HTN ICD9, meds, BP, and RegEx
Sum of HTN ICD9, meds, BP, and concepts doc normalized
Sum of HTN ICD9, meds, BP, and concepts
concept category (random forest)
All doc HTN concept count doc normalized
All doc HTN concept count concept normalized
HTN meds count unique normalized
HTN ICD9 count unique normalized
All doc HTN concept count
HTN ICD9 count
HTN ICD9 count normalized
HTN meds count
All doc HTN RegEx doc matches doc normalized
HTN meds count normalized
RegEx category (random forest)
ICD9 category (random forest)
unique HTN ICD9 count
All doc HTN RegEx doc matches
unique HTN ICD9 count unique normalized
unique HTN ICD9 count normalized
unique HTN meds count unique normalized
meds category (random forest)
unique HTN meds count
vitals category (random forest)
hypertensive BP count (days)
outpatient hypertensive BP count (days)
unique HTN meds count normalized
hypertensive BP count normalized
median systolic (all)
outpatient hypertensive BP count normalized
max age
median systolic (outpatient)
inpatient hypertensive BP count (days)
meds count
inpatient hypertensive BP count normalized
median systolic (inpatient)
All doc concept count
unique meds count
date-ICD9 count
outpatient visits with vitals count (days)
All doc count
visits with vitals count (days)
unique ICD9 count
median diasystolic (all)
median diasystolic (inpatient)
median diasystolic (outpatient)
median pulse (all)

The best random forest effectively separates most cases from controls



We created a portable Konstanz Information Miner (KNIME) module



Machine learning and summing algorithms successfully replicate at Marshfield Clinic

	Model	Vanderbilt (n=631) AUC	Vanderbilt (n=631) Sens.	Vanderbilt (n=631) PPV	Replication Marshfield (n=100) AUC	Replication Marshfield (n=100) Sens.	Replication Marshfield (n=100) PPV
Random Forests	ICD9, meds, all BP (random forest)	0.955	0.844	0.954	0.922	0.966	0.919
	ICD9, meds, all vitals (random forest)	0.961	0.858	0.954	0.910	0.966	0.905
	ICD9, meds, all vitals, RegEx (random forest)*	0.967	0.866	0.954	**0.934**	0.966	0.934
	ICD9, meds, all vitals, concept (random forest)	**0.976**	0.902	0.952	0.873	0.966	0.864
	ICD9, meds, all vitals, RegEx, concepts (random forest)*	0.968	0.877	0.954	0.898	0.966	0.891

***Marshfield Clinic inputs to random forest models did not include regular expression (RegEx) information**

Machine learning and summing algorithms successfully replicate at Marshfield Clinic

	Model	Vanderbilt (n=631) AUC	Vanderbilt (n=631) Sens.	Vanderbilt (n=631) PPV	Replication Marshfield (n=100) AUC	Replication Marshfield (n=100) Sens.	Replication Marshfield (n=100) PPV
Random Forests	ICD9, meds, all BP (random forest)	0.955	0.844	0.954	0.922	0.966	0.919
	ICD9, meds, all vitals (random forest)	0.961	0.858	0.954	0.910	0.966	0.905
	ICD9, meds, all vitals, RegEx (random forest)*	0.967	0.866	0.954	**0.934**	0.966	0.934
	ICD9, meds, all vitals, concept (random forest)	**0.976**	0.902	0.952	0.873	0.966	0.864
	ICD9, meds, all vitals, RegEx, concepts (random forest)*	0.968	0.877	0.954	0.898	0.966	0.891
Category Counts	Count of HTN ICD9, med, and BP 2 of 3	0.833	0.952	0.822	0.646	1.000	0.670
	Count of HTN ICD9, med, and BP 3 of 3	0.877	0.798	0.967	**0.914**	0.949	0.918
	Count of HTN ICD9, med, BP, and concept 3 of 4	**0.910**	0.925	0.924	0.711	0.983	0.716

***Marshfield Clinic inputs to random forest models did not include regular expression (RegEx) information**

Machine learning and summing algorithms successfully replicate at Marshfield Clinic

	Model	Vanderbilt (n=631) AUC	Sens.	PPV	Replication Marshfield (n=100) AUC	Sens.	PPV
Random Forests	ICD9, meds, all BP (random forest)	0.955	0.844	0.954	0.922	0.966	0.919
	ICD9, meds, all vitals (random forest)	0.961	0.858	0.954	0.910	0.966	0.905
	ICD9, meds, all vitals, RegEx (random forest)*	0.967	0.866	0.954	**0.934**	0.966	0.934
	ICD9, meds, all vitals, concept (random forest)	**0.976**	0.902	0.952	0.873	0.966	0.864
	ICD9, meds, all vitals, RegEx, concepts (random forest)*	0.968	0.877	0.954	0.898	0.966	0.891
Category Counts	Count of HTN ICD9, med, and BP 2 of 3	0.833	0.952	0.822	0.646	1.000	0.670
	Count of HTN ICD9, med, and BP 3 of 3	0.877	0.798	0.967	**0.914**	0.949	0.918
	Count of HTN ICD9, med, BP, and concept 3 of 4	**0.910**	0.925	0.924	0.711	0.983	0.716
Category Sums	**Sum of normalized HTN ICD9, meds, and BP**	0.915	1.000	0.673	**0.949**	1.000	0.702
	Sum of normalized HTN ICD9, meds, BP, and concept	**0.929**	1.000	0.663	**0.949**	1.000	0.702

***Marshfield Clinic inputs to random forest models did not include regular expression (RegEx) information**

Conclusions

- Combining categories improves performance
  - 0.976 AUC for best random forest
  - 0.908 AUC billing codes
  - 0.854 AUC hypertensive blood pressure readings

- Concepts are the most valuable individual category on Vanderbilt data

- All random forest models and summing algorithms performed comparably on Marshfield Clinic data