Recent Advances in Predictive Modeling with Electronic Health Records

Jiaqi Wang, Junyu Luo, Muchao Ye, Xiaochen Wang, Yuan Zhong, Aofei Chang, Guanjie Huang, Ziyi Yin, Cao Xiao, Jimeng Sun, **Fenglong Ma**

College of Information Sciences and Technology

The Pennsylvania State University

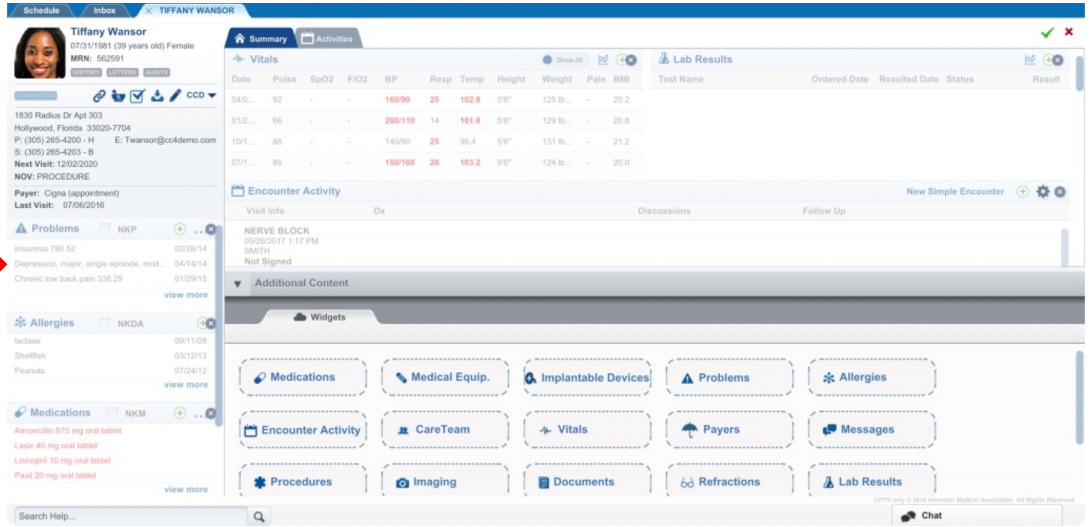
jqwang@psu.edu

08/2024

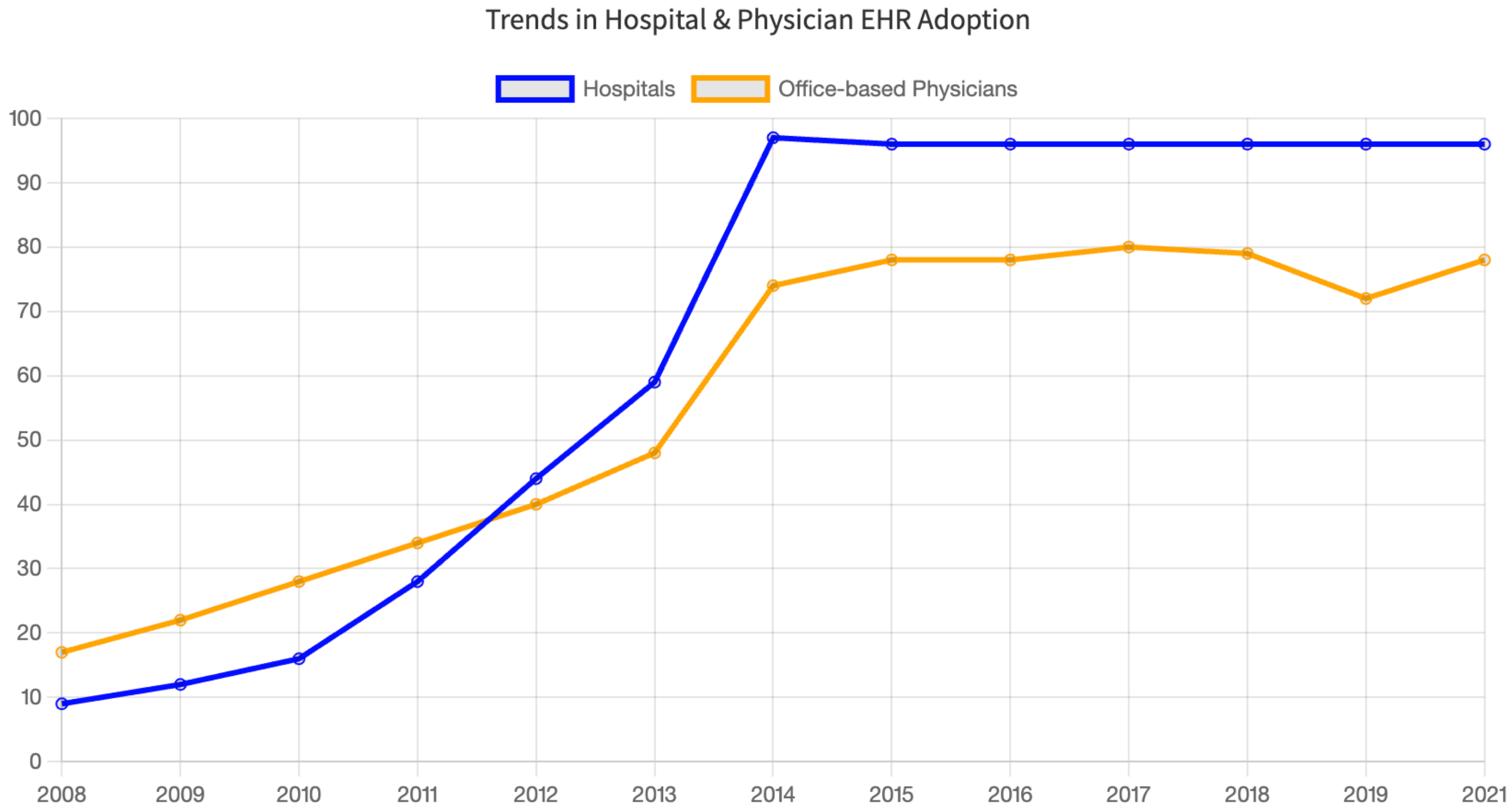
PennState

Electronic Health Records (EHR)

- Digital versions of patients' paper charts
- A patient's medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory and test results



National Trends in Hospital and Physician Adoption of Electronic Health Records (EHR)

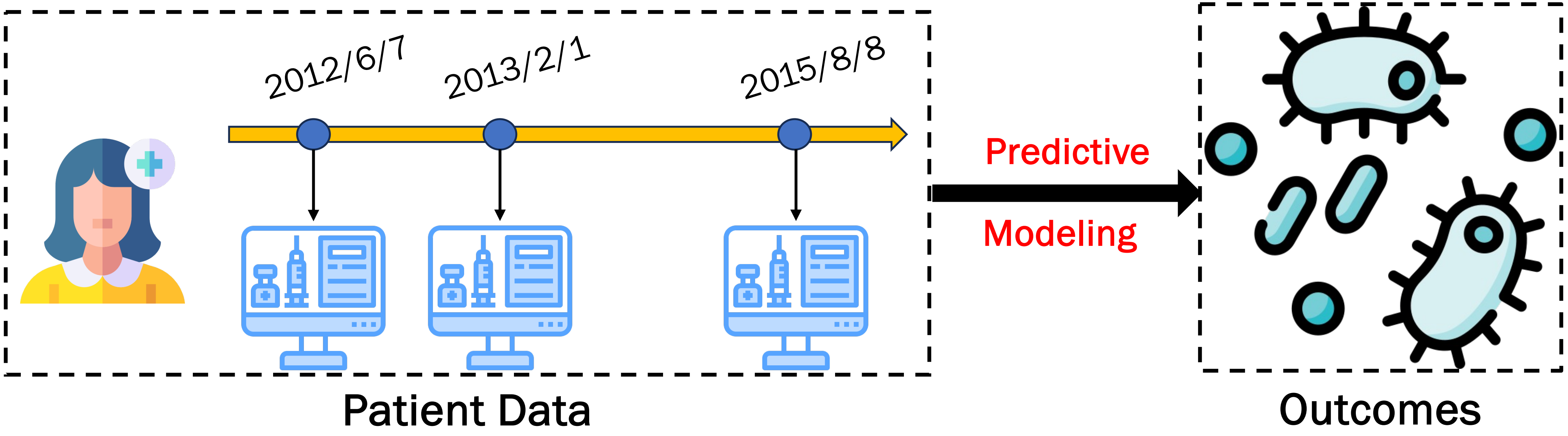


As of 2021, nearly 4 in 5 office-based physicians (78%) and nearly all non-federal acute care hospitals (96%) adopted a certified EHR. This marks substantial 10-year progress since 2011 when 28% of hospitals and 34% of physicians had adopted an EHR.

Office of the National Coordinator for Health Information Technology. 'National Trends in Hospital and Physician Adoption of Electronic Health Records,' Health IT Quick-Stat #61.

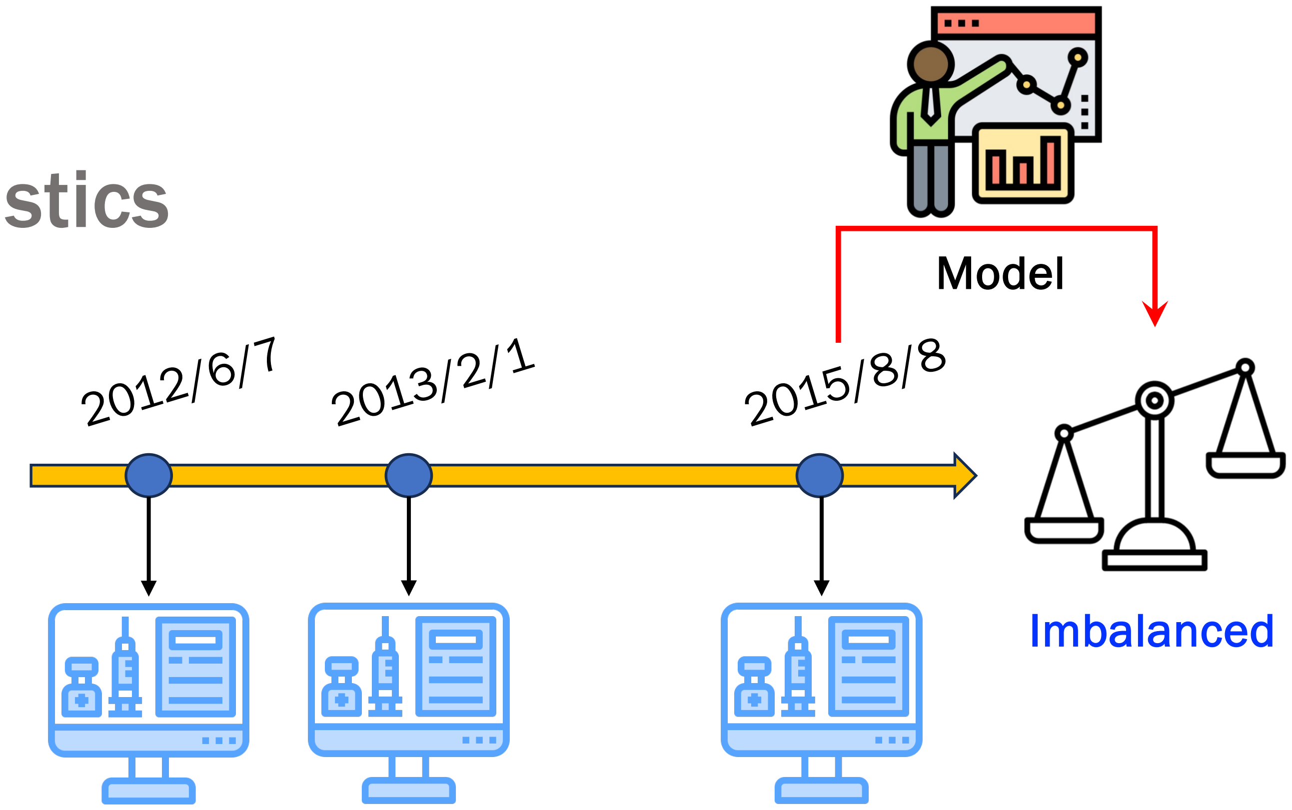
Predictive Modeling

- Using machine learning techniques to analyze patients' historical data along with current observations to support diagnosis or make predictions



EHR Data Unique Characteristics

- Temporal Dynamics
- Multimodalities and Heterogeneity
- High Dimensionality
- Imbalanced Data
- Clinical Explainability



International Classification of Diseases (ICD)

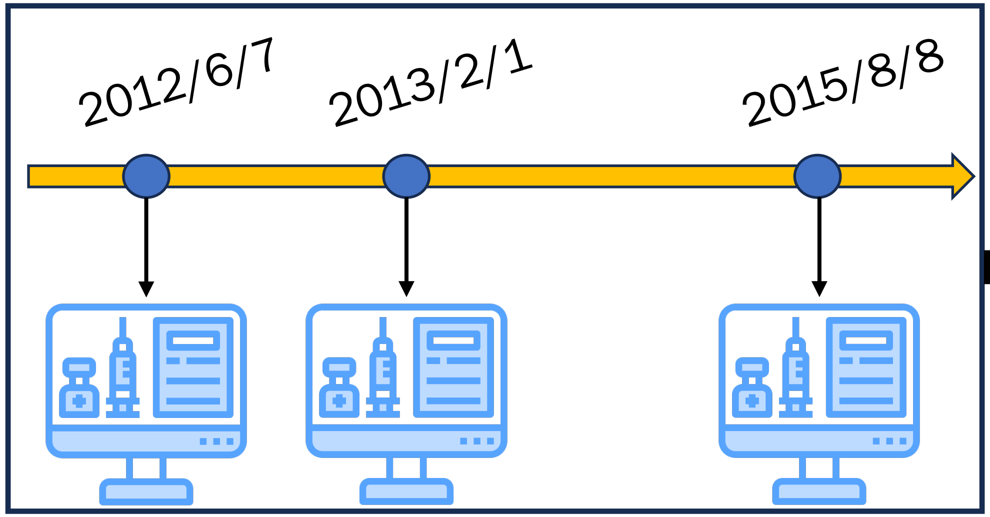
Version 9: ~18,000
Version 10: ~138,000

Existing Progress

- Basic Deep Learning-based Predictive Models
- Time–aware Predictive Modeling
- Predictive Modeling with Multimodal Data
- AutoML-based Predictive Modeling
- Knowledge-Enhanced Predictive Modeling
- Predictive Modeling with Imbalanced Classes
- Interpretable Predictive Modeling

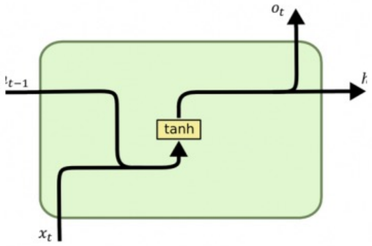
Basic Deep Learning-based Predictive Models

Temporal Dynamics

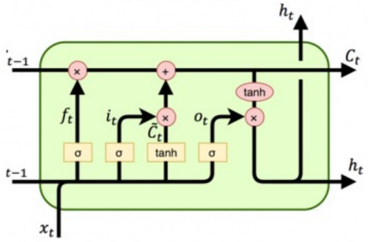


Retain (NeurIPS'16)
Dipole (KDD'17)

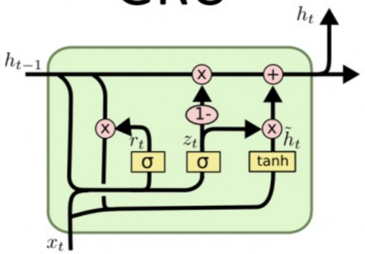
RNN



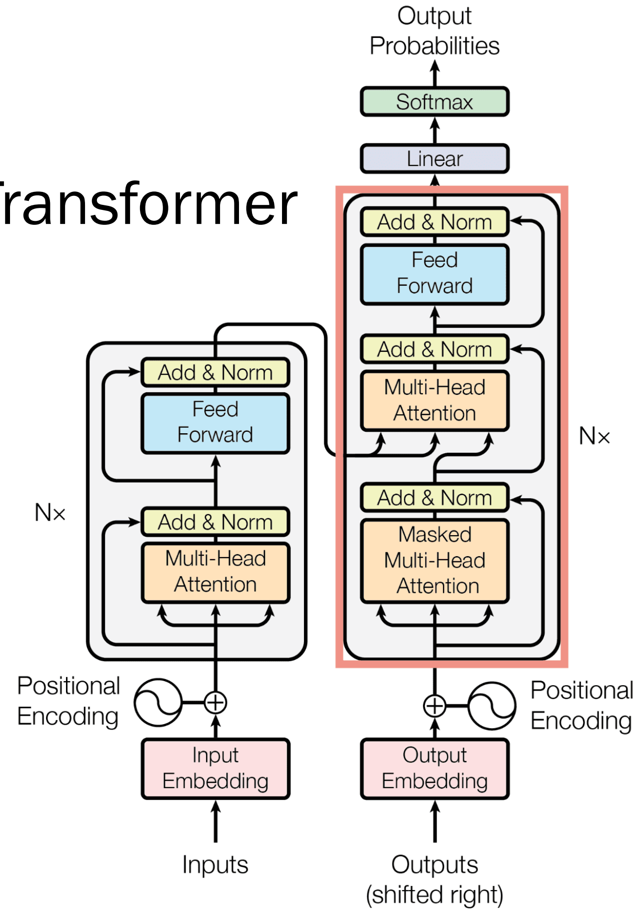
LSTM



GRU



Transformer

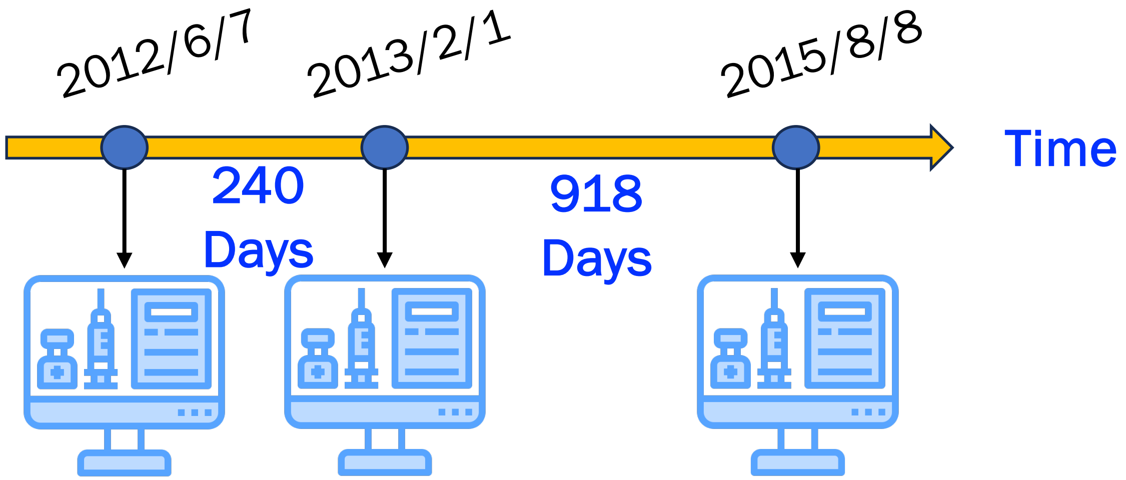


HiTANet (KDD'20)

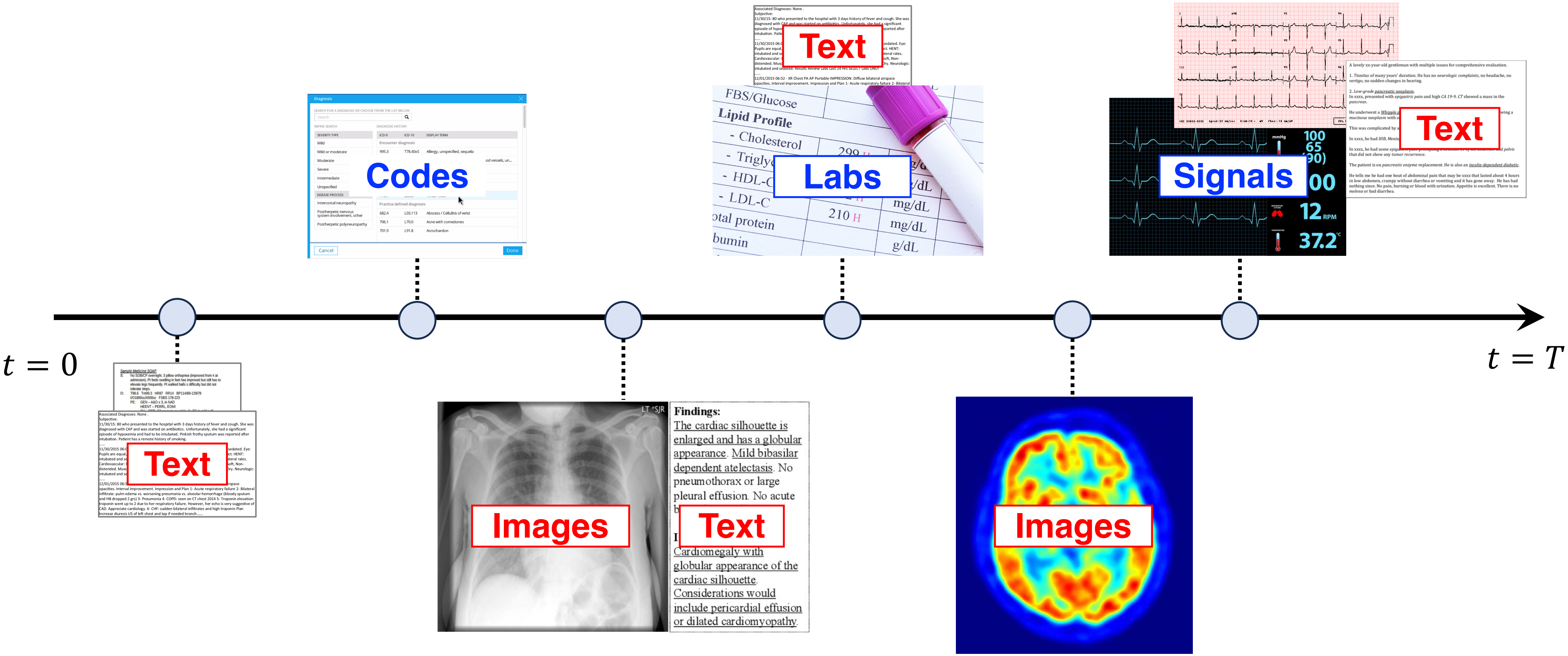
Time-aware Predictive Modeling

- Contrasting with textual data, the sequence of EHR data hinges on time information, and the intervals between recordings often vary. Accurately modeling this time aspect is essential for evaluating the impact of each patient visit.

T-LSTM (KDD'17)
TimeLine (KDD'18)
HiTANet (KDD'20)



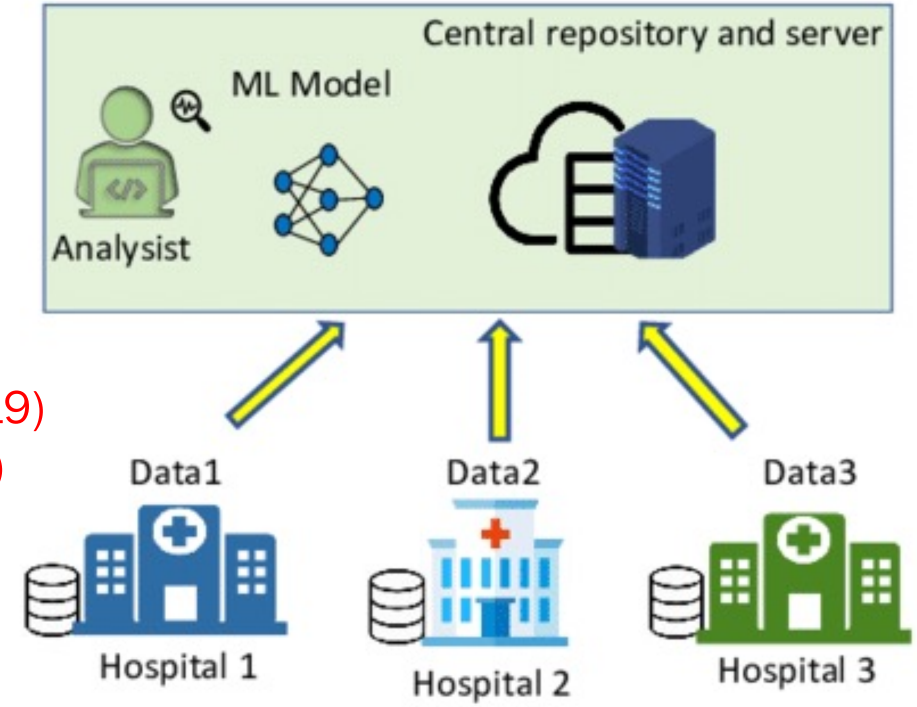
Predictive Modeling with Multimodal Data

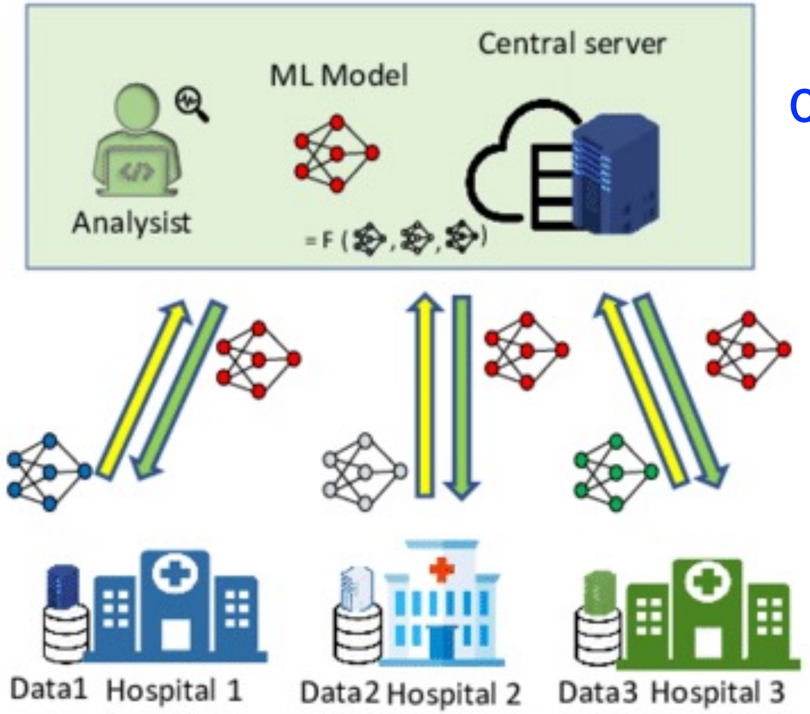


Predictive Modeling with Multimodal Data

Modality-level Feature Interaction/ Fusion

DCMN (ICDM'19)
RAIM (KDD'18)

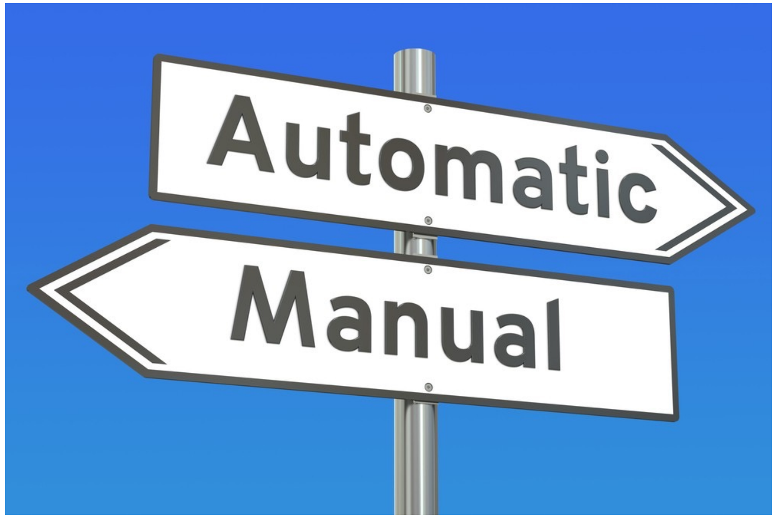


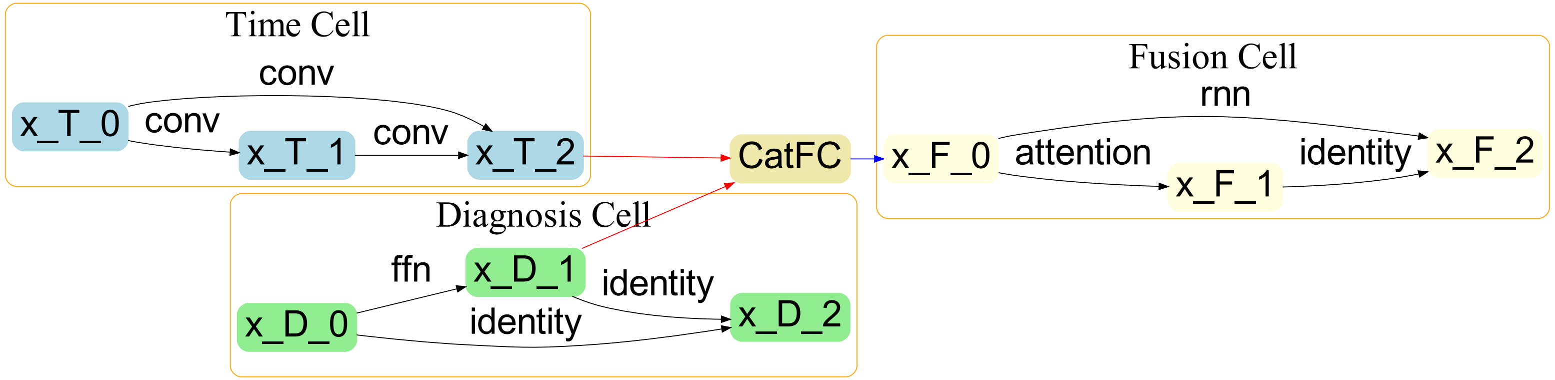


Client Data Privacy

FedCOVID (ECML-PKDD'21)

AutoML-based Predictive Modeling

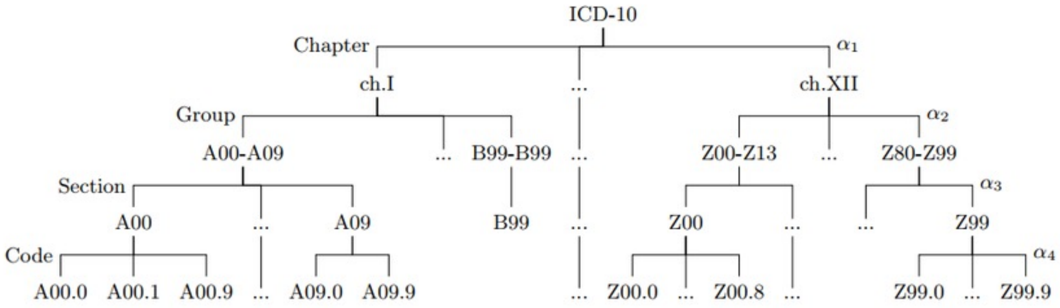




MUFASA (AAAI'21) AutoMed (BIBM'22) AutoFM (SDM'24)

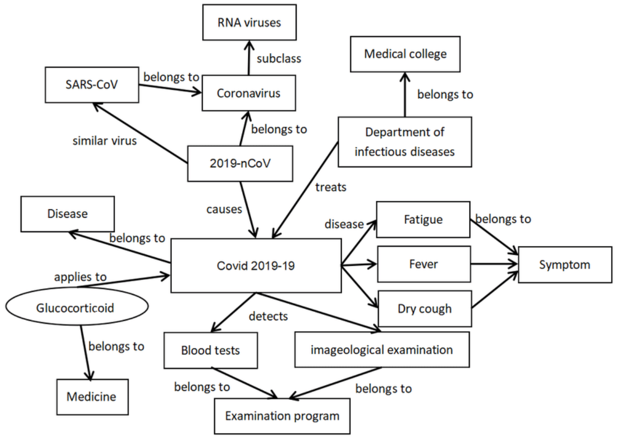
Knowledge-Enhanced Predictive Modeling

- Structured Knowledge

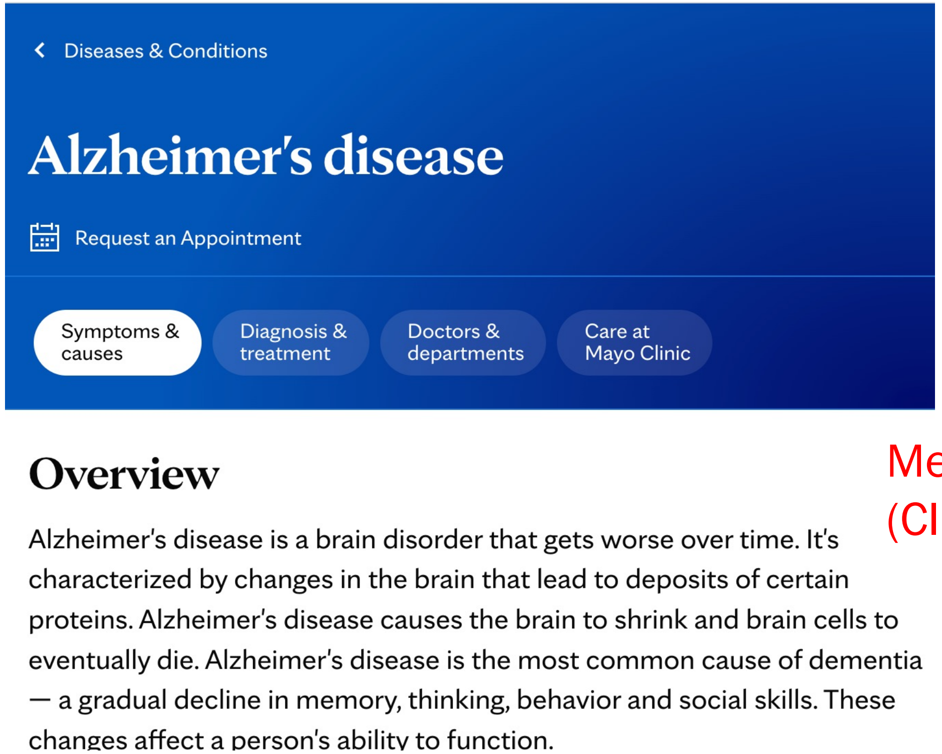


GRAM (KDD'17)
KAME (CIKM'18)
PRIME (KDD'18)

MedPath (WWW'21)



- Unstructured Knowledge

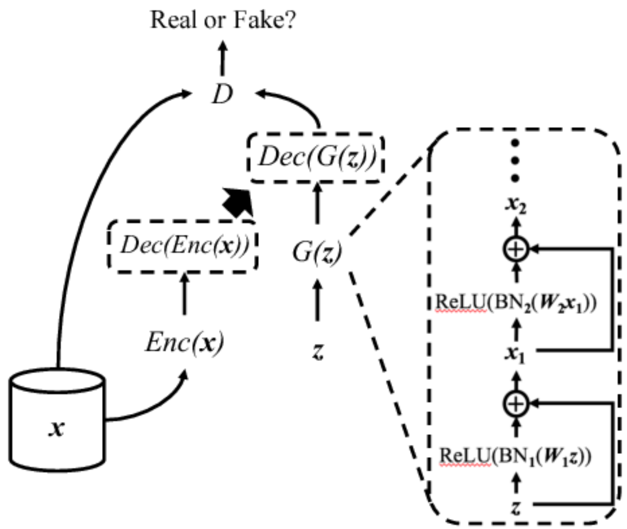


MedRetriever (CIKM'21)

Predictive Modeling with Imbalanced Classes

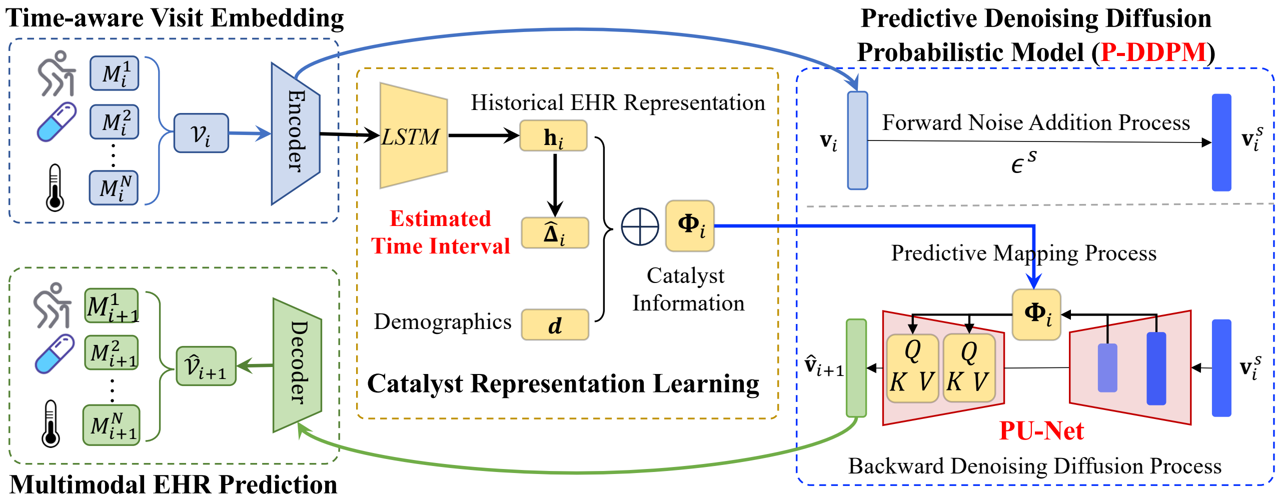
- Oversampling and Undersampling Techniques
- Generative Techniques

MedGAN (MLHC'17)
MaskEHR (SDM'20)
synTEG (JAMIA'21)



Generative Adversarial Networks (GAN)

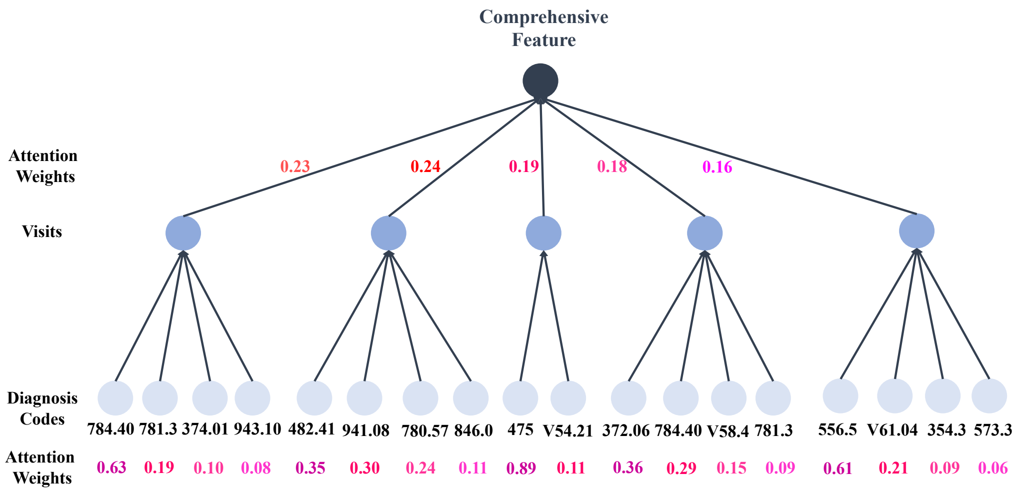
MedDiffusion (SDM'24) EHRPD (KDD'24)



Diffusion Models

Interpretable Predictive Modeling

- Attention-based Interpretation



LSAN (CIKM'20)

- Personalized Knowledge Graph-based Interpretation

EHR Data		**Visit 1**: 250.02 (Diabetes mellitus); **Visit 2**: 585.9 (Chronic kidney disease) and 780.79 (Fatigue); **Visit 3**: 244.9 (Hypothyroidism), 272.4 (Hyperlipidemia), and 401.1 (Benign essential hypertension); **Visit 4**: 585.9 (Chronic kidney disease); **Visit 5**: 585.9 (Chronic kidney disease); **Visit 6**: 585.9 (Chronic kidney disease) and 244.9 (Hypothyroidism)
1st Highest Attention Weighted Path	Weight: 0.0189	Hypothyroidism $\xrightarrow[E1]{CAUSES}$ Hypertensive disease $\xrightarrow[E2]{CAUSES}$ Left heart failure
	Evidence E1	*Animal studies suggest that hypertension leads to cardiac tissue hypothyroidism a condition that can by itself lead to heart failure.*
	Evidence E2	*Left ventricular failure in some SA/OHS patients may be the result of hypertensive cardiac disease.*
2nd Highest Attention Weighted Path	Weight: 0.0178	Hyperlipidemia $\xrightarrow[E3]{CAUSES}$ Hypertensive disease $\xrightarrow[E4]{CAUSES}$ Left heart failure
	Evidence E3	*A literature search indicates that Anglo-Saxon countries report alarming hyperplastic changes particularly in the liver blood clots hyperlipidemia leading to high blood pressure porphyria atypical leiomyomas and cervical hyperplasia.*
	Evidence E4	*Left ventricular failure in some SA/OHS patients may be the result of hypertensive cardiac disease.*
3rd Highest Attention Weighted Path	Weight: 0.0150	Fatigue $\xrightarrow[E5]{CAUSES}$ Cessation of life $\xrightarrow[E6]{CAUSES}$ Left heart failure
	Evidence E5	*In light of the magnitude of this sleep debt it is not surprising that fatigue is a factor in 57% of accidents leading to the death of a truck driver and in 10% of fatal car accidents and results in costs of up to 56 billion dollars per year.*
	Evidence E6	*Though rare death due to myocardial stunning and LV power failure can occur during ICD insertion.*
One of the Lowest Attention Weighted Path	Weight: 0.0000	Heart failure $\xrightarrow[E7]{CAUSES}$ Hypertensive disease $\xrightarrow[E8]{CAUSES}$ Left heart failure
	Evidence E7	*These findings suggest that the ATF3 activator tBHQ may have therapeutic potential for the treatment of pressure-overload heart failure induced by chronic hypertension or other pressure overload mechanisms.*
	Evidence E8	*Left ventricular failure in some SA/OHS patients may be the result of hypertensive cardiac disease.*

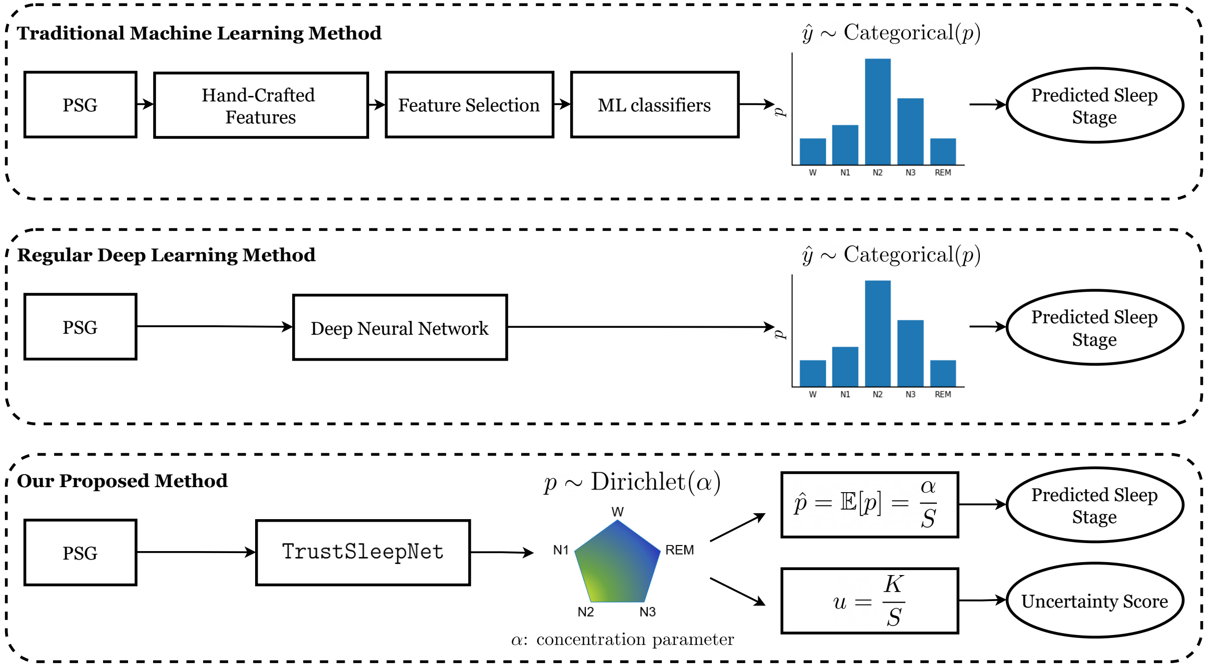
MedPath (WWW'21)

Interpretable Predictive Modeling

- Medical Text-based Explicit Interpretation
- Uncertainty-based Interpretation

EHR	**Visit 1:** Esophageal reflux (530.81), Acute conjunctivitis (372.00), Asthma (493.90) **Visit 2:** Conjunctivitis (372.30) **Visit 3:** Other mucopurulent conjunctivitis (372.03) **Visit 4:** Lumbago (724.2), Unspecified contraceptive management (V25.9) **Visit 5:** Lumbago (724.2), Asthma (493.90), Nausea with vomiting (787.01)
Target Guidance	1. Asthma, a chronic inflammatory airway disease, may be a risk factor for developing COPD. The combination of asthma and smoking increases the risk of COPD even more. *(Weight: 0.034482)* 2. Exposure to tobacco smoke. The most significant risk factor for COPD is long-term cigarette smoking. The more years you smoke and the more packs you smoke, the greater your risk. Pipe smokers, cigar smokers and marijuana smokers also may be at risk, as well as people exposed to large amounts of secondhand smoke. *(Weight: 0.034479)*
Text Memory (Visit 1-4)	1. Proper treatment makes a big difference in preventing both short-term and long-term complications caused by asthma. *(Weight: 0.05109)* 2. Exposure to various irritants and substances that trigger allergies (allergens) can trigger signs and symptoms of asthma, including: Respiratory infections such as the common cold, Physical activity, Air pollutants and irritants such as smoke, Strong emotions and stress, Gastroesophageal reflux disease (GERD) and etc. *(Weight: 0.05107)* 3. Signs that your asthma is probably worsening include: Asthma signs and symptoms that are more frequent and bothersome, Increasing difficulty breathing, The need to use a quick-relief inhaler more often and etc. *(Weight: 0.05106)* 4. Asthma complications include: Signs and symptoms that interfere with sleep, work and other activities, Sick days from work or school during asthma flare-ups, A permanent narrowing of the tubes that carry air to and from your lungs (bronchial tubes), which affects how well you can breathe. *(Weight: 0.05105)* 5. Conditions that can increase your risk of GERD include: Obesity, Pregnancy, Connective tissue disorders, such as scleroderma and etc *(Weight: 0.05104)*
Text Memory (Visit 5)	1. Exposure to various irritants and substances that trigger allergies (allergens) can trigger signs and symptoms of asthma, including: Respiratory infections such as the common cold, Physical activity, Air pollutants and irritants such as smoke, Strong emotions and stress, Gastroesophageal reflux disease (GERD) and etc. **(appears twice)** *(Weight: 0.05015)* 3. Asthma complications include: Signs and symptoms that interfere with sleep, work and other activities, Sick days from work or school during asthma flare-ups, A permanent narrowing of the tubes that carry air to and from your lungs (bronchial tubes), which affects how well you can breathe. *(Weight: 0.05012)* 4. Asthma signs and symptoms include: Shortness of breath, Chest tightness or pain, Wheezing when exhaling, which is a common sign of asthma in children, Trouble sleeping caused by shortness of breath, coughing or wheezing, Coughing or wheezing attacks that are worsened by a respiratory virus, such as a cold or the flu. *(Weight: 0.05012)* 5. A number of factors are thought to increase your chances of developing asthma. They include: Being a smoker, Exposure to secondhand smoke, Exposure to exhaust fumes or other types of pollution and etc. *(Weight: 0.05011)*

MedRetriever (CIKM'21)

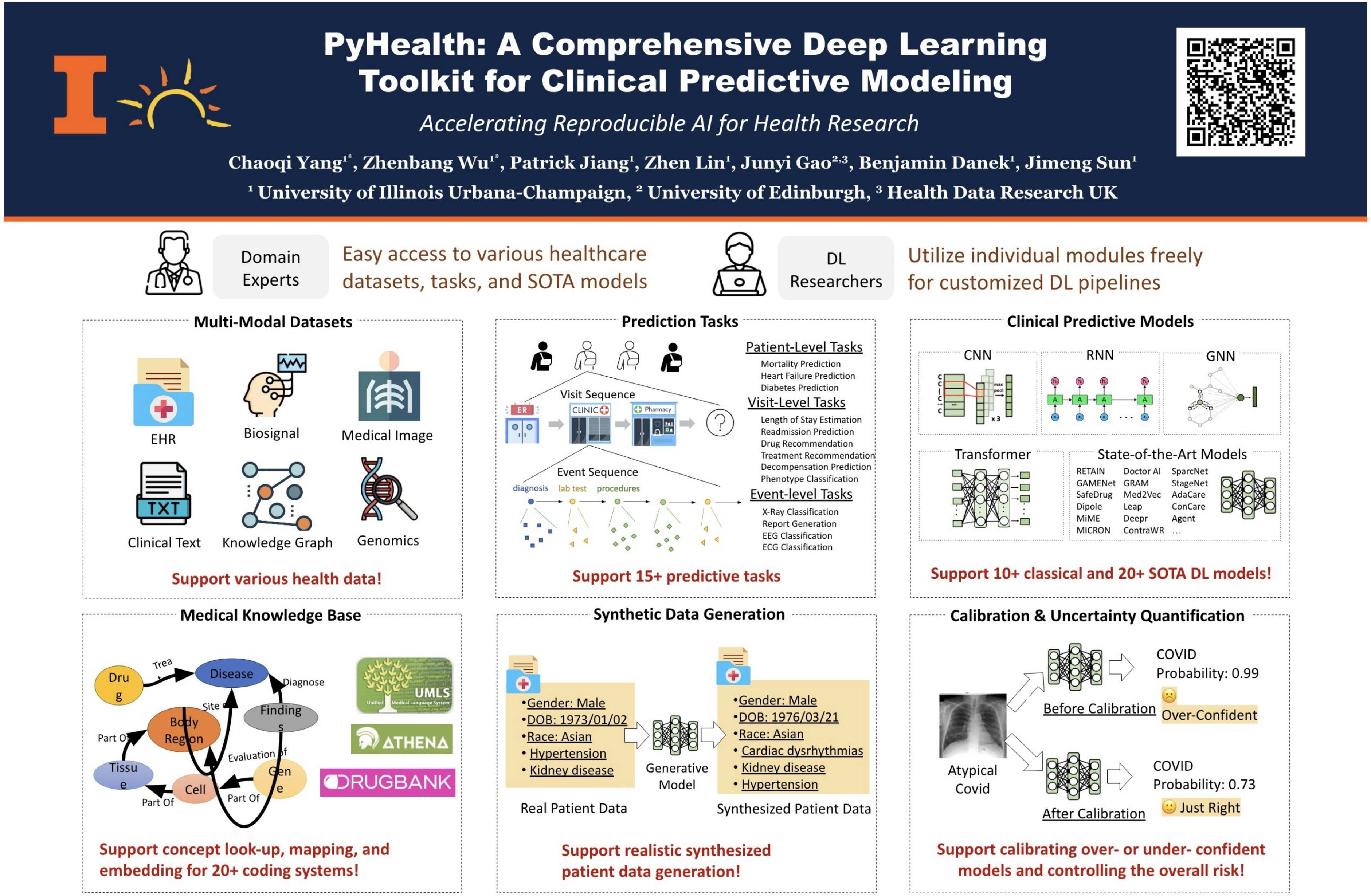


TrustSleepNet (BHI'22)

Benchmarks

Name	Data Type	# of Data	Modalities	Link
MIMIC-III	Real	38,597 patients	Demographics, vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data	https://physionet.org/content/mimiciii/1.4/
MIMIC-IV	Real	40,000+ patients	Demographics, vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data	https://physionet.org/content/mimiciv/2.2/
MIMIC-CXR	Real	377,110 images 227,835 reports	Electronic health record data, images (chest radiographs), and natural language (free-text reports)	https://physionet.org/content/mimic-cxr/2.0.0/
eICU	Real	200,000+ admissions	Vital sign measurements, care plan documentation, severity of illness measures, diagnosis information, treatment information, and more	https://physionet.org/content/mimic-cxr/2.0.0/
PPMI	Real	2,230 patients	Subject characteristics, biospecimen, images, medical history, etc.	https://www.ppmi-info.org/
ADNI	Real	2,775 patients	Subject characteristics, genetic data, images, medical history, neuropathology, etc.	https://adni.loni.usc.edu/
Apnea-ECG	Real	70 recordings	Subject characteristics, electrocardiogram	https://physionet.org/content/apnea-ecg/1.0.0/
MIT-BIH PSG	Real	18 recordings	Subject characteristics, electrocardiogram, electroencephalography, electrooculography, electromyography, etc.	https://physionet.org/content/slpdb/1.0.0/
SHHS	Real	6,441 patients	Subject characteristics, electrocardiogram, electroencephalography, electrooculography, electromyography, airflow, etc.	https://sleepdata.org/datasets/shhs
Newcastle-Accel	Real	28 patients	Subject characteristics, acceleration, polysomnography.	https://zenodo.org/records/1160410#.YLqiSC1h1qt
Sleep-Accel	Real	31 patients	Acceleration, heart rate, steps.	https://physionet.org/content/sleep-accel/1.0.0/
EMRBOTS	Synthetic	100,000 patients	Patients' admissions, demographics, socioeconomics, labs, medications, etc.	http://www.emrbots.org/
Project Data Sphere	Real	242 studies	Data provider, sponsor, study phase, linked data, tumor type, access, etc.	https://www.projectdatasphere.org

Toolkits



https://pyhealth.readthedocs.io/en/latest/

Open Challenges and Future Directions

- Trustworthy Predictive Modeling

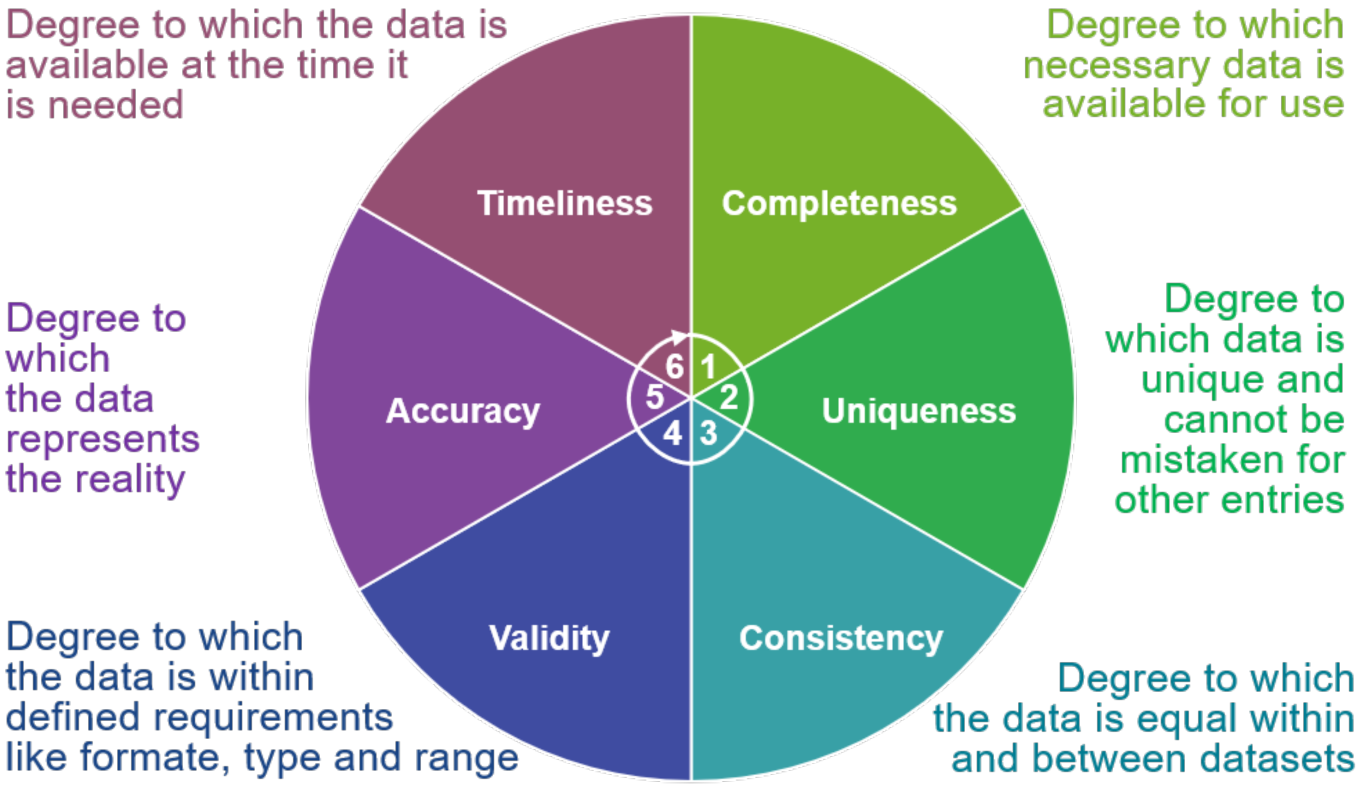
LLM-driven interpretable model design

Ethical model design

Human-in-the-loop learning

Open Challenges and Future Directions

- Data Scarcity/Sparsity



Open Challenges and Future Directions

- Pre-training Across Multiple Data Sources

Wang et al., [Unity in Diversity: Collaborative Pre-training Across Multimodal Medical Sources](), ACL'24

Open Challenges and Future Directions

- Federated Training for Foundation Models

Wang et al., *FedMeKI: A Benchmark for Scaling Medical Foundation Models via Federated Knowledge Injection*, under review

Wang et al., *FedKIM: Adaptive Federated Knowledge Injection into Medical Foundation Models*, under review

Paper

Lab

Personal

Thank You.

Any questions, please feel free contact Jiaqi Wang or Fenglong Ma via jqwang@psu.edu or fenglong@psu.edu.