UNC | GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

PHAIR

Electronic Health Records

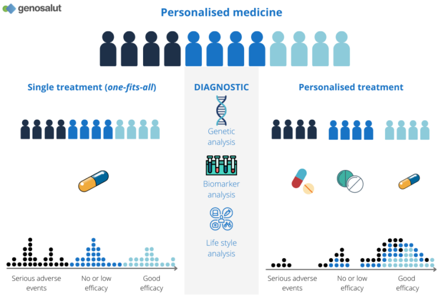
Yajie He, Yuanyuan Yan, Yang Ruan

9/6/2024

Content

An Motivating Example[1]

- Precisione Medicine offers more accurate diagnosis and personalized treatment
- This requires integrating **large-scale clinical** and **omics data**
- Highlights the need for large dataset and techniques including **data integration**, **genotyping**, **phenotyping**, and **ML/DL**



Why Electronic Health Records (EHR)?

- Provides detailed, longitudinal clinical data
- Not available in administrative or survey datasets
- Crucial for realizing precision medicine

[1] Ahmed Elhussein et al. "A framework for sharing of clinical and genetic data for precision medicine applications". In: *Nature Medicine* (2024), pp. 1–12.

An Motivating Example

Genotype

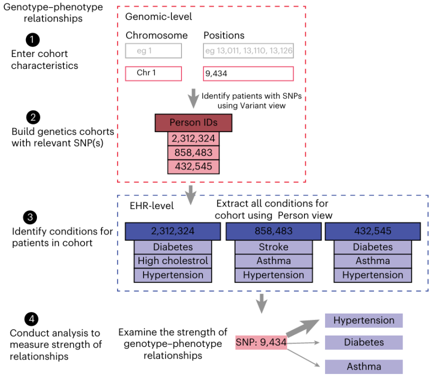
- combination of alleles (different forms of a gene) people inherits from parents
- Example: "Bb", brown eyes "B"and blue eyes "b", unobservable

Phenotype

- influenced by both its genotype and environmental factors
- Example: Brown eyes, observable

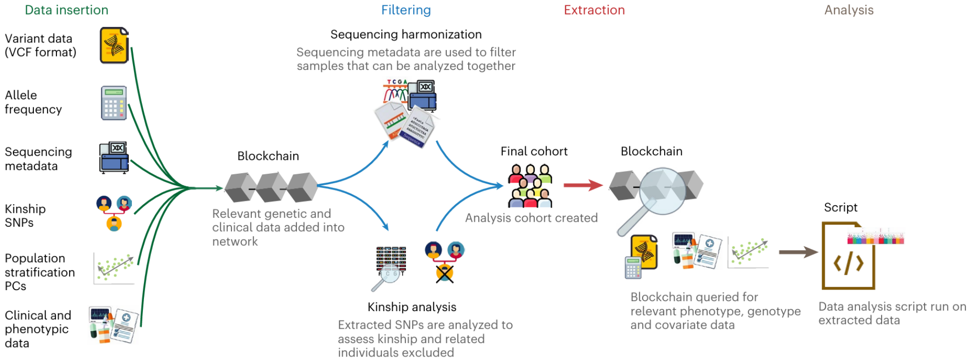
Single Nucleotide polymorphism (SNP)

- are variations in a single nucleotide
- can be associated with particular traits or diseases.



Why genotype–phenotype relationships: understand how specific genetic variants lead to particular health outcomes - whether increase disease risk or affect treatment responses $\Rightarrow$ tailor treatments based on personalized genetic profile

An Motivating Example



Data Insertion ⇒ Filtering ⇒ Extraction ⇒ Analysis (e.g. Genotype-phenotype association studies)

Ensures that the analysis is conducted on high-quality, well-filtered data

Increases the reliability and precision of the results

Content

EHR - Definition

What is Electronic Health Records (EHR) ?

- An EHR is an electronic version of a patients medical history, that is maintained by the provider over time (Centers for Medicare & Medicaid Services)

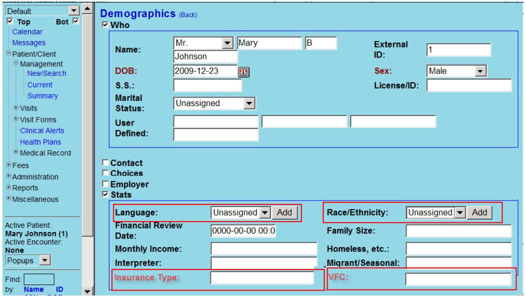
EHR contains patient health information, such as:

- Patient demographics
- Progress notes
- Vital signs
- Medical histories
- Diagnoses
- Medications
- Immunization dates
- Allergies
- Radiology images
- Lab and test results

Data Elements[2]

Demographic

- MRN: medical record number
- Basic information (static): Age, Sex, Race/Ethnicity
- Time varying elements: Payer, address



Contextualizing Information for Encounters

- **When** someone seen
- **Who** the patient saw
- **Where** the patient was seen
- **What** happened

Encounter Type

Three Basic Encounters:

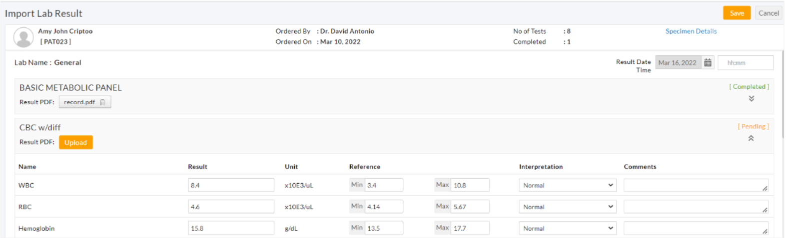
- Outpatient (AV- Ambulatory Visit)
- Inpatient (IP)
- Emergency Department (ED)

[2] Salomeh Keyhani et al. "Electronic health record components and the quality of care". In: *Medical care* 46.12 (2008), pp. 1267–1272.

Data Elements

Laboratory Tests Results

Outcomes of clinical lab tests: blood, urine, or tissue analysis, etc.



Vitals Signs

Key physiological indicators: blood pressure, heart rate, temperature, and respiratory rate, etc.

Social Health

Data related to a patient's social determinants of health: smoking status, drug and alcohol use, employment status, marital status, etc.

Data Elements[3]

International Classification of Diseases (ICD) Codes

- What is ICD codes?
  Hierarchical system to code all diagnoses that are made during a health encounter

- ICD-9 vs ICD-10
  ICD-9(previously used) had approximately 13,000 unique codes, ICD-10 has approximately 68,000

- Codes don't always represent the primary concern, they are used as billing codes.

[3] Keyhani et al., "Electronic health record components and the quality of care".

Data Elements

The Structure of ICD Codes .

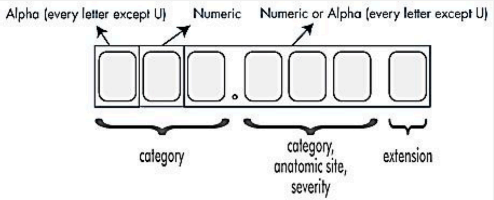


Figure: Structure of ICD-10 Codes

- **Category**: The first three characters represent the diagnostic category.
- **Etiology, Anatomic Site, and Severity**: The next characters, positions 4-6, provide further details regarding the cause, anatomic site, and severity.
- **Extension**: A 7th character may be added for additional specificity, such as for initial or subsequent encounters, or complications from conditions.

Data Elements

Example of ICD-10

For example, the code ‘S52.521A’ represents:

- **S52** - Fracture of forearm
- **S52.52** - Fracture of lower end of radius
- **S52.521** - Fracture of lower end of right radius
- **S52.521A** - Initial encounter for closed fracture of right radius

Data Elements

Current Procedural Terminology (CPT) Codes

Category/Subcategory	Code Numbers
Team conference services	
Team conferences with patient/family	99366*
Team conferences without patient/family	99367
Behavior Change Interventions	
Smoking and tobacco use cessation	99406-99407
Alcohol and/or Substance abuse structured screening and brief intervention	99408-99409
Non-Face-to-Face Physician Services*	
Telephone services	99441-99443
On-Line Medical Evaluation	99444
Basic Life and/or Disability Evaluation Services	99450
Work Related or Medical Disability Evaluation Services	99455-99456

Figure: CPT Codes

- CPT is coding system for what happened during an encounter, e.g., surgeries, x-rays, etc.
- Approximately 10,000 unique codes in use
- Also tied to reimbursements and similar systems for organizing CPTs as ICDs

Data Elements

Medications[4]

- Medications element refers to the documentation and management of a patient's medication history.

- In EHR systems, medications are organized into three main categories:

  Prescribed Medications: The drugs ordered by healthcare providers for patient treatment.
  Administered Medications: Records of medications that have been given to patients during care (e.g., in hospital).
  Reconciliation: A process that ensures consistency in medication records, particularly during transitions of care, such as hospital admission or discharge.

[4] Alice S Tang et al. "Harnessing EHR data for health research". In: *Nature Medicine* (2024), pp. 1–9.

Content

What We Can Do with EHR

Risk Prediction

- Disease Management and Clinical Decision Support[5]:
  - Predict the likelihood of developing chronic diseases
  - Provide clinical support, risk alerts and treatment recommendations
- Adverse Drug Reactions[6]:
  - Identify patients at high risk for adverse drug reactions through machine learning models

Population Health[7]

- Public Health Surveillance:
  - Monitor disease rates & track trends & early detection of health threats
- Public Health Interventions:
  - Design targeted health interventions

---

[5] Che Ngufor et al. "Development and validation of a risk stratification model using disease severity hierarchy for mortality or major cardiovascular event". In: *JAMA network open* 3.7 (2020), e208270–e208270.

[6] Hae Reong Kim et al. "Analyzing adverse drug reaction using statistical and machine learning methods: A systematic review". In: *Medicine* 101.25 (2022), e29387.

[7] Daniel J Friedman, R Gibson Parrish, and David A Ross. "Electronic health records and US public health: current realities and future promise". In: *American journal of public health* 103.9 (2013), pp. 1560–1567.

What We Can Do with EHR

Precision Health[8]

- Personalized Treatment Plans:
  - Integrate genomics data into EHR
  - Enhance effectiveness of medical interventions

Genomic Medicine[9]

- Genomic Research and Discovery:
  - Leveraging EHR-linked biobanks to conduct Genome-Wide Association Studies (GWAS)
  - Facilitate discovery of genetic variants associated with diseases

Clinical Implementation[10]

- Pharmacogenomics in Clinical Care:
  - Guide drug prescribing in clinical trials

---

[8] Noura S Abul-Husn and Eimear E Kenny. "Personalized medicine and the power of electronic health records". In: *Cell* 177.1 (2019), pp. 58–69.

[9] Jodell E Linder et al. "The role of electronic health records in advancing genomic medicine". In: *Annual review of genomics and human genetics* 22.1 (2021), pp. 219–238.

[10] J Kevin Hicks et al. "Integrating pharmacogenomics into electronic health records with clinical decision support". In: *American journal of health-system pharmacy* 73.23 (2016), pp. 1967–1976.

Content

The Advantages of Using EHR Data for Clinical Research

- Data Readily Available: the use of data that are already captured, thus reducing administrative efforts, costs.
- Large Sample Sizes: EHRs contain vast amounts of patient data, enabling researchers to study rare diseases, uncommon side effects, and conduct subgroup analyses with greater statistical power.
- Comprehensive and Multidimensional Data: EHRs contain information collected over a variety of fields and include more comprehensive data from patients as they evolve over time.

Content

Challenges: Sample Selection Bias[11]

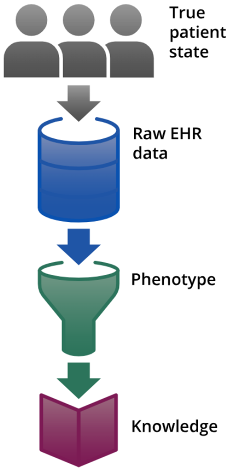
Step: Cohort Building

Patient Classification:

- Require careful phenotyping to identify cases and control groups
- Introduce bias if not rigorous or consistent phenotype
- May lead to misclassification or exclusion of certain patients

How to avoid:

- Use cohort definitions that have been extensively validated
- Perform sensitivity analyses to check the robustness of your case definitions



[11] Christopher M Sauer et al. "Leveraging electronic health records for data science: common pitfalls and how to avoid them". In: *The Lancet Digital Health* 4.12 (2022), e893–e898.

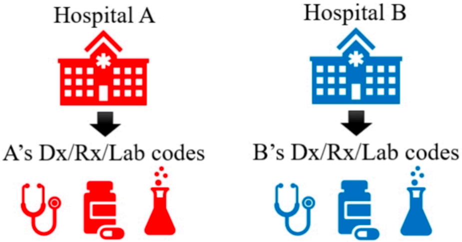
Challenges with EHR Data

Step: Model Building

Challenges III: Heterogenous EHR Formats[12]

- **What is heterogenous in EHR**
  Variation in data structures, coding standards, and information representation across different EHR systems.

- **Problem**
  Poses challenges for data integration, analysis, and interoperability, making it difficult to share and utilize patient information seamlessly across different healthcare settings.



[12] Kyunghoon Hur et al. "Unifying heterogeneous electronic health records systems via text-based code embedding". In: *Conference on Health, Inference, and Learning*. PMLR. 2022, pp. 183–203.

Challenges: Heterogenous EHR Formats

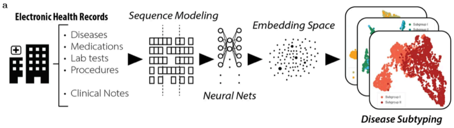
Deal with heterogenous formats

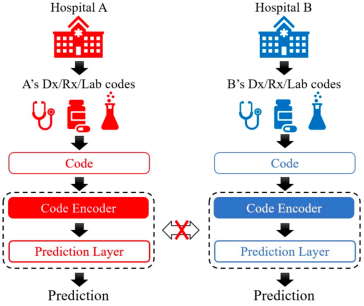
Code-Based Embedding:

- **How it works**: converting structured data, like medical codes (ICD codes, procedure codes, etc.), into a dense vector representation
- **Advantage**: enabling their use in machine learning models for tasks like patient outcome prediction or treatment recommendation.

Problem:

- The code encoders and the prediction layers cannot be shared among different hospitals due to heterogeneity of the code systems.



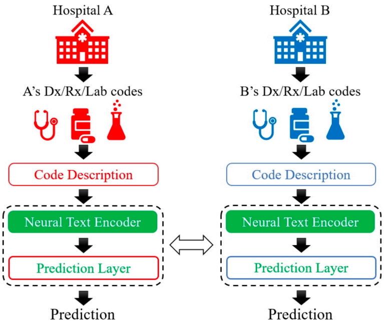


Challenges: Heterogenous EHR Formats

Deal with heterogenous formats

Text-Based Embedding:

- **How it works**: converting unstructured or semi-structured clinical text, such as clinical notes, medical code descriptions, and free-text entries, into numerical vector representations.
- **Advantage**: both the text encoders and the prediction layers can be transferred between different hospitals.



Common Data Models (CDMs):

- **How it works**: It provides a uniform framework for collecting, sharing, and analyzing healthcare data from various sources like hospitals, clinics, and research institutions.

Challenges: Overfitting and Generalisability[13]

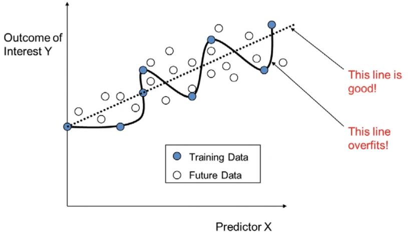
Step: Study and Results Validation

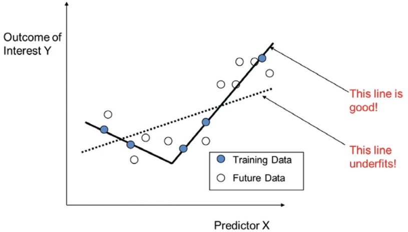
Model Overfitting:

- Overfitting can lead to reduced generalisability across institutions
- Local practice patterns might influence model performance

How to avoid:

- Use causal diagrams to infer generalisability across clinical settings
- Avoid using metrics sensitive to class imbalance (e.g., accuracy)
- Tailor performance measures to the specific use case, screening vs treatment recommendation





[13] Constantin Aliferis and Gyorgy Simon. "Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI". In: *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls.* Springer, 2024, pp. 477–524.

Content

Introductory to Frailty[14]

Frailty Measurement (General Approach)

(1) Phenotype based

Example: Fried Frailty Phenotype: Identified by the presence of three or more of the following components: Shrinking (unintentional weight loss), Weakness, Poor endurance and energy, Slowness, Low physical activity level.

(2) Frailty Index Based on Accumulation of Deficits

Example: Counting the number of health deficits (e.g., diseases, symptoms, disabilities) and expressing it as a proportion of the total number of possible deficits.

Meaning of the Measurement

Its predictive ability and potential as a therapeutic target for the prevention of falls, disability, and unnecessary or burdensome hospitalizations.

[14] Dae Hyun Kim. "Measuring frailty in health care databases for clinical care and research". In: *Annals of geriatric medicine and research* 24.2 (2020), p. 62.

Frailty Measurement in EHR[15]

Population

- Data from July 1, 2014 to July 1, 2016.
- Patients included were at least 65 years of age as of July 1, 2016.
- Data sources: Encounter, diagnosis code, laboratory, medication, and Medicare Annual Wellness Visit (AWV) data from the EHR.

[15] Nicholas M Pajewski et al. “Frailty screening using the electronic health record within a Medicare accountable care organization”. In: *The Journals of Gerontology: Series A* 74.11 (2019), pp. 1771–1777.

Frailty Measurement in EHR

Composition of the Adapted EHR-Based Frailty Index (eFI)

(1) Deficit Selection

The eFI is an index to measure frailty based on 54 deficits. These deficits are derived from:

- **Diagnosis Codes**: ICD-9-CM and ICD-10-CM codes from various healthcare sources (outpatient, inpatient, emergency department, etc.)

  If a patient has no observed diagnosis codes during the 2-year look-back period, those deficits are marked as missing (NA).
- **Laboratory Measures and Vital Signs**: Data from outpatient encounters, following the lab-based FI of Howlett and colleagues.

  If multiple measurements are available during the look-back period, each is scored individually, and the scores are averaged to obtain a single deficit score.
- **Functional Data**: Information from AWV.

Frailty Measurement in EHR

Composition of the Adapted EHR-Based Frailty Index (eFI)

(2) Calculation of eFI

- A minimum of 30 non-missing deficits is required to compute the eFI, i.e, NA<30.
- $n =$ count of deficits
- eFI $= \frac{n}{54}$

(3) eFI Classification

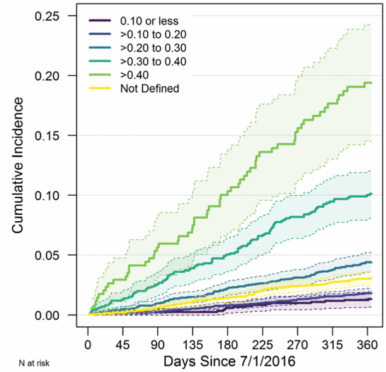
- Non-frail: eFI ≤ 0.1
- Pre-frail: 0.1 < eFI < 0.2
- Frailty: eFI ≥ 0.2

Frailty Measurement in EHR

Statistical Analysis

(1) Association with All-Cause Mortality

- Kaplan–Meier estimates of all-cause mortality stratified by eFI score



Frailty Measurement in EHR

- Multivariable Cox regression model

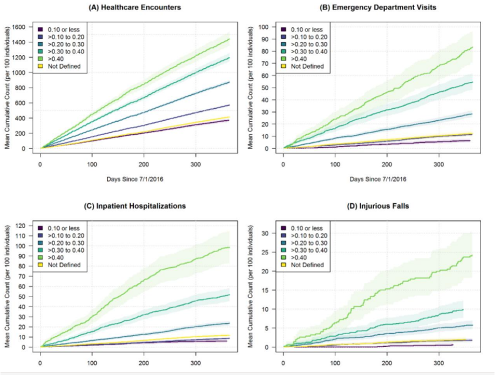
Variable	Hazard Ratio (95% CI)	p Value	Explained Relative Risk
Age (per 5 year increase)	1.23 (1.15–1.32)	<.001	4.3%
Sex (male)			1.6%
Female	0.68 (0.55–0.84)	<.001	
Race (white)			0.1%
Nonwhite	0.85 (0.64–1.14)	.285	
No. of outpatient encounters in past 2 years (<5)			1.0%
5 to <10	1.04 (0.64–1.67)	.881	
10 or more	0.76 (0.48–1.20)	.240	
No. of ED visits or inpatient encounters in past 2 years (0)			6.2%
1	1.69 (1.24–2.32)	.001	
2 or more	2.43 (1.79–3.31)	<.001	
Weighted Charlson Comorbidity Index	1.17 (1.13–1.21)	<.001	11.2%
eFI (per 0.1 increase)	1.33 (1.15–1.53)	<.001	3.5%

The c-statistic for a model based only on the eFI was 0.740 and 0.790 for the full multivariable model.

Frailty Measurement in EHR

(2) Associations with incident health care utilization and injurious falls

- Unadjusted mean cumulative count (MCC) estimates for incident health care utilization and injurious falls.



The c-statistics were 0.724 (inpatient), 0.691 (ED), and 0.749 (injurious falls) based on the eFI.

Frailty Measurement in EHR

- Adjusted estimates of the association between the eFI and utilization or falls, adjusting for age, comorbidity, and past health care utilization.

	Health Care Encounters		Emergency Department Visits	
Variable	Hazard Ratio (95% CI)	*p* Value	Hazard Ratio (95% CI)	*p* Value
No. of outpatient encounters in past 2 years (<5)				
5 to <10	1.24 (1.16–1.33)	<.001	1.17 (0.89–1.53)	.255
10 or more	1.97 (1.84–2.12)	<.001	1.42 (1.11–1.83)	.006
No. of ED visits or inpatient encounters in past 2 years (0)				
1	1.07 (1.04–1.11)	<.001	1.97 (1.66–2.33)	<.001
2 or more	1.15 (1.10–1.20)	<.001	3.75 (3.16–4.44)	<.001
Weighted Charlson Comorbidity Index	1.02 (1.01–1.02)	<.001	1.00 (0.98–1.02)	.952
eFI (per 0.1 increase)	1.20 (1.17–1.22)	<.001	1.36 (1.26–1.46)	<.001
	Inpatient Encounters		Injurious Falls	
Variable	Hazard Ratio (95% CI)	*p* Value	Hazard Ratio (95% CI)	*p* Value
No. of outpatient encounters in past 2 years (<5)				
5 to <10	1.14 (0.85–1.51)	.383	1.22 (0.60–2.46)	.583
10 or more	1.14 (0.87–1.50)	.352	1.04 (0.53–2.05)	.914
No. of ED visits or inpatient encounters in past 2 years (0)				
1	1.75 (1.48–2.08)	<.001	1.99 (1.38–2.86)	<.001
2 or more	2.32 (1.95–2.76)	<.001	3.16 (2.23–4.50)	<.001
Weighted Charlson Comorbidity Index	1.05 (1.02–1.07)	<.001	0.99 (0.95–1.04)	.749
eFI (per 0.1 increase)	1.62 (1.50–1.76)	<.001	1.66 (1.42–1.93)	<.001

The c-statistics were 0.741 (inpatient), 0.739 (ED), 0.791 (injurious falls).

Frailty Measurement in EHR

(3) Contribution of functional data from Annual Wellness Visits

Variable	Healthcare Encounters Hazard Ratio (95% CI)	Healthcare Encounters *p* Value	Emergency Department Visits Hazard Ratio (95% CI)	Emergency Department Visits *p* Value
eFI (per 0.1 increase)[a]	1.24 (1.20–1.28)	<.001	1.29 (1.13–1.46)	<.001
≥1 Deficit from Medicare AWV[b]	1.03 (0.98–1.07)	.275	1.17 (0.93–1.48)	.172
Variable	**Inpatient Encounters Hazard Ratio (95% CI)**	**Inpatient Encounters *p* Value**	**Injurious Falls Hazard Ratio (95% CI)**	**Injurious Falls *p* Value**
eFI (per 0.1 increase)[a]	1.78 (1.52–2.07)	<.001	1.45 (1.11–1.89)	.007
≥1 Deficit from Medicare AWV[b]	1.44 (1.13–1.84)	.003	1.85 (1.07–3.21)	.028

Conclusion In this study, eFI demonstrated reasonable discriminative ability (c-statistics > 0.70), so it is feasible to adapt an EHR-based FI. within the context of a Medicare ACO population

Acknowledgements

We would like to thank Dr. Smith for his invaluable guidance and support in completing this presentation.

Advisor:

- Patrick J. Smith, PhD, MPH

PHAIR Lab members:

- Yajie He, MS.
- Yuanyuan Yan, MS.
- Yang Ruan, MS.

PHAIR lab

Precision Health and Artificial Intelligence Research

Thank you!

You can contact us at: **yajie@unc.edu, yuanyyan@unc.edu, yangruan@unc.edu**