Neural topic modelling for automatic phenotyping from EHR data and PheWAS analysis

Yue Li
Assistant Professor
School of Computer Science
Mila - Quebec AI Institute
McGill Quantitative Life Sciences
McGill University

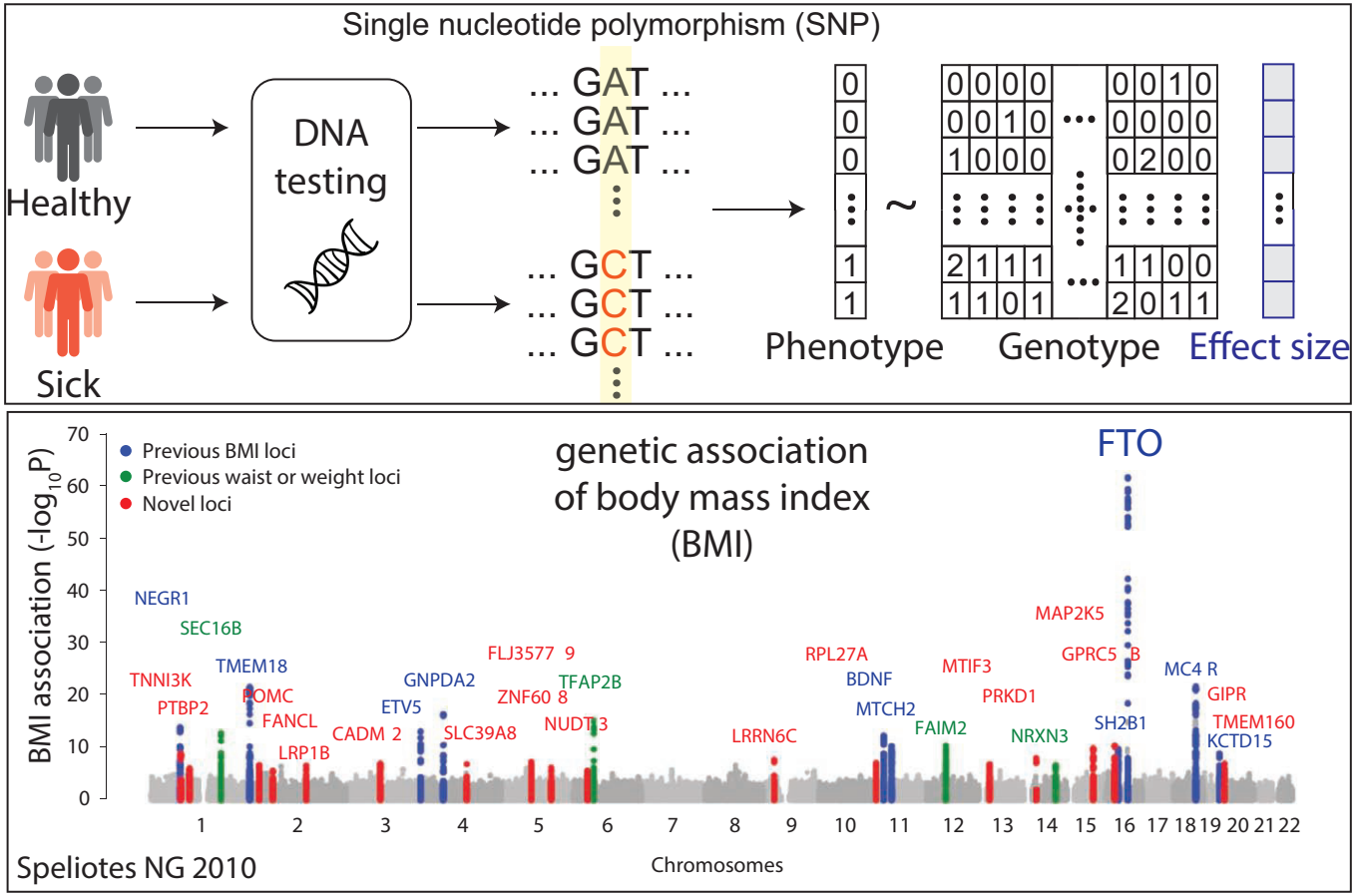
Outline

Phenome-wide association studies using EHR data

Graph-informed EHR topic modeling
- GETM: Graph-embedded topic model
- GAT-ETM: an end-to-end graph-topic model

Phecode-guided EHR topic modeling
- MixEHR-guided: a phecode-guided multimodal topic model
- MixEHR-seed: a seed-guided VAE-EM hybrid topic model

Genome-wide association studies (GWAS)



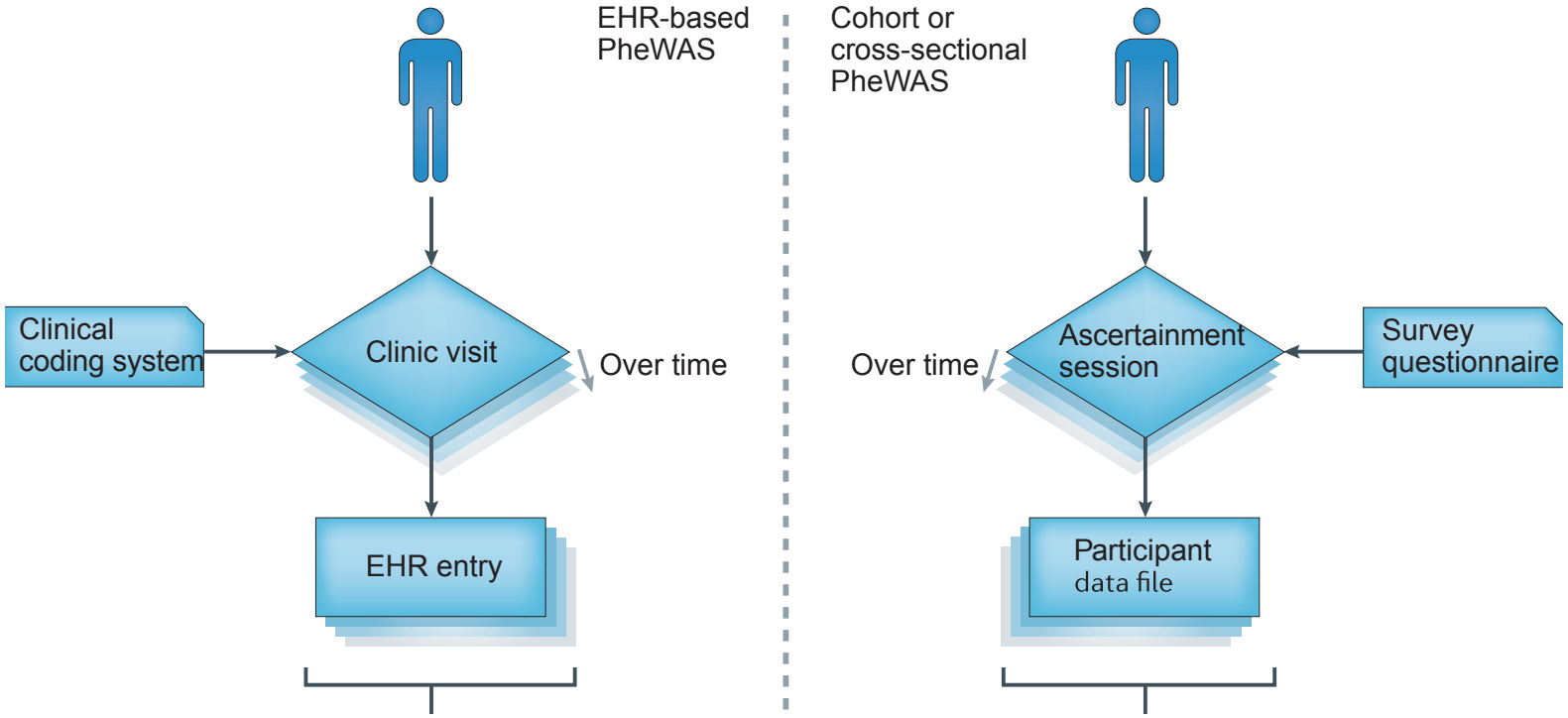
- Only one phenotype is investigated yet many traits share causal SNPs
- Many genetically correlated or the upstream causal phenotypes are often unknown

Phenome-wide association studies (PheWAS) design

Step 1: "deep" phenotyping [Bush et al., 2016]

Prospective or observational

Acertainment of group of phenotypes



Patient EHR-based measures

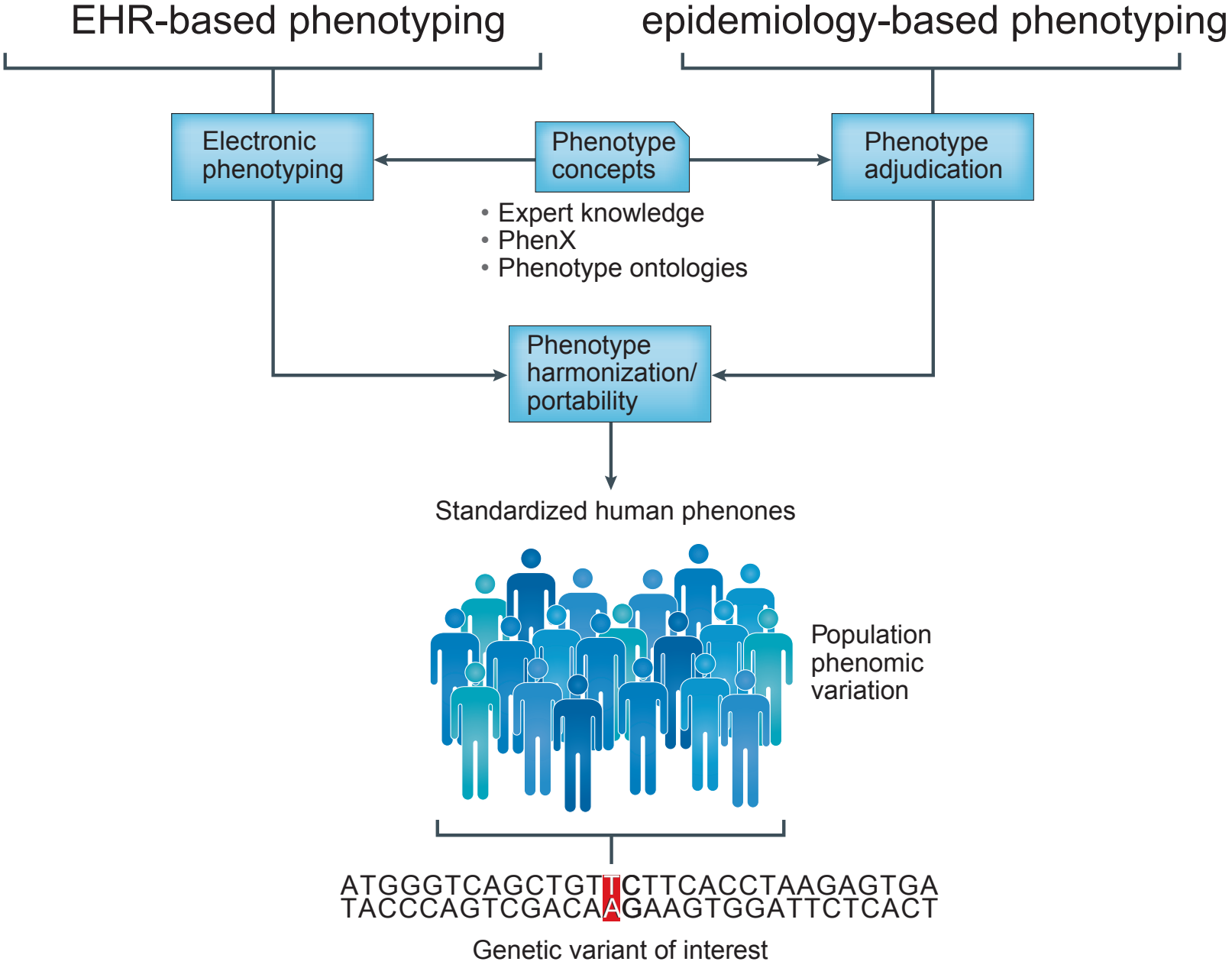
250.0 T2DM	Yes
411.1 coronary syndrome	Yes
414.01 coronary artery disease	No
278.01 obesity	Yes
Alanine aminotransferase	15.6 units per l
Blood albumin	3.7 g per dl
Aspartate aminotrasferase	22 units per l
Bicarbonate (HCO_3)	24 mEq per l
Carbon dioxide (CO_2)	27 mEq per l
Blood cholesterol	240 mg per dl
Blood creatinine	1.2 mg per dl

Participant epidemiology-based measures

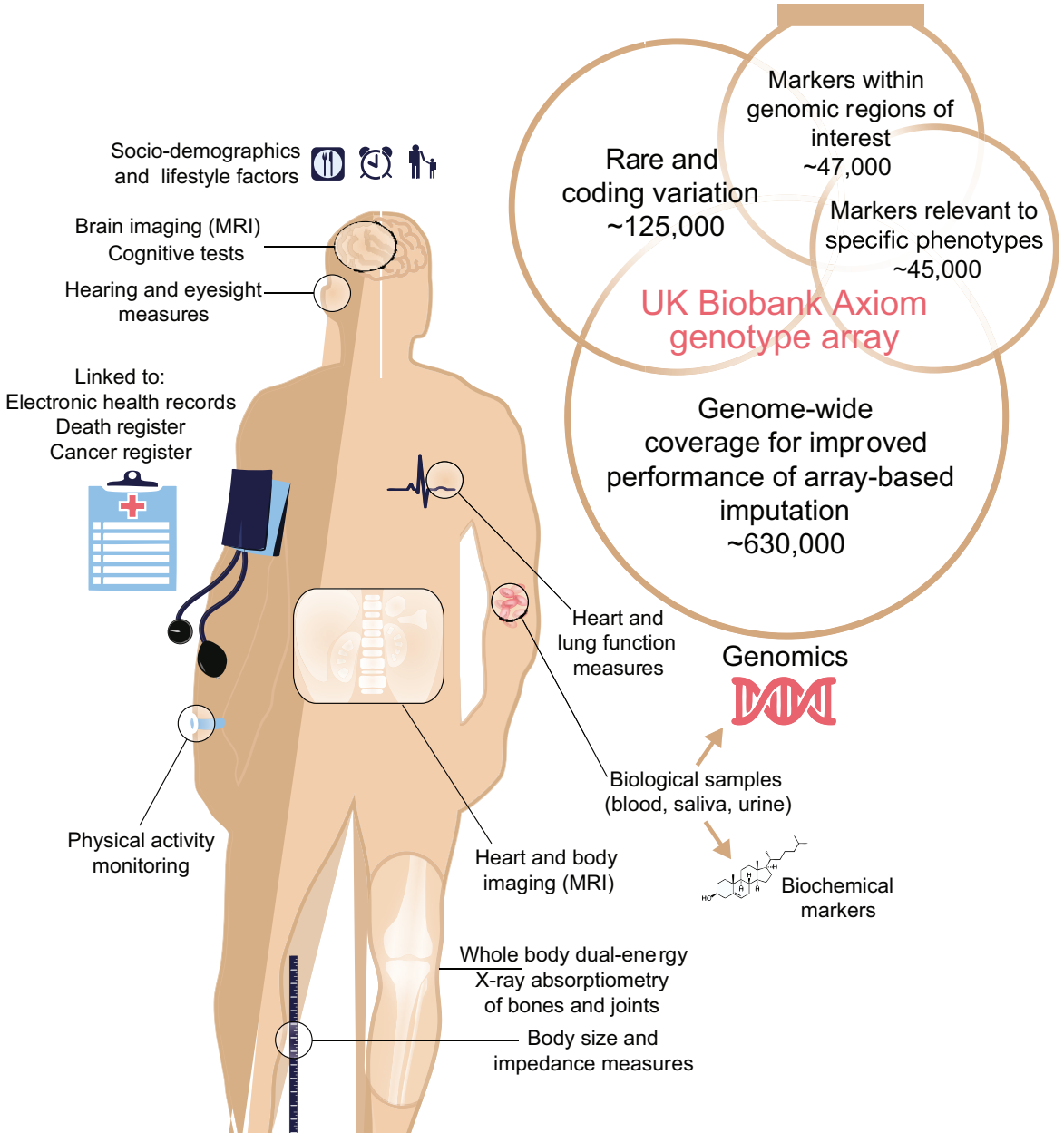
Ever had diabetes?	Yes
Cancer ever diagnosed?	Yes
Ever smoked?	No
Allergic to gluten?	No
Allergic to peanuts?	Yes
Current weight	240 lb
Current height	5'8"
Green vegetables per week	2–4 servings
Red meat per week	6–8 servings
Blood cholesterol	275 mg per dl
Exercise time per week	30 min

Phenome-wide association studies (PheWAS) design

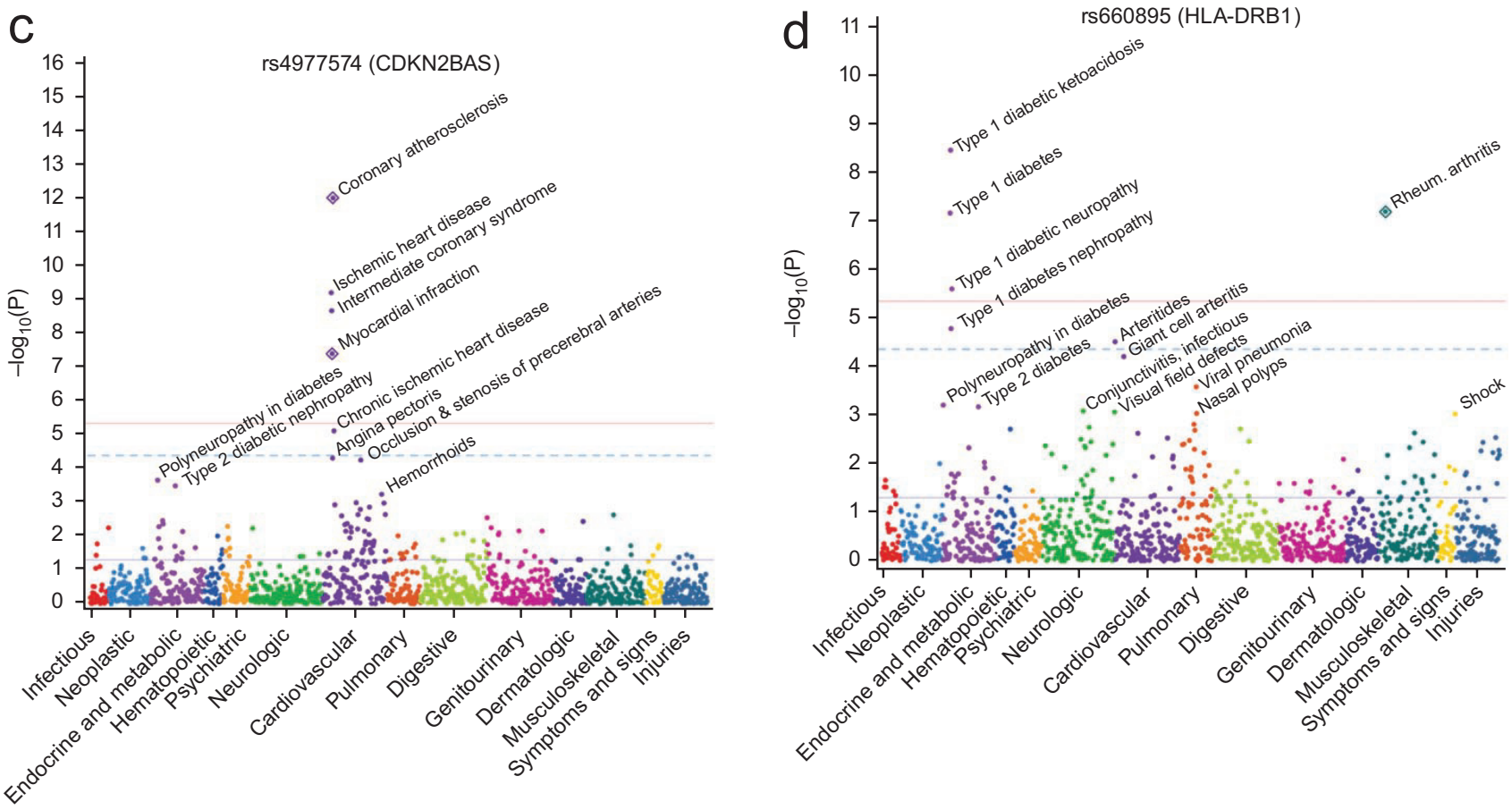
Step 2: genotyping [Bush et al., 2016]



UK Biobank pheno/genotyping of half million individuals [Bycroft et al., 2018]

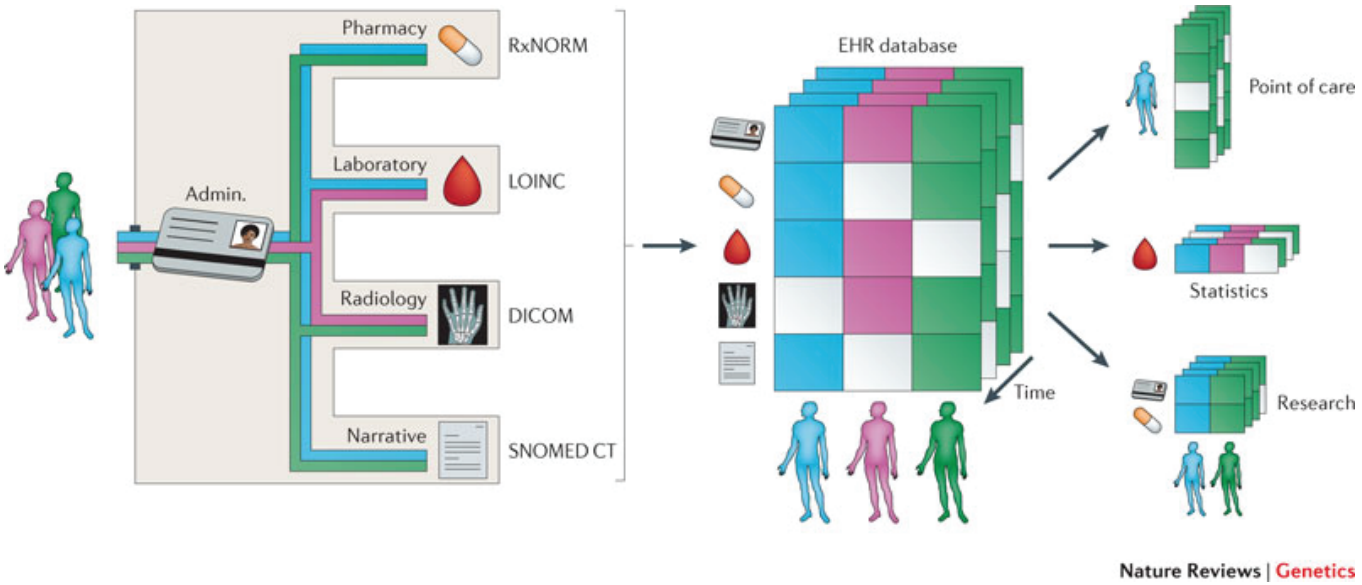


PheWAS reveals pleiotropic SNPs



Electronic Medical Records and Genomics (eMERGE) Network [Denny et al., 2013]

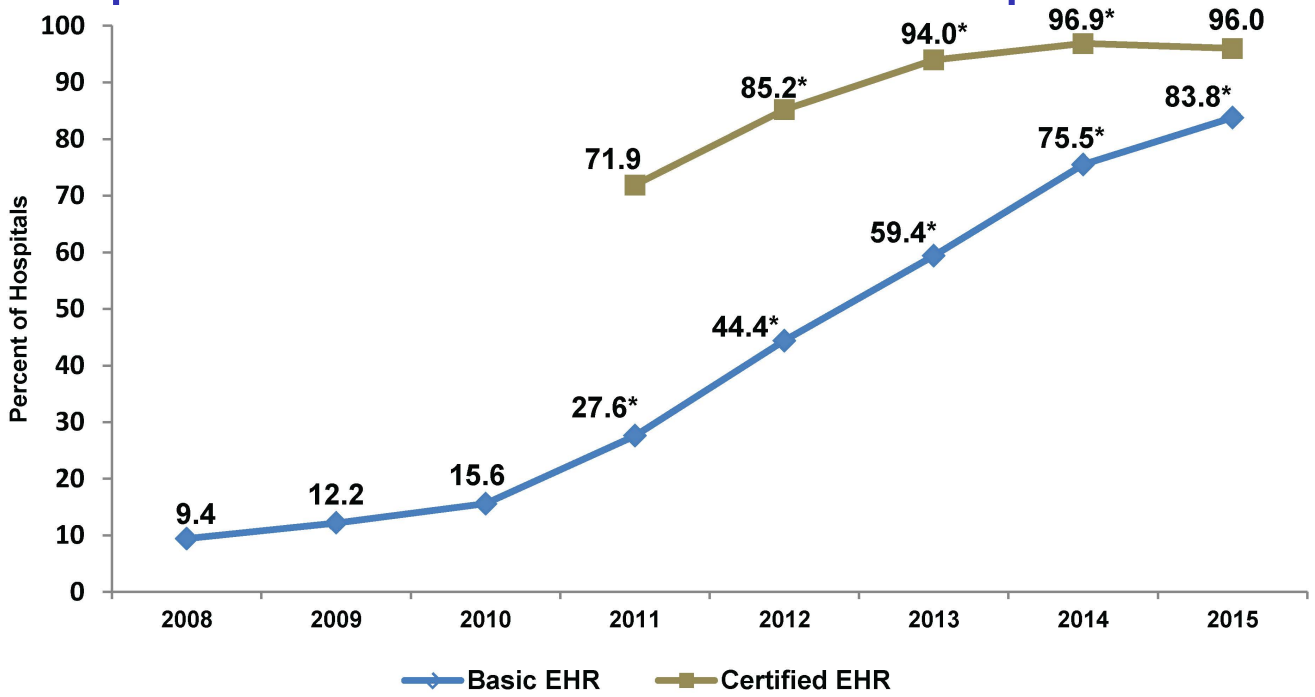
Electronic health records contain rich patient-level data [Jensen et al., 2012]

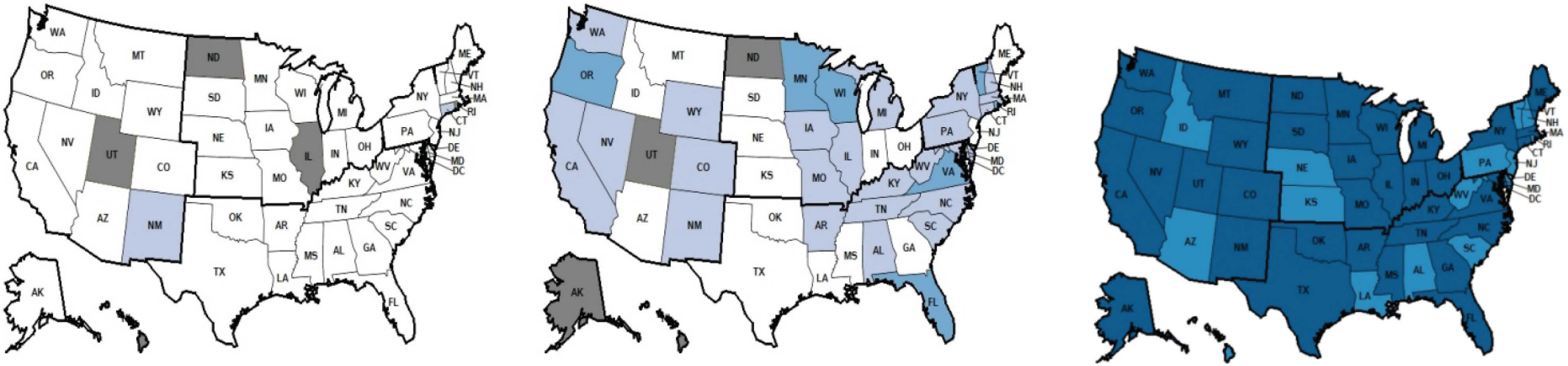


Jensen et al., Nature Rev. Gen. 2012

- Clinical notes (unstructured free-form text)
- Billing code: International Classification of Disease (ICD-CM)
- Billing code: ICD Current Procedural Terminology (ICD-CPT)
- Lab tests: Logical Obs. Identifiers Names & Codes (LOINC)
- Pharmaceutical: Prescription data (RxNorm)
- Radiology, electrocardiogram, MRI, etc

Rapid adoption of EHR in the US hospitals 2008-2015

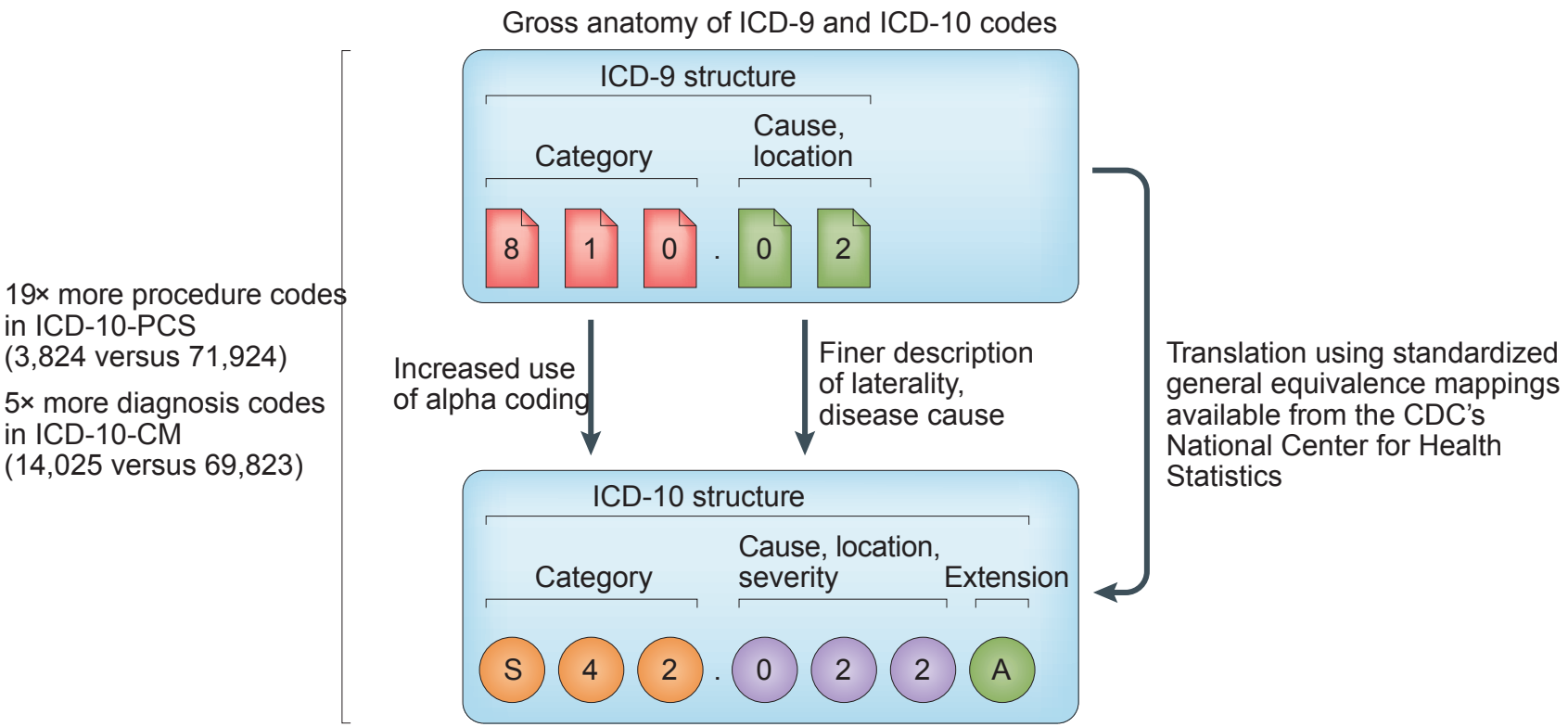




2008 **2011** **2015**

NR | 0-19% | 20-39% | 40-59% | 60-79% | 80-100%

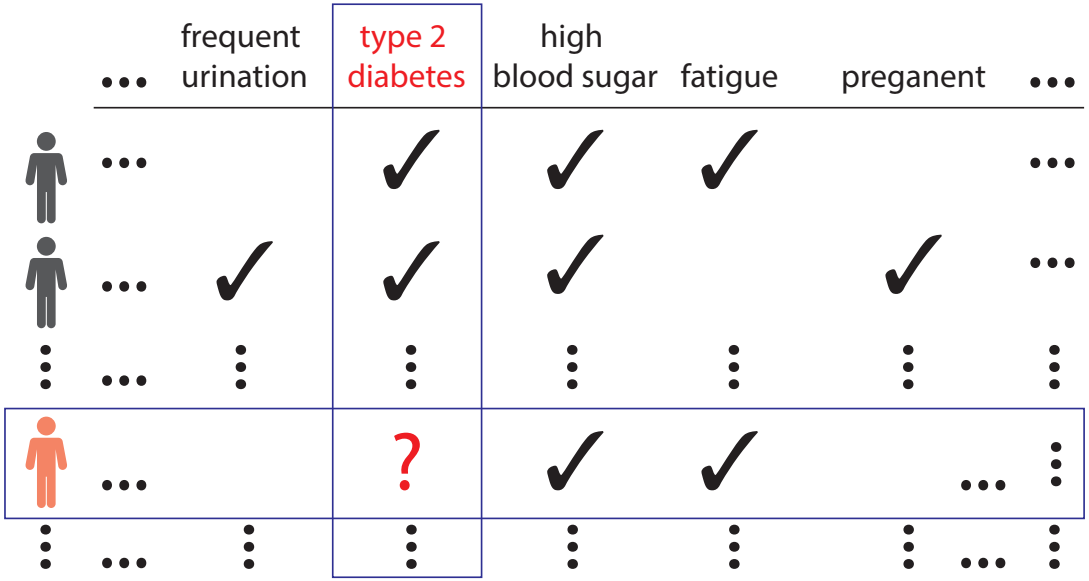
International Classification of Diseases (ICD)



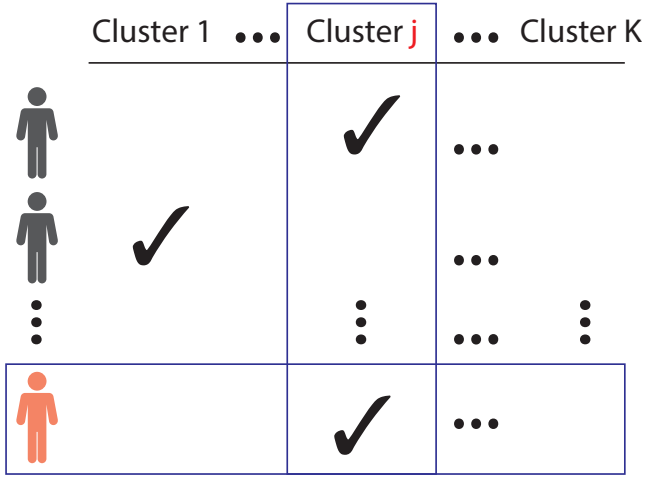
ICD-9 taxonomy: https://icdlist.com/icd-9/index

ICD-10 taxonomy: https://icdlist.com/icd-10/index

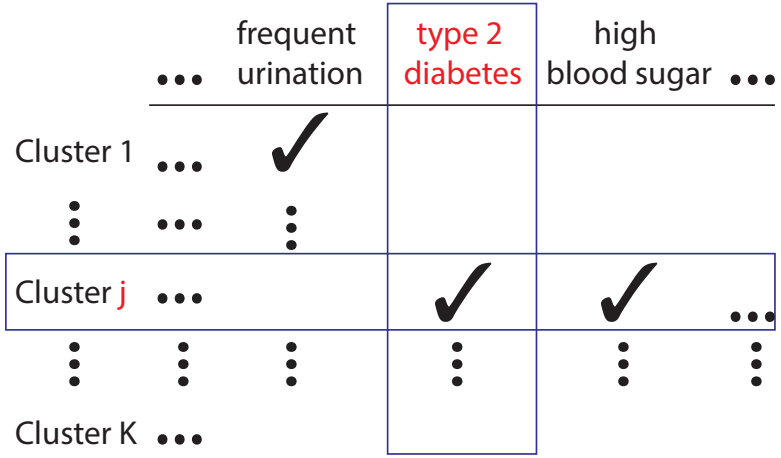
Focus of this talk: unsupervised learning of disease topics to aid phenome-wide association studies



Patient clusters



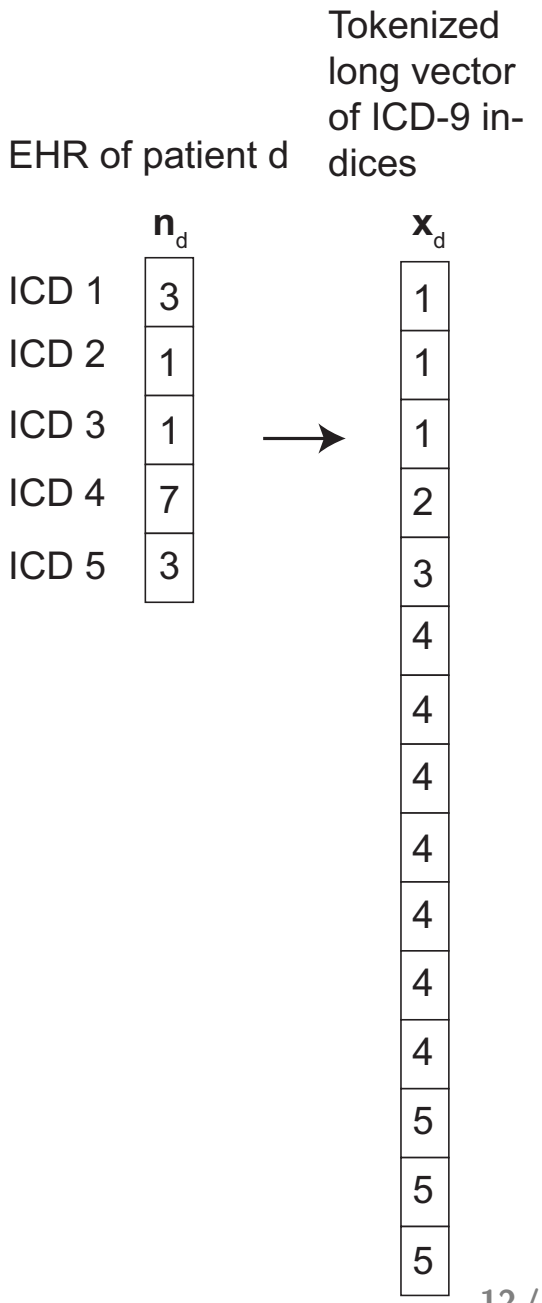
Phenotype clusters



Represent EHR as a bag of words

- We can expand EHR code count vector $\mathbf{n}_d$ in patient d into a long vector of code indices $\mathbf{x}_d$ of length equal to N_d
- Each patient EHR profile is a "document"
- Each record code is a "token"
- The ith token in document d is the ith EHR code from patient d
- The total count of EHR "word" w in patient document d is the sum of the tokens that are word w:

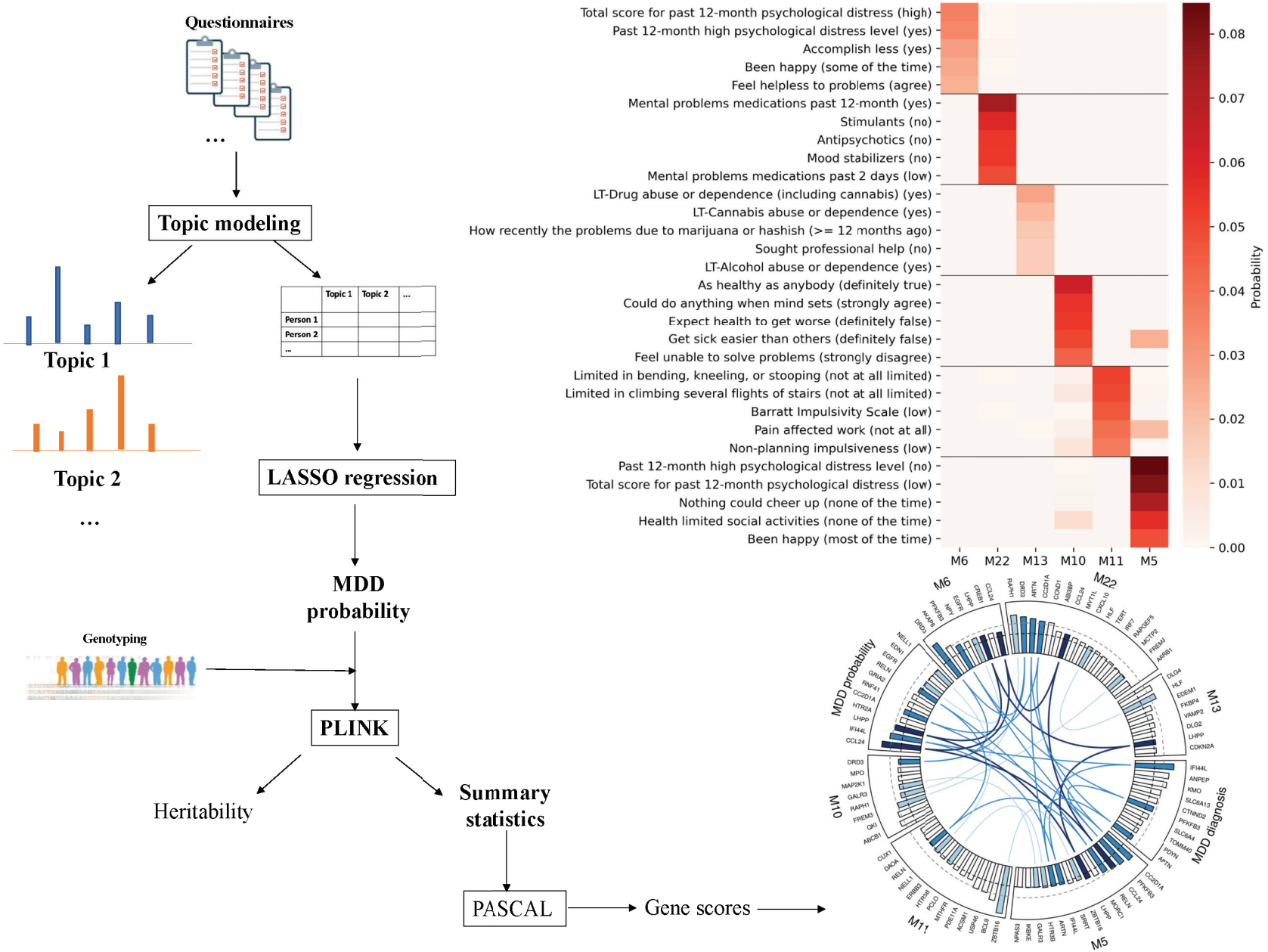
$$n_{wd} = \sum_i [x_{id} = w]$$



Document exhibits mixture of topics [Blei et al., 2003]

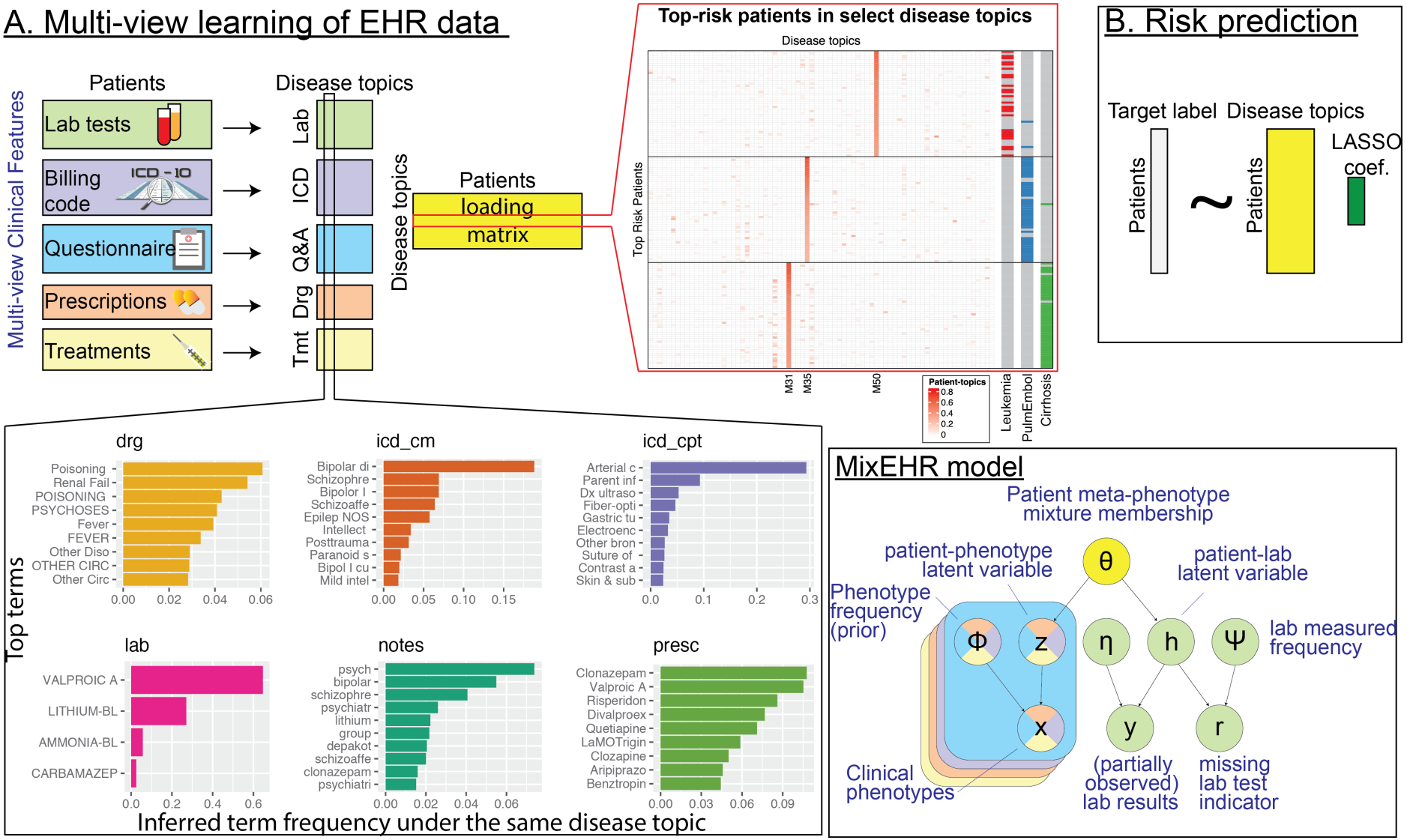
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.



Meng, X.*, Wang, M., . . . , **Li, Y.*** (2022) Integrative PheWAS analysis in risk categorization of major depressive disorder and identifying their associations with genetic variants using a latent topic model approach. *Translational Psychiatry*

Inferring multimodal topics from EHR[1]



[1] **Li, Y.*** et al. Inferring multimodal latent topics from electronic health records. *Nat Commun* 11, 2536 (2020). [Li et al., 2020]

Learning accurate phenotypes from EHR data

Challenges:

- noisy and sparse EHR
- topic interpretability and identifiability

Three strategies (trainees, . . . , *correspondence):

Modelling specialist-specific decision process:

- Song, Z., . . . , **Li, Y.*** (2021) Supervised multi-specialist topic model with applications on large-scale EHR data. In 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB)

Leverage taxonomical knowledge graphs:

- Wang, Y., . . . , **Li, Y.*** (2022) A graph-embedded topic model enables characterization of diverse pain phenotypes among UK Biobank individuals. *iScience* 104390
- Zou Y., . . . , **Li, Y.*** (2022) Modeling electronic health record data using a knowledge-graph-embedded topic model. arXiv.

Leverage expert-curated phenotype definitions as guides:

- Anjuha, Y., . . . , **Li, Y.***. (2022) MixEHR-Guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using EHR. (in rev.)
- Song, Z., . . . , **Li, Y.*** (2022) Automatic phenotyping by a seed-guided topic model. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining

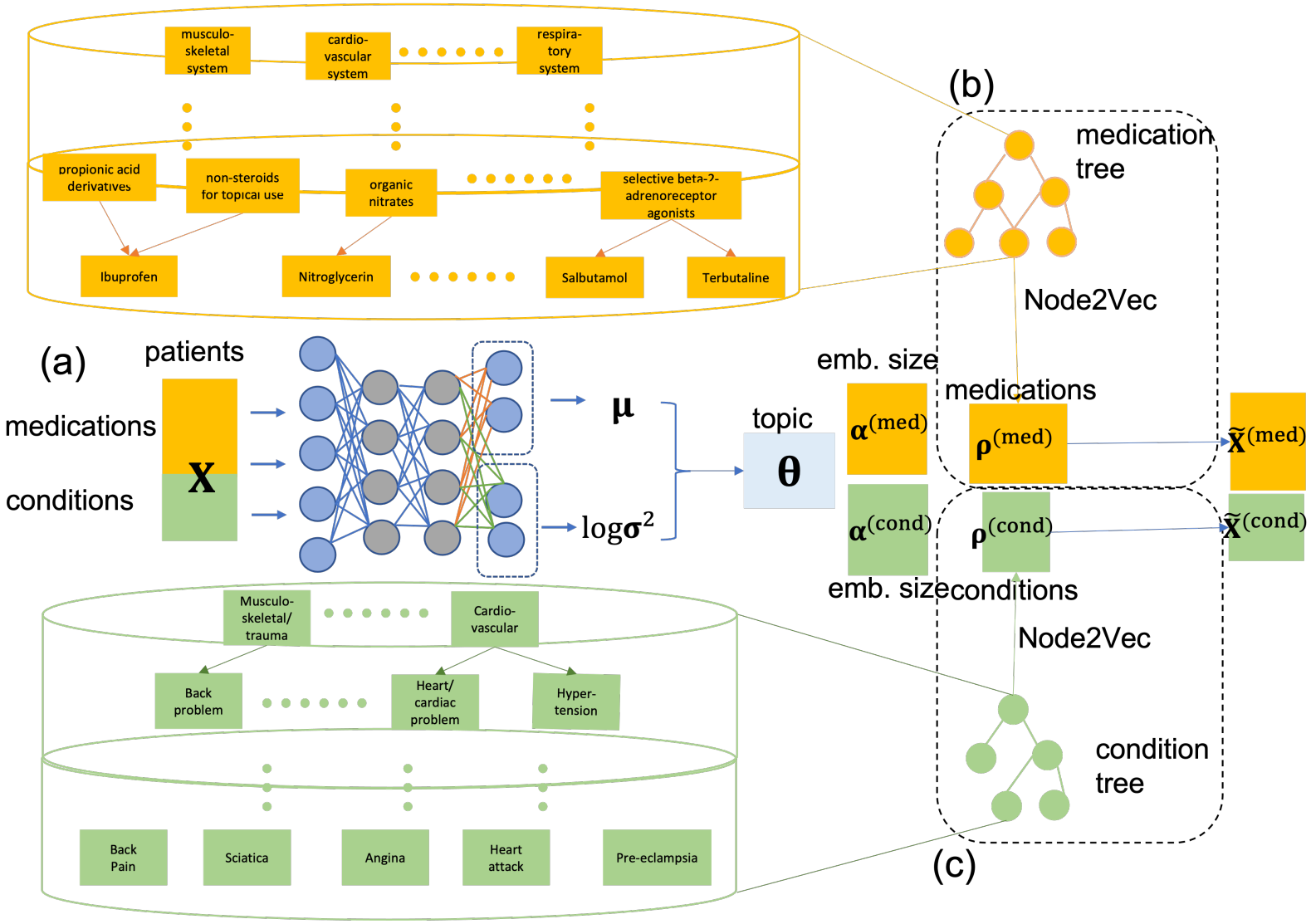
Outline

Phenome-wide association studies using EHR data

Graph-informed EHR topic modeling
- GETM: Graph-embedded topic model
- GAT-ETM: an end-to-end graph-topic model

Phecode-guided EHR topic modeling
- MixEHR-guided: a phecode-guided multimodal topic model
- MixEHR-seed: a seed-guided VAE-EM hybrid topic model

Graph-ETM (GETM)[2]



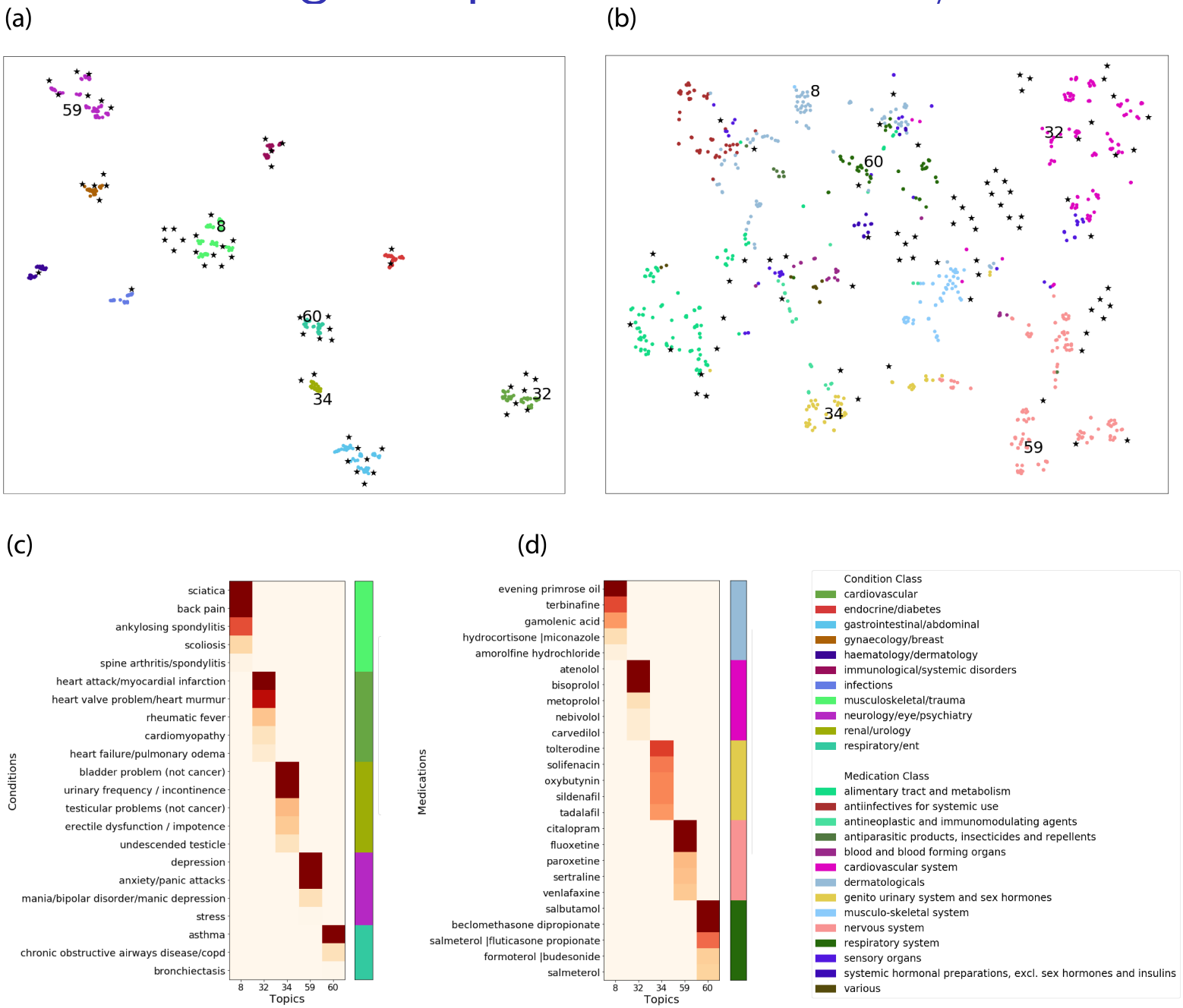
[2]Wang, Y., . . . , & **Li, Y.** (2022) A graph-embedded topic model enables characterization of diverse pain phenotypes among UK Biobank individuals. *iScience*

Modeling conditions and medications data of 450K individuals from UK Biobank [Bycroft et al., 2018]

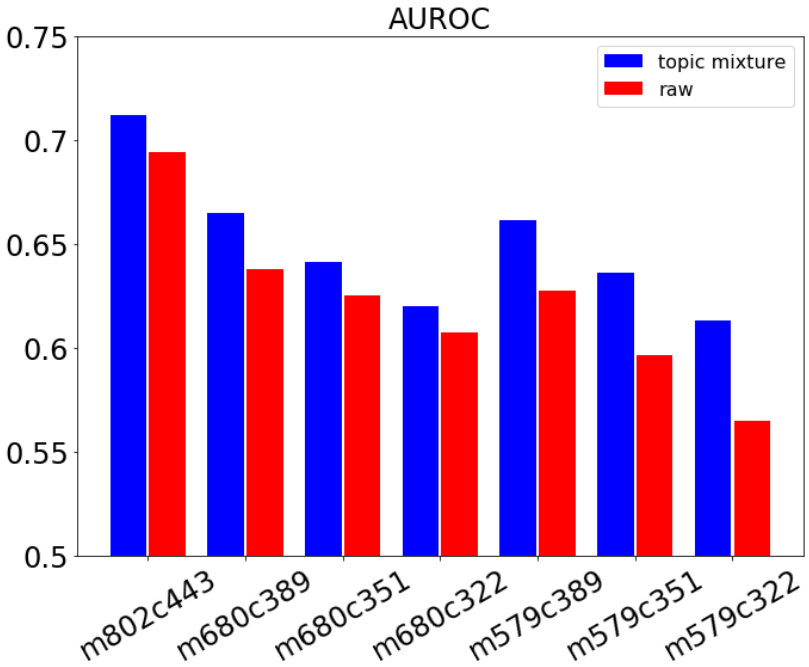
- 457,461 individuals of European descent individuals to reduce confounding caused by different ethic groups
- 802 active ingredients for medications
- 443 phenotypic conditions

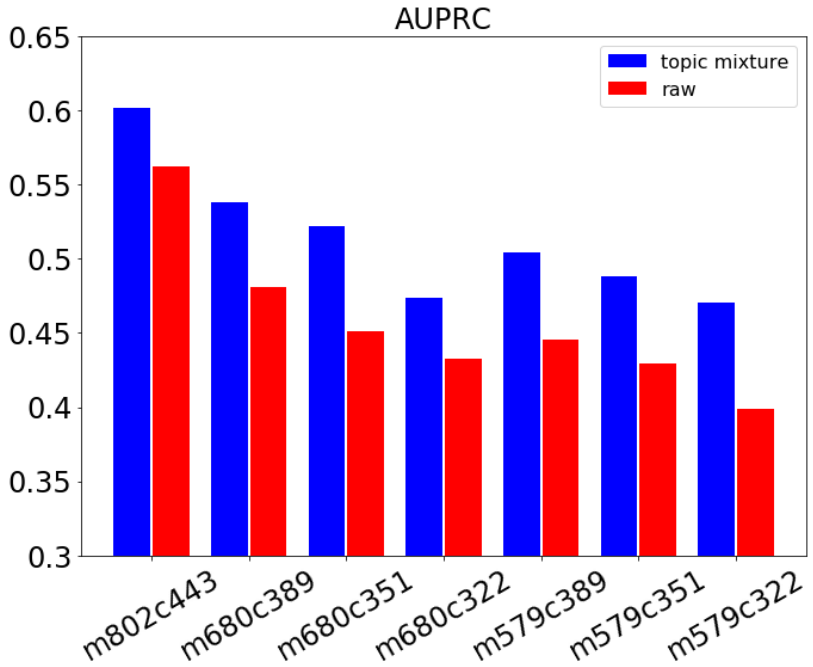
In collaboration with Audrey Grant at the Department of Anesthesia

Visualize embedding of topics and conditions/medications



Prediction performance on chronic musculoskeletal pain

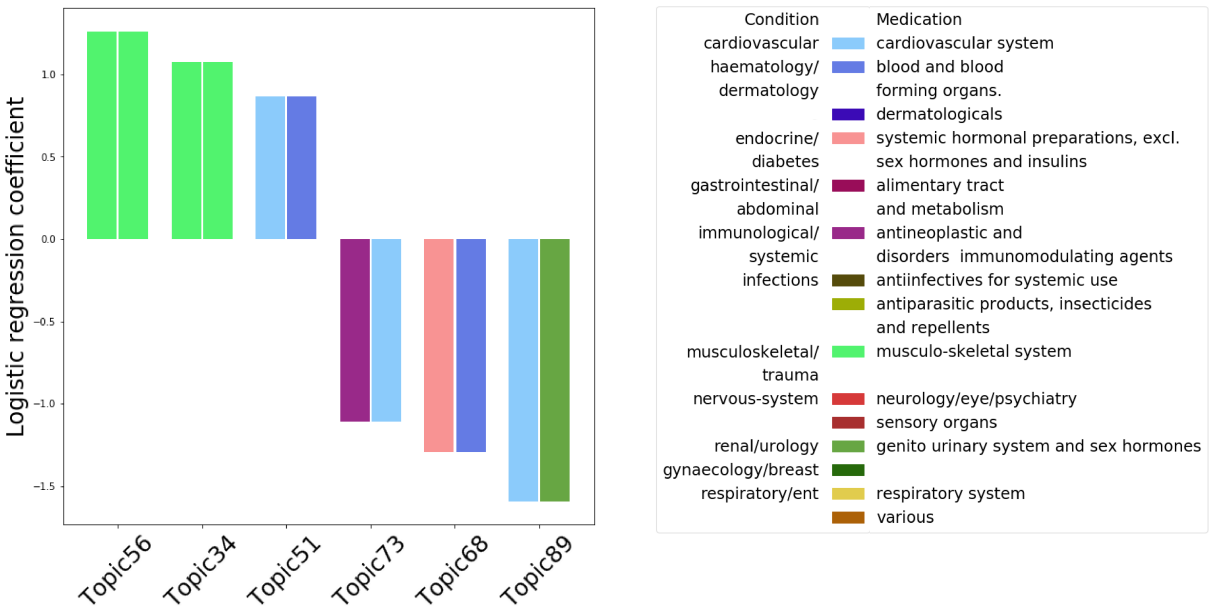




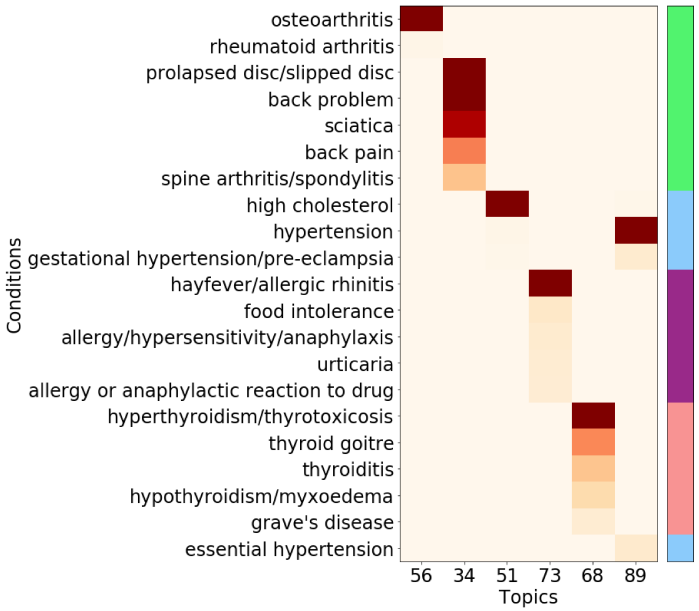
- Logistic regression was performed using θ obtained from GETM with 128 topics to predict CMK pain.
- The baseline used raw conditions and medications data as input features.
- We experimented on seven data configurations with different condition sets and medication sets as indicated by x-axis.

Top topics for chronic musculoskeletal pain

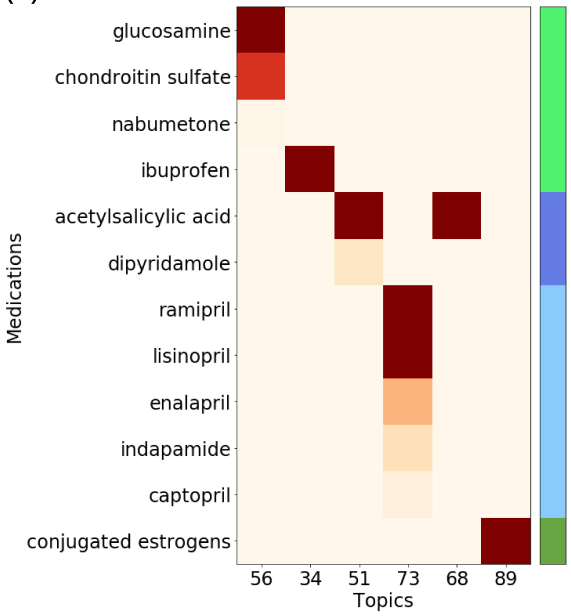
(a)



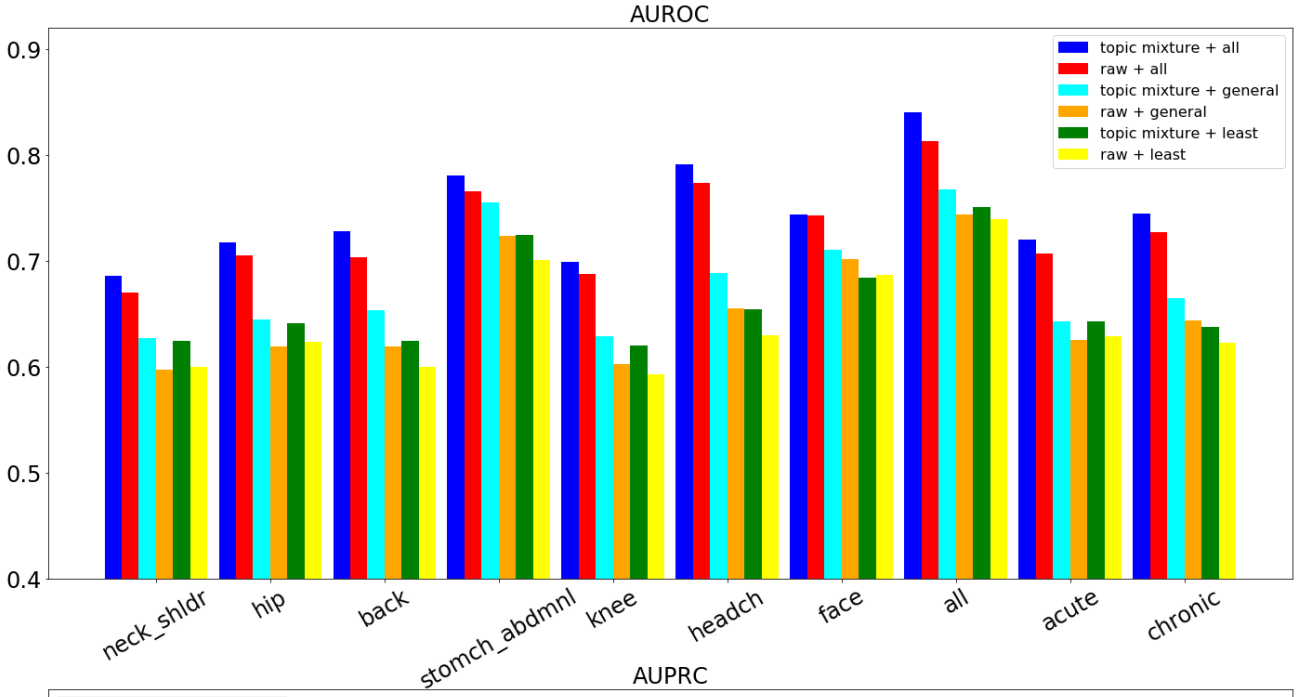
(b)

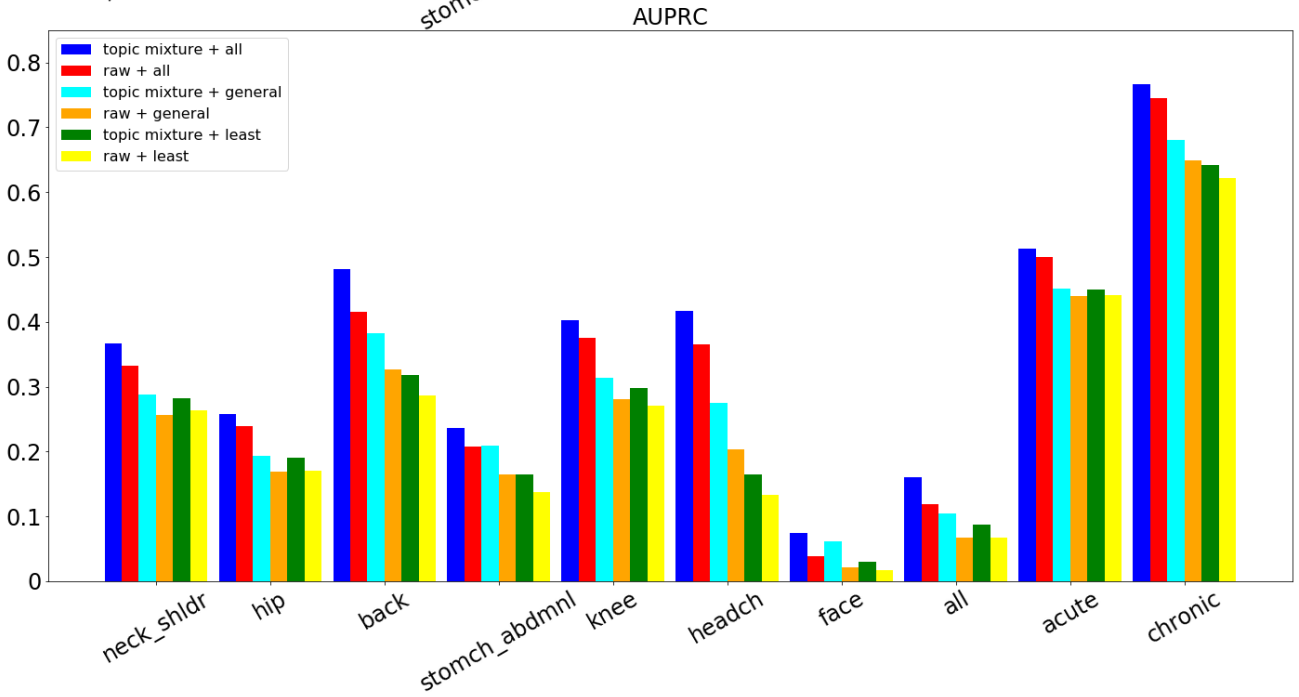


(c)



Predicting chronic pain types on different body sites





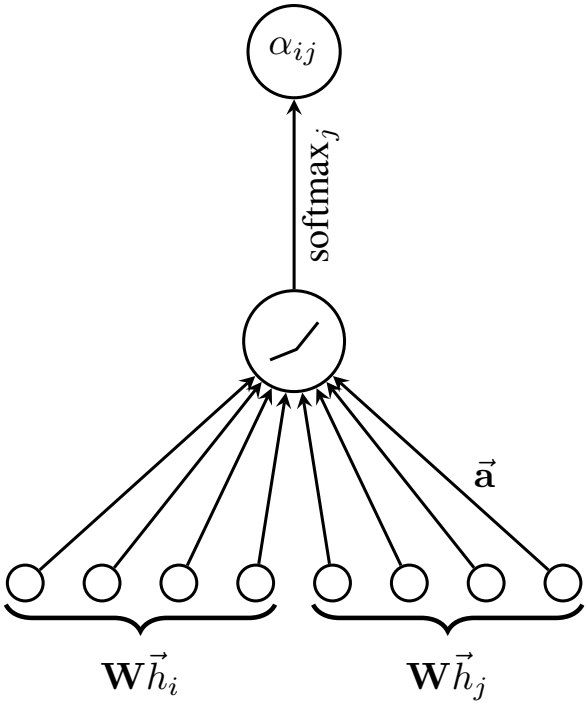
Outline

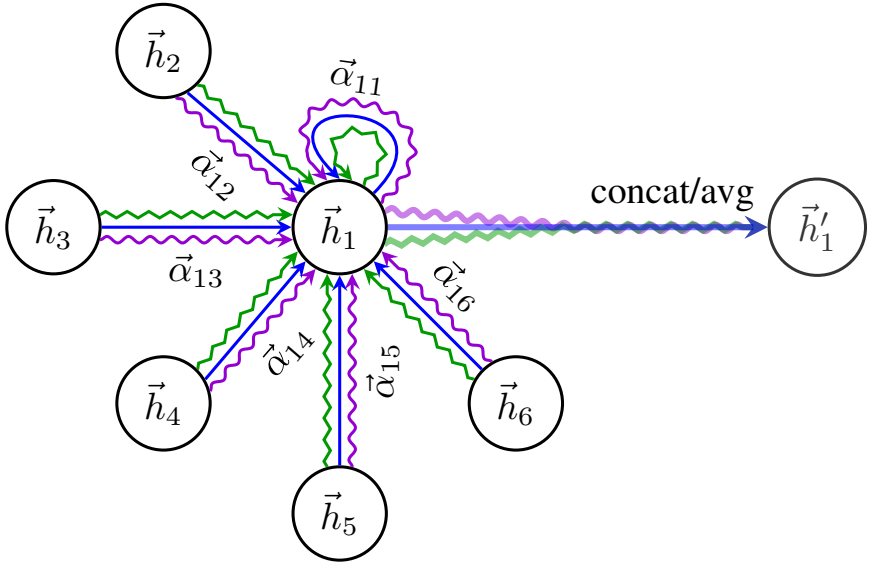
Phenome-wide association studies using EHR data

Graph-informed EHR topic modeling
- GETM: Graph-embedded topic model
- GAT-ETM: an end-to-end graph-topic model

Phecode-guided EHR topic modeling
- MixEHR-guided: a phecode-guided multimodal topic model
- MixEHR-seed: a seed-guided VAE-EM hybrid topic model

Graph Attention neTworks (GAT) [Cucurull et al., 2017]

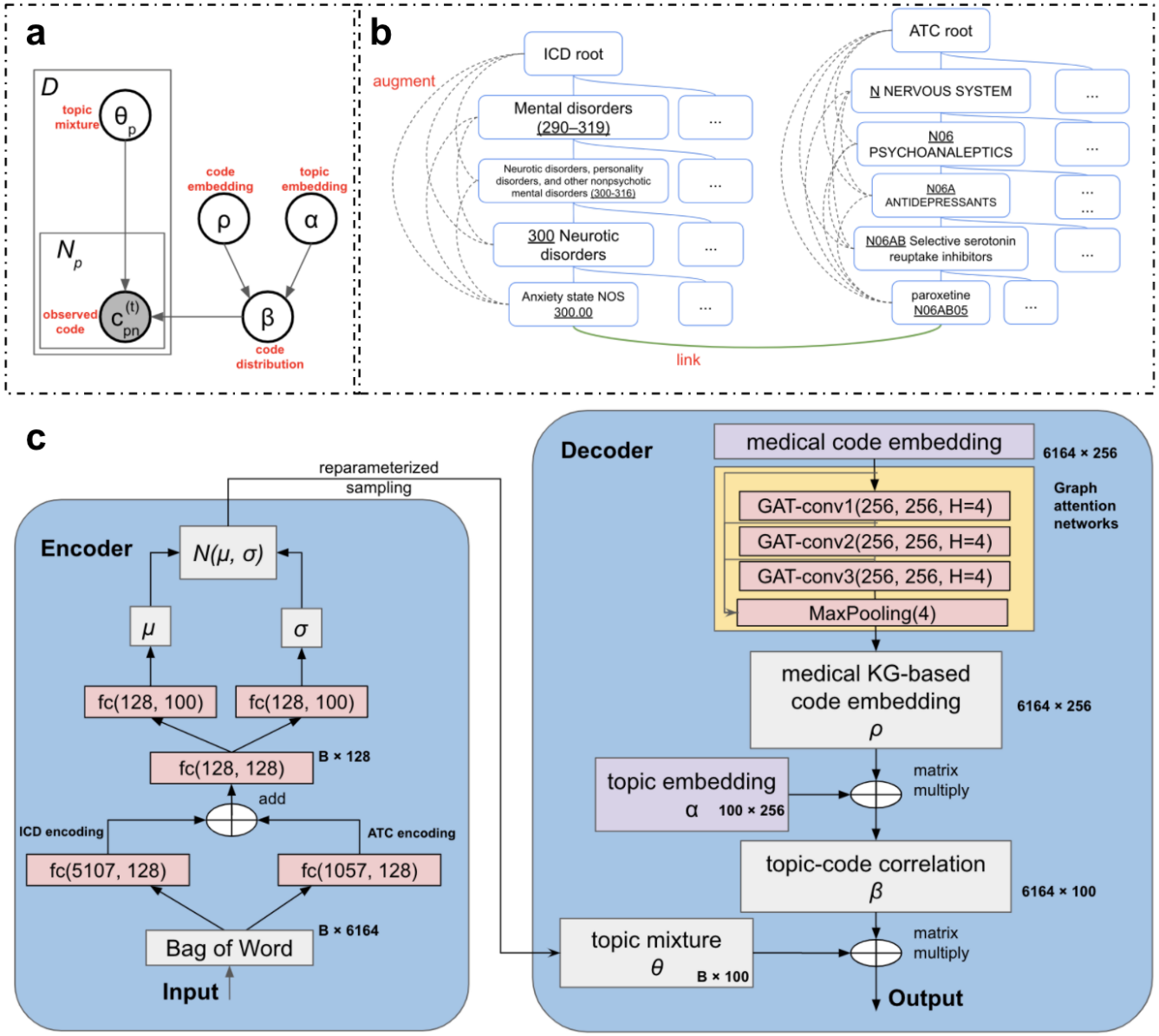




$$\alpha_{ij} = \frac{\exp(\mathbf{a}^\top[\mathbf{W}\mathbf{h}_i, \mathbf{W}\mathbf{h}_j])}{\sum_{k\in\mathcal{N}_i}\exp(\mathbf{a}^\top[\mathbf{W}\mathbf{h}_i, \mathbf{W}\mathbf{h}_k]}, \quad \mathbf{h}_i = \sigma(\sum_{j\in\mathcal{N}_i}\alpha_{ij}\mathbf{W}\mathbf{h}_j)$$

$$h_i = \|_{k=1}^{K}\sigma(\sum_{j\in\mathcal{N}}\alpha_{ij}^k\mathbf{W}^k\mathbf{h}_j), \quad h_i^{(f)} = \sigma(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}_i}\alpha_{ij}^k\mathbf{W}^k\mathbf{h}_j)$$

End-to-end training of GAT-ETM[3]



[3]Zou Y., . . . , & **Li, Y.** (2022) Modeling EHR data using GAT-ETM. arXiv.

GAT-ETM evaluation on Montreal PopHR data[4]

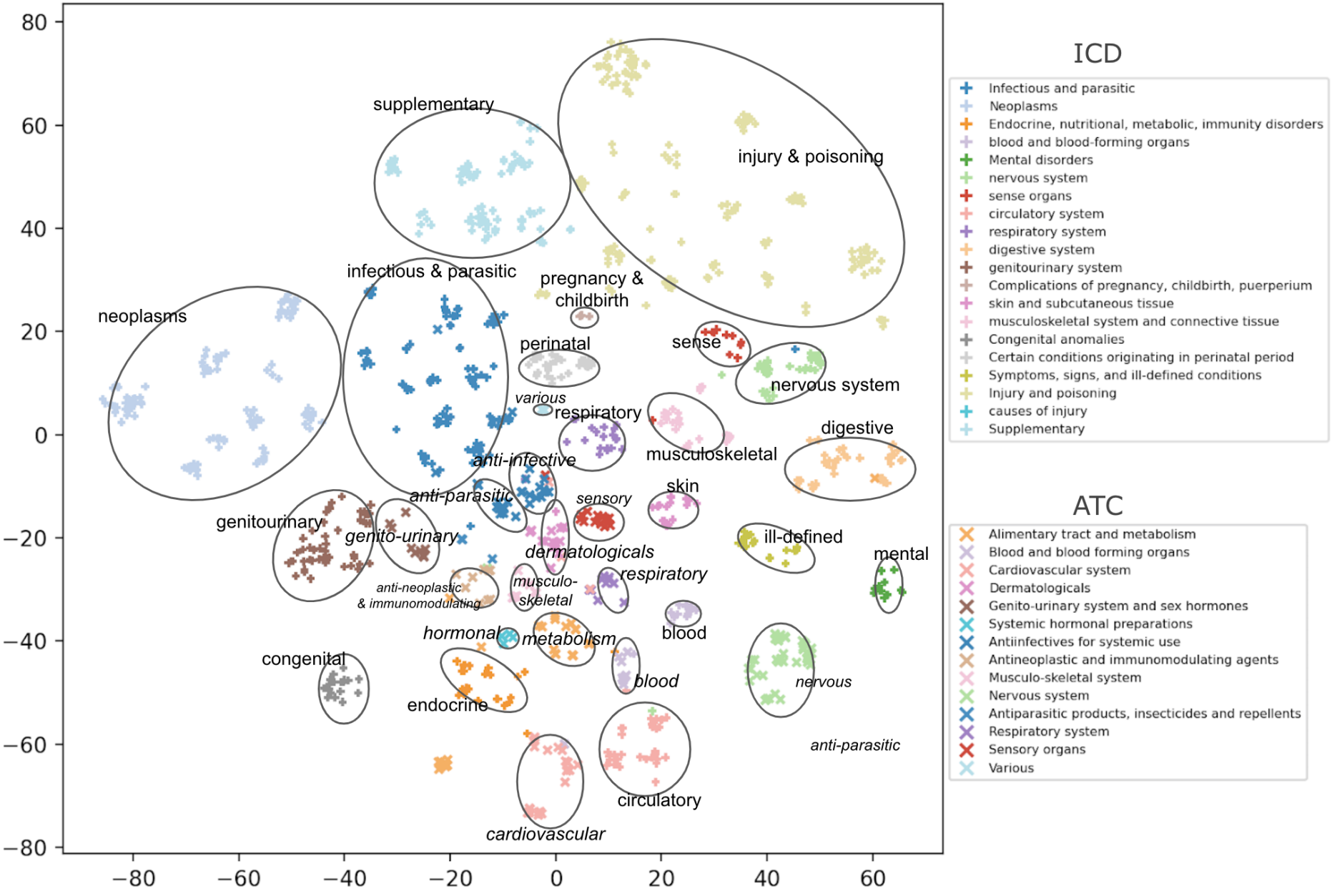
Montreal PopHR Dataset [Shaban-Nejad et al., 2017]:

- 5107 unique ICD codes
- 1057 unique ATC (i.e., medication) codes
- 1.2 million patients (6/2/2 training/validation/testing)

Model	Recon.	Topic Quality [ICD,ATC]			
	NLL.	topic coherence	topic diversity	topic quality	TQ(ave.)
ETM	198.26	0.113, 0.233	0.373, 0.423	0.0421, 0.0986	0.0704
GETM	184.32	0.167, 0.271	**0.86**, **0.83**	**0.1436**, 0.2249	0.1843
GAT-ETM	**172.69**	**0.18**, **0.314**	0.76, 0.787	0.1368, **0.2471**	**0.1920**

[4] PopHR data accessed via collaboration with David Buckeridge from School of Public Health at McGill

Embedding of EHR codes generated by the GAT



Outline

Phenome-wide association studies using EHR data

Graph-informed EHR topic modeling
- GETM: Graph-embedded topic model
- GAT-ETM: an end-to-end graph-topic model

Phecode-guided EHR topic modeling
- MixEHR-guided: a phecode-guided multimodal topic model
- MixEHR-seed: a seed-guided VAE-EM hybrid topic model

Key Idea: based on patients' ICD-9 codes from a particular dataset, infer 1500 Phecode-guided topics

Phecode Map 1.1 with ICD-9 Codes

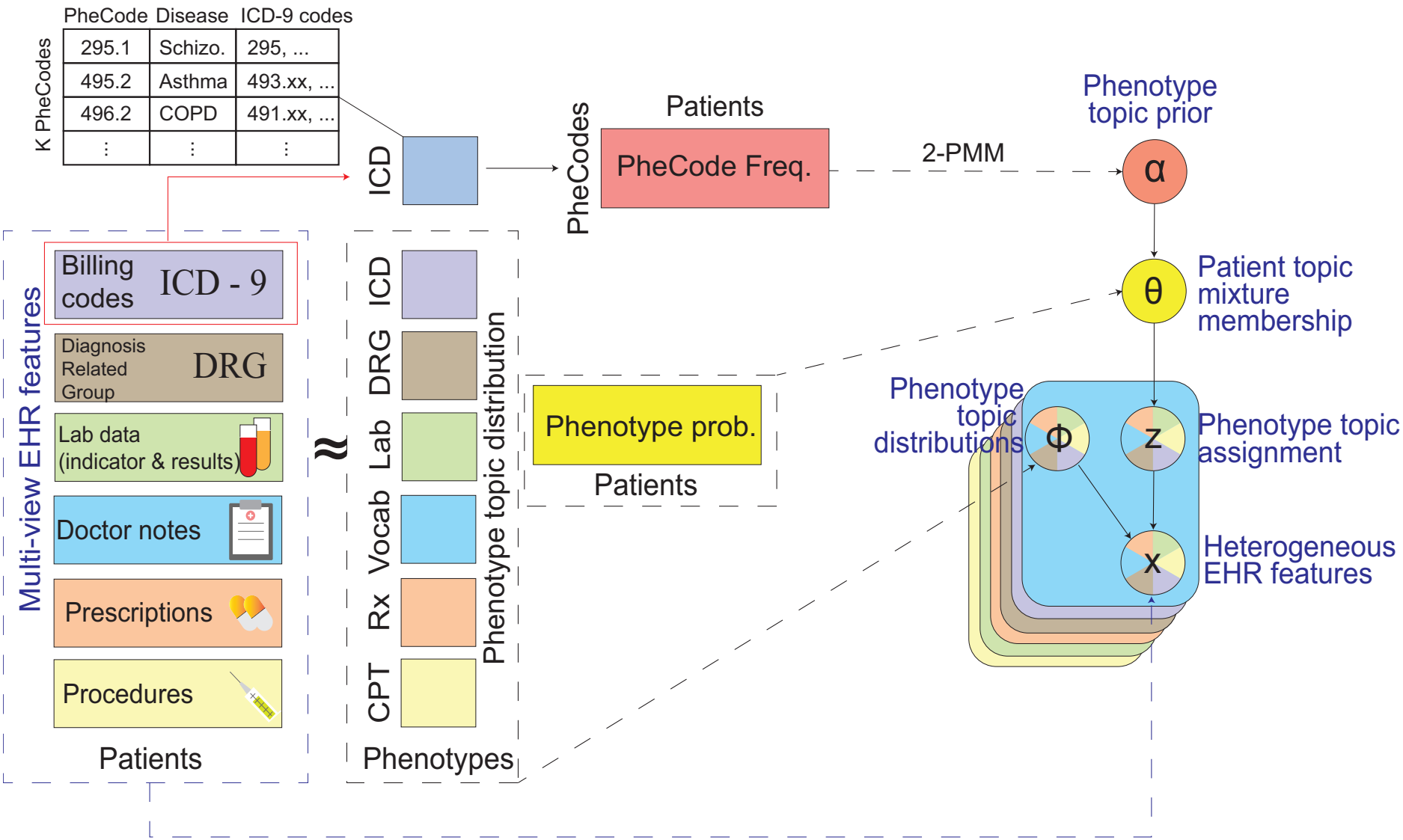
This is the previous version of the map used in the HLA analysis. You can download this with the Export All button.

Clear Filters | Export All | Export Visible

ICD9	ICD9 String	PheCode	Phenotype	Excl. Phecodes	Excl. Phenotypes
icd9	description	code	phenotype	excl. phecode	excl. range name
001	Cholera	008	Intestinal infection	001-009.99	Intestinal infection
001.0	Cholera due to Vibrio cholerae	008	Intestinal infection	001-009.99	Intestinal infection
001.1	Cholera due to Vibrio cholerae el tor	008	Intestinal infection	001-009.99	Intestinal infection
001.9	Cholera NOS	008	Intestinal infection	001-009.99	Intestinal infection
002	Typhoid and paratyphoid fevers	008	Intestinal infection	001-009.99	Intestinal infection
002.0	Typhoid fever	008.5	Bacterial enteritis	001-009.99	Intestinal infection
002.1	Paratyphoid fever A	008	Intestinal infection	001-009.99	Intestinal infection
002.2	Paratyphoid fever B	008	Intestinal infection	001-009.99	Intestinal infection
002.3	Paratyphoid fever C	008	Intestinal infection	001-009.99	Intestinal infection
002.9	Paratyphoid fever NOS	008	Intestinal infection	001-009.99	Intestinal infection
003	Other salmonella infections	008.5	Bacterial enteritis	001-009.99	Intestinal infection
003.0	Salmonella gastroenteritis	008.5	Bacterial enteritis	001-009.99	Intestinal infection
003.1	Salmonella septicemia	038.1	Gram negative septicemia	010-041.99	bacterial infection
003.2	Localized salmonella infections	008.5	Bacterial enteritis	001-009.99	Intestinal infection
003.20	Localized salmonella infection, unspeci…	008.5	Bacterial enteritis	001-009.99	Intestinal infection
003.21	Salmonella meningitis	320	Meningitis	320-326.9	INFLAMMATORY DISEASES OF THE C…
003.22	Salmonella pneumonia	480.1	Bacterial pneumonia	480-488.99	Pneumonia and influenza
003.23	Salmonella arthritis	711	Arthropathy associated with infections	710-716.99	Arthropathies
003.24	Salmonella osteomyelitis	710.1	Osteomyelitis	710-716.99	Arthropathies
003.29	Other localized salmonella infections	008.5	Bacterial enteritis	001-009.99	Intestinal infection
003.8	Other specified salmonella infections	008.5	Bacterial enteritis	001-009.99	Intestinal infection
003.9	Salmonella infection NOS	008.5	Bacterial enteritis	001-009.99	Intestinal infection
004	Shigellosis	008.5	Bacterial enteritis	001-009.99	Intestinal infection
004.0	Shigella dysenteriae	008.5	Bacterial enteritis	001-009.99	Intestinal infection

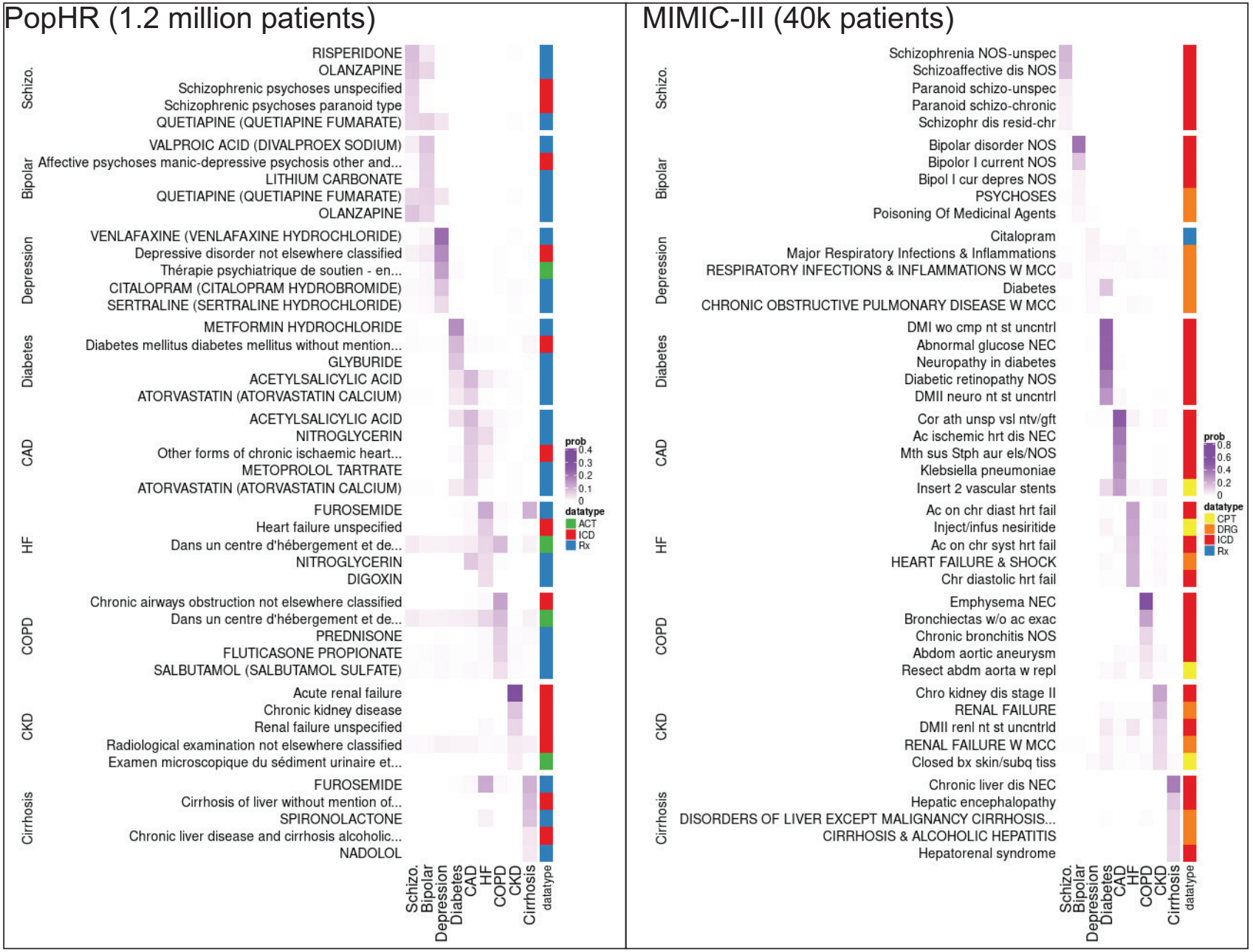
1 / 623 25 items per page 1 - 25 of 15558 items

MixEHR-Guided[5]

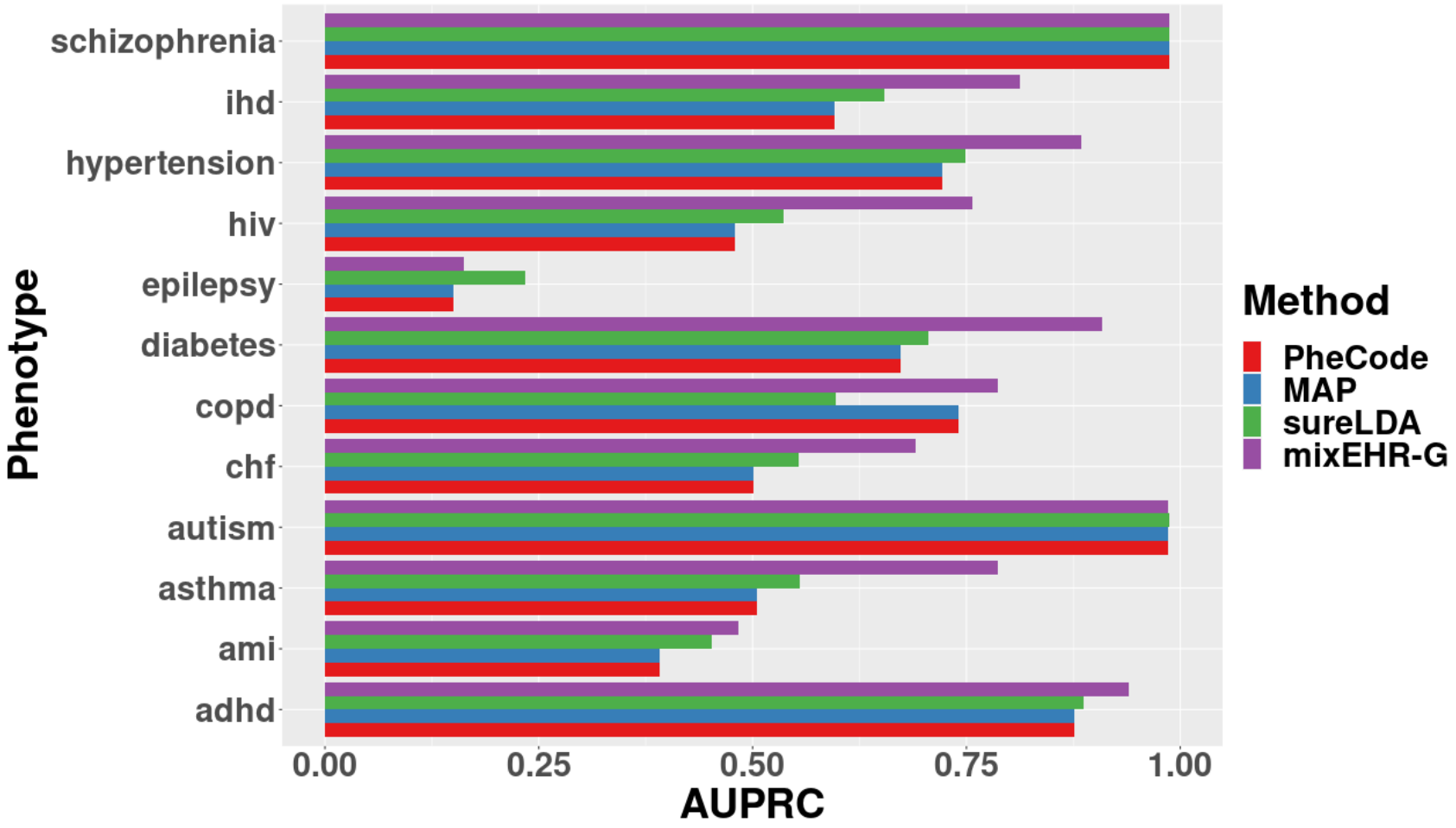


[5]Anjuha, Y., . . . , **Li, Y.***. MixEHR-Guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using EHR. (in rev.)

Top 5 features for each of 9 diverse disease phenotypes



Automatic phenotyping performance



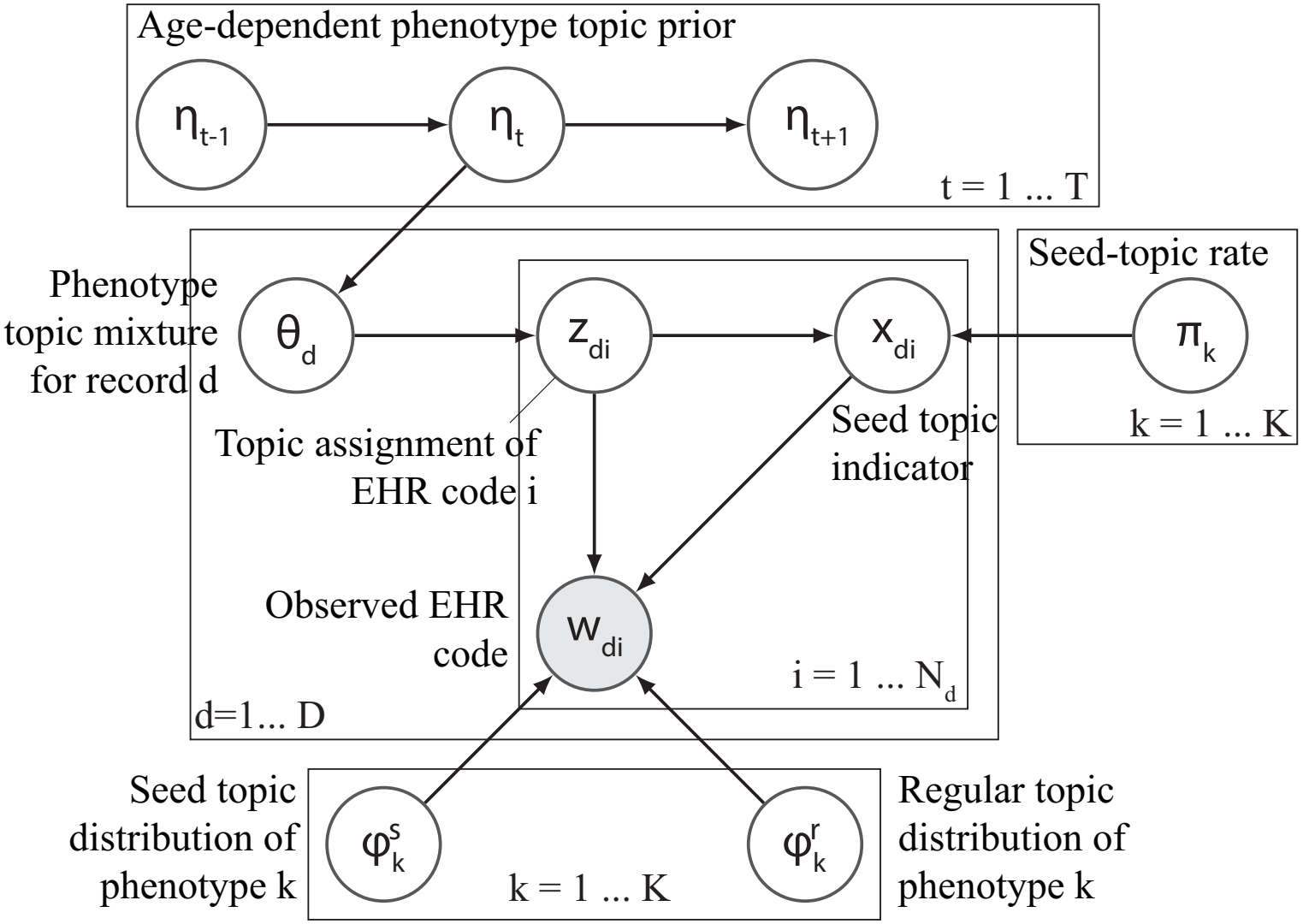
Outline

Phenome-wide association studies using EHR data

Graph-informed EHR topic modeling
- GETM: Graph-embedded topic model
- GAT-ETM: an end-to-end graph-topic model

Phecode-guided EHR topic modeling
- MixEHR-guided: a phecode-guided multimodal topic model
- MixEHR-seed: a seed-guided VAE-EM hybrid topic model

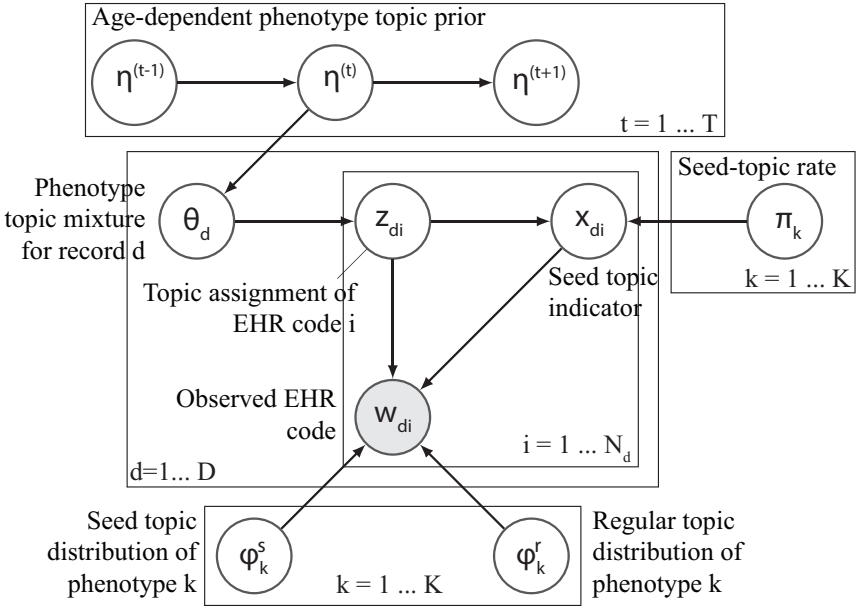
MixEHR-seed model[6]



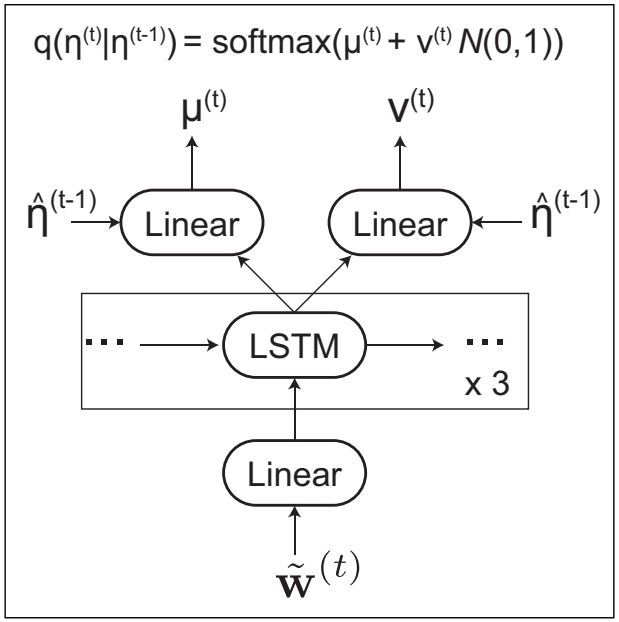
[6]Song, Z., . . . , & **Li, Y.** (2022) Automatic phenotyping by a seed-guided topic model. In Proceedings of the 28th ACM SIGKDD Conference

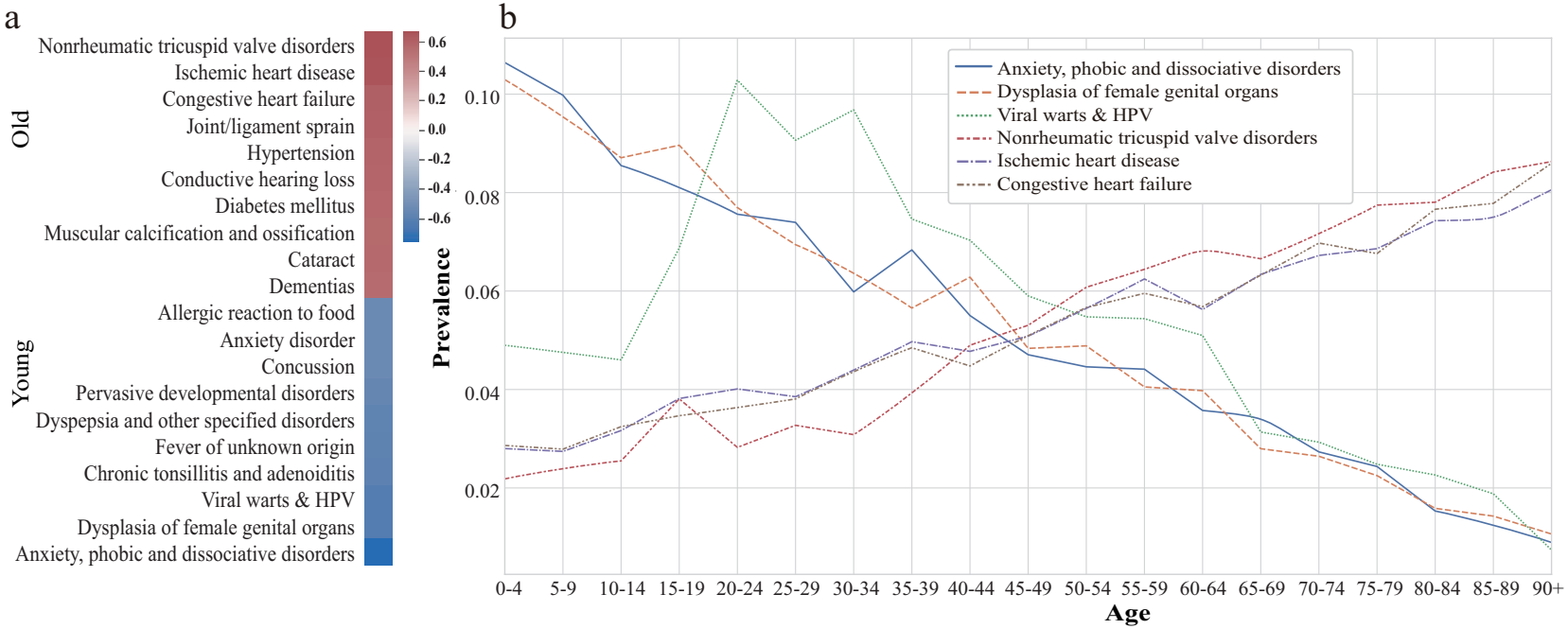
Inferred age-dependent phenotypes in 1/4 Montreal PopHR

a. MixEHR-seed PGM



b. Amortized variational inference of topic prior





Application on UK Biobank data (unpublished & prelim.)

ICD-10 processing:

- 500,000 UKB subjects (including all races)
- 6807 unique ICD-10 codes are mapped 1484 PheCodes
- Remove PheCodes with frequency < 10 subjects
- 6.12 million observed ICD-10 records

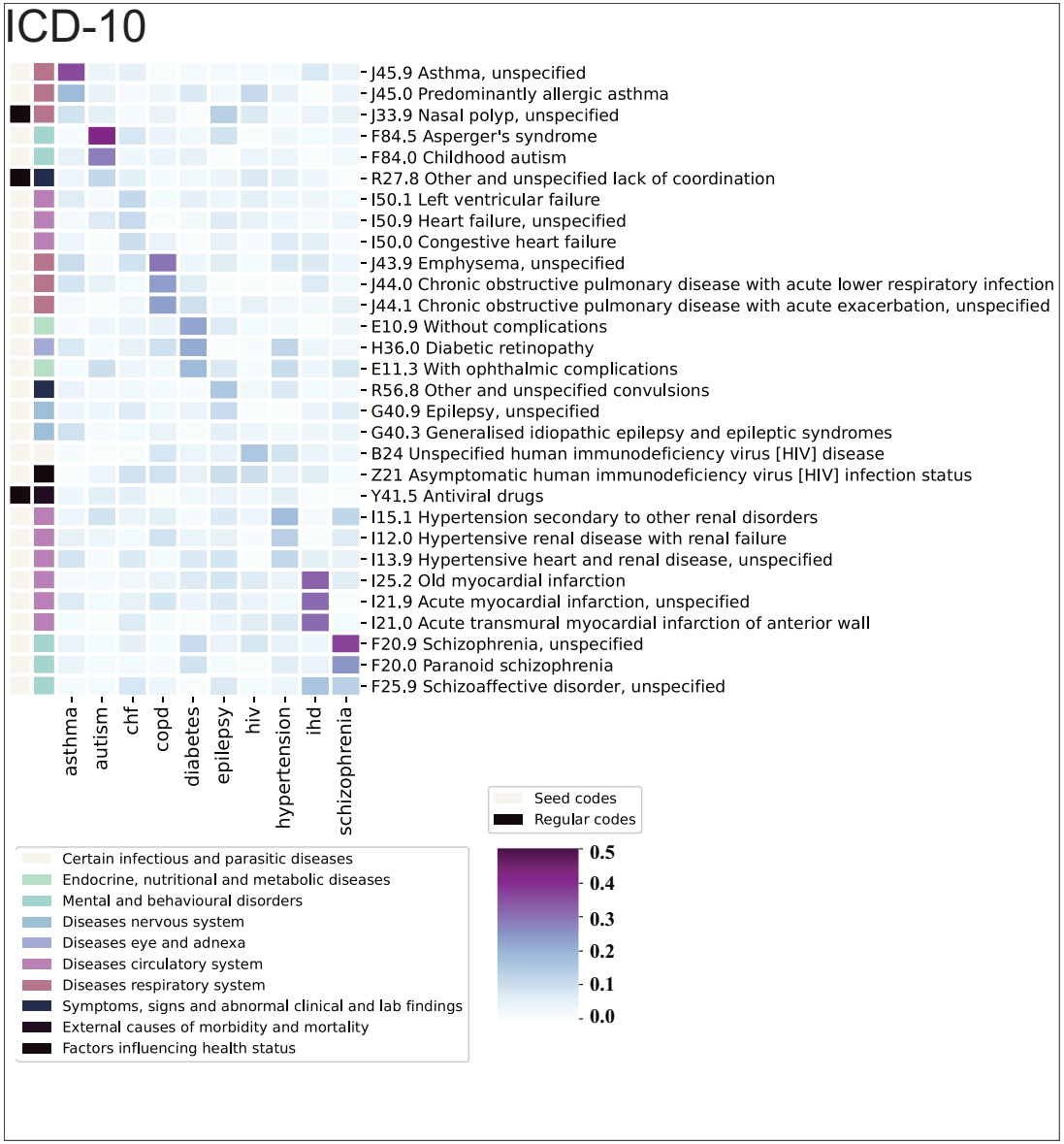
Drug code processing:

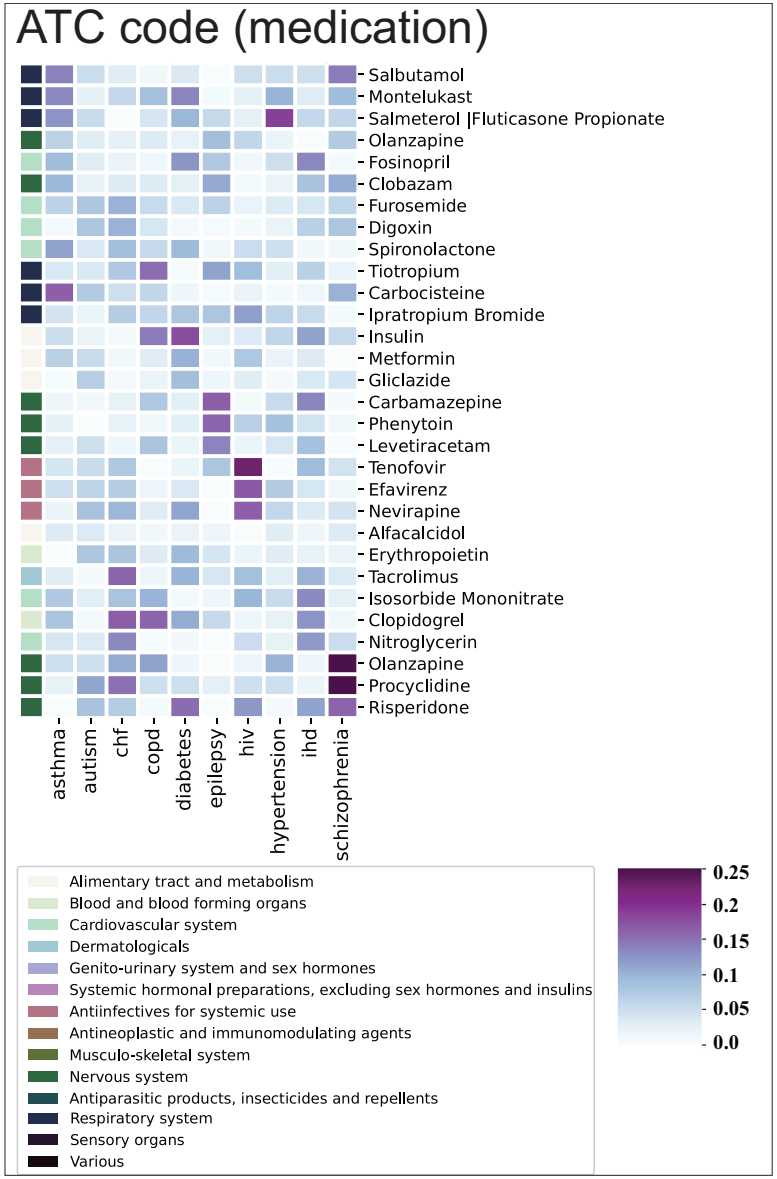
- Group all same drugs with different dosage, tablet/liquid to a unique ATC codes
- Remove ATC with frequency < 10
- 803 unique ATC codes
- 1.19 million ATC records

Drug usage prediction:

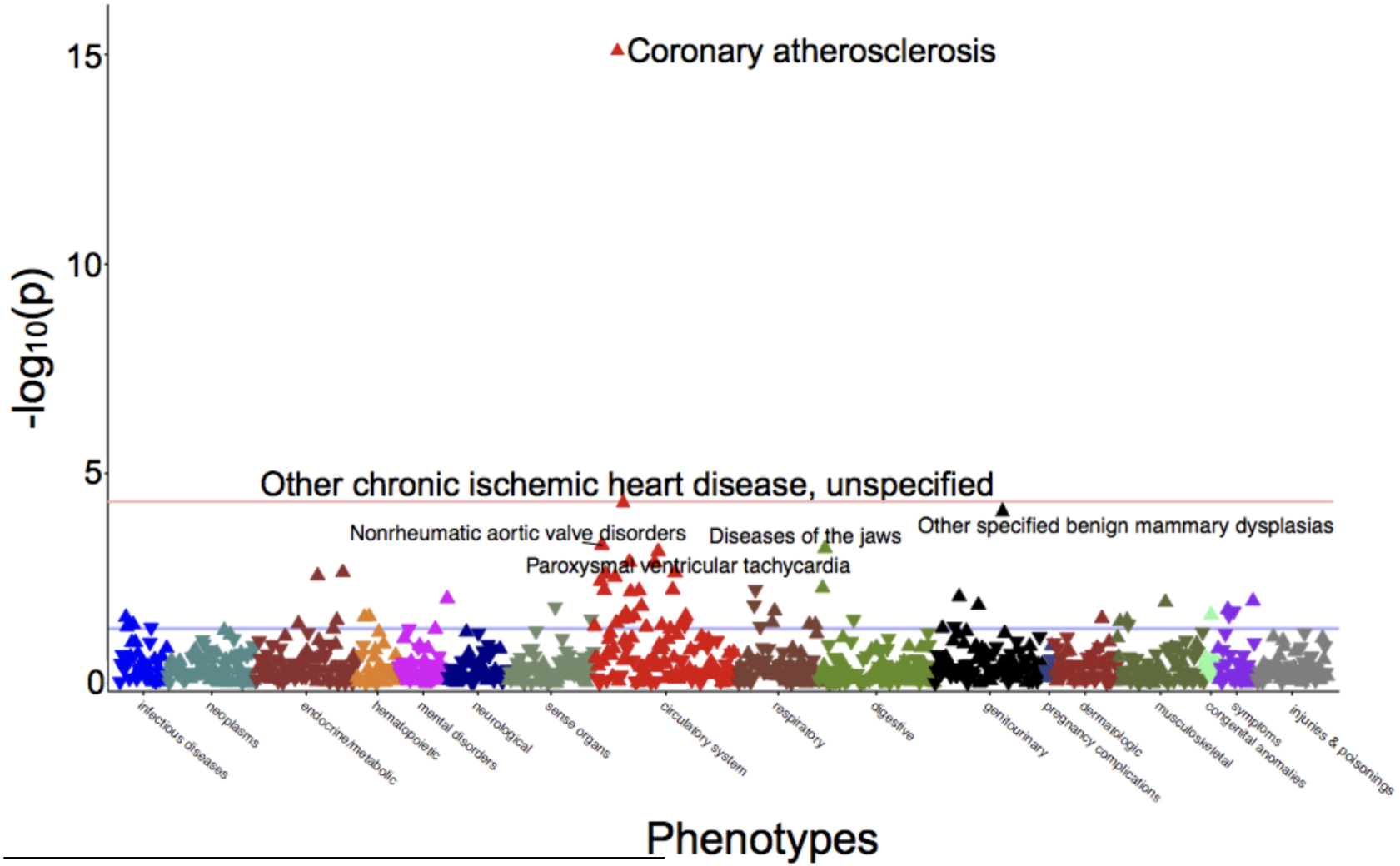
- 139 PheCodes have at least one known drug treatment
- Remove patients that use any of those drugs in the first visit
- For patients in the following visits, they were labelled as positive if they took the phecode-linked drugs
- Average AUPRC: 60% (in contrast to 40% using 2-PMM or 20% using only PheCode)

Select phecode-guided topics inferred from the UKB data



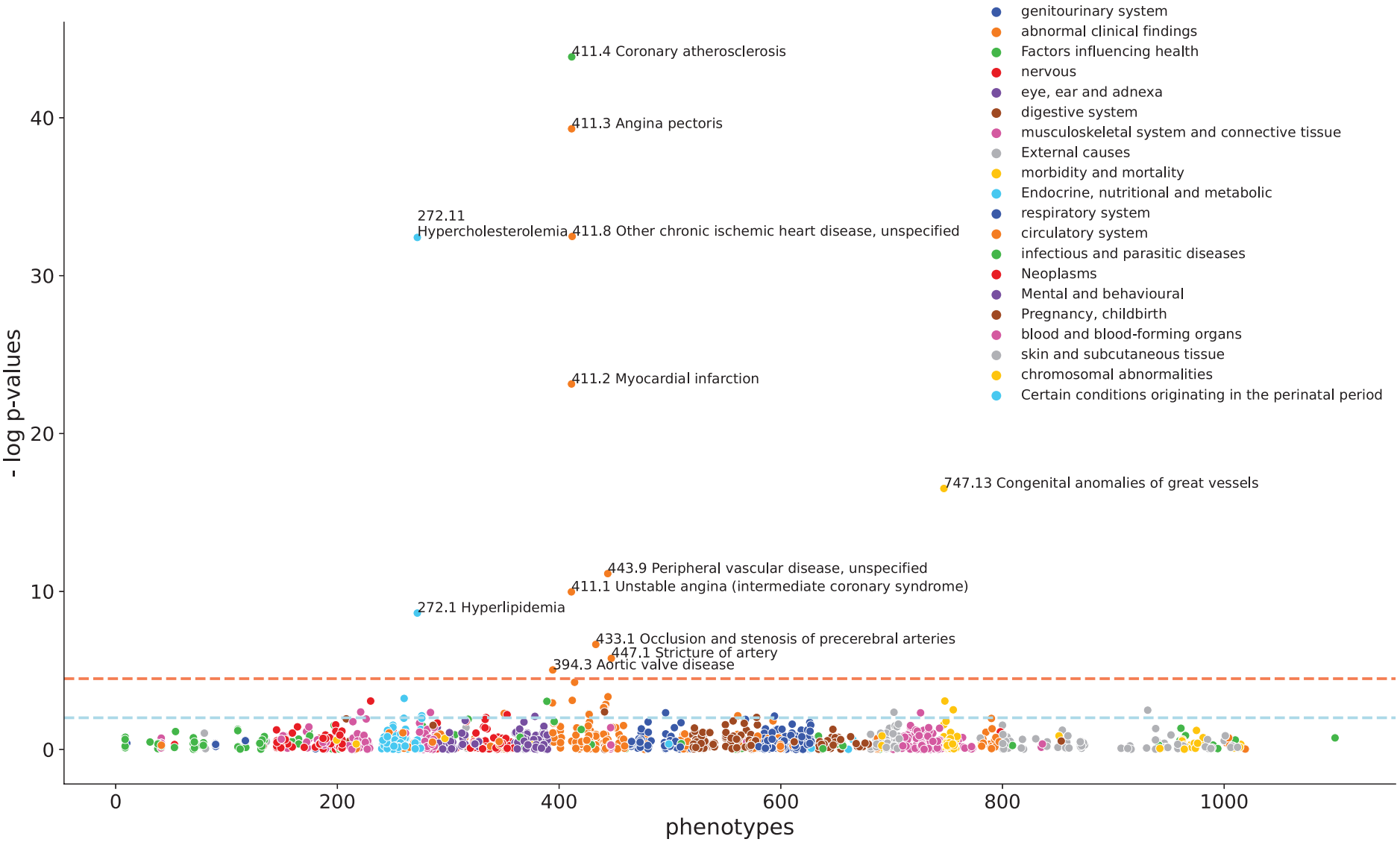


PheWAS of lipoprotein(a) (LPA) genetic variant rs10455872 using non-UKB data[7]

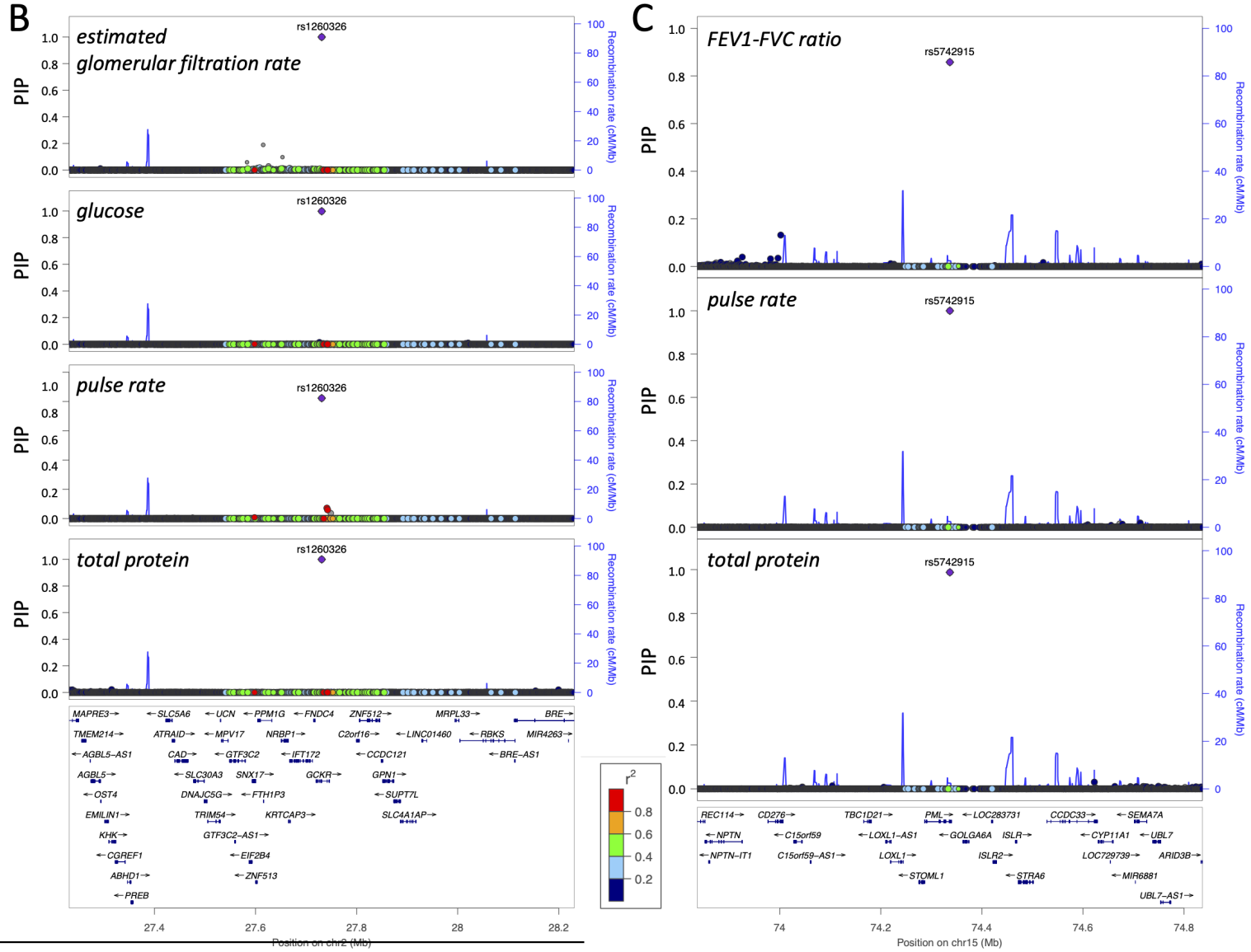


[7] 13,900 adults from DNA biobank at Vanderbilt University Medical Center; Wu, P. et al. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. Jmir Medical Informatics 7, e14325 (2019). [Wu et al., 2019]

Phenome-topic-wide associations with LPA variant rs10455872 from UK Biobank

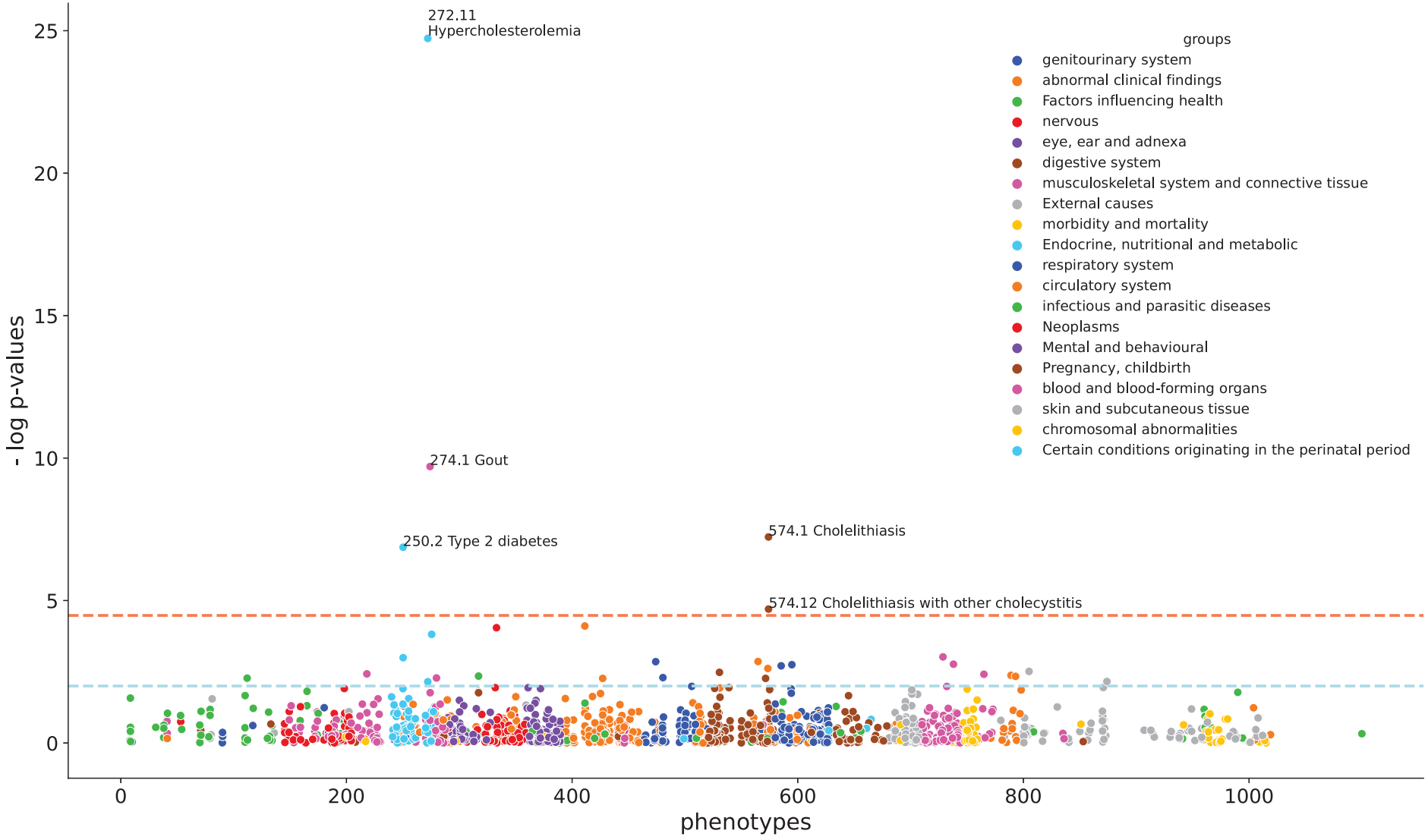


Fine-mapped SNPs for metabolic measurements[8]



[8] Zhang, W., Najafabadi, H., **Li, Y.*** SparsePro: an efficient genome-wide fine-mapping method integrating summary statistics and functional annotations. bioRxiv (under review)

PhenoTopicWAS on the fine-mapped SNP rs1260326



Summary

- Modelling multi-modal EHR data allows us to better quantify phenotypic risk as the topic probabilistic score
- Harnessing knowledge graph in representational learning help deriving interpretable topics from otherwise sparse and noisy EHR data
- Anchoring 1500 phecode-defined phenotypes enables inferring identifiable and interpretable phenotypic topics that can be used for downstream PheWAS

Acknowledgements

Li Lab (since 2019):

- GETM: Yuening Wang (M.Sc.)
- GAT-ETM: Yuesong Zou (M.Sc.)
- MixEHR-guide: Yuri Ahuja (PhD, Harvard Medical School)
- MixEHR-seed:
  - Ziyang Song (PhD cand.)
  - Yuanyi Hu (undergrad)
- UK Biobank data analysis:
  - Ziqi Yang (undergrad)
  - Wenmin Zhang (PhD cand.)

Collaborators

- Dr. Audrey Grant
- Dr. David Buckeridge
- Dr. Xiangfei Meng

Funding

- NSERC Discovery Grant (RGPIN-2019-06216)
- FRQNT New Career (NC-268592)
- Canada First Research Excellence Fund Healthy Brains for Healthy Life (HBHL) Initiative New Investigator start-up award (G24959)
- McGill initiative of Computational Medicine (MiCM)

We are recruiting! Welcome to apply: thesis-based master, PhD, postdoc
Homepage: https://www.cs.mcgill.ca/~yueli/

References I

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993–1022, March 2003.

William S Bush, Matthew T Oetjens, and Dana C Crawford. Unravelling the human genome– phenome relationship using phenome-wide association studies. Nature reviews Genetics, 17(3):129–145, February 2016.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. Nature, 562(7726):203–209, October 2018.

Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. arXiv.org, October 2017.

J C Denny, L Bastarache, M D Ritchie, and R J Carroll. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nature, 31(12):1102–1111, 2013.

References II

Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. Nature reviews Genetics, 13(6):395–405, May 2012.

Yue Li, Pratheeksha Nair, Xing Han Lu, Zhi Wen, Yuening Wang, Amir Ardalan Kalantari Dehaghi, Yan Miao, Weiqi Liu, Tamas Ordog, Joanna M Biernacka, Euijung Ryu, Janet E Olson, Mark A Frye, Aihua Liu, Liming Guo, Ariane Marelli, Yuri Ahuja, Jose Davila-Velderrain, and Manolis Kellis. Inferring multimodal latent topics from electronic health records. Nature Communications, 11(1):1–17, May 2020.

Arash Shaban-Nejad, Maxime Lavigne, Anya Okhmatovskaia, and David L Buckeridge. Pophr: a knowledge-based platform to support integration, analysis, and visualization of population health data. Annals of the New York Academy of Sciences, 1387(1):44–53, 2017.

Patrick Wu, Aliya Gifford, Xiangrui Meng, Xue Li, Harry Campbell, Tim Varley, Juan Zhao, Robert Carroll, Lisa Bastarache, Joshua C Denny, Evropi Theodoratou, and Wei-Qi Wei. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. JMIR Medical Informatics, 7(4):e14325, 2019. ISSN 2291-9694. doi: 10.2196/14325.