BMI 702: Biomedical Artificial Intelligence

Foundations of Biomedical Informatics II, Spring 2023

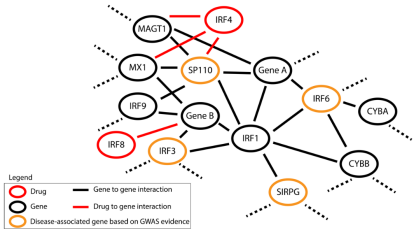
Lecture 9: Clinical trial site identification, patient trial matching, clinical trial recruitment

HARVARD
MEDICAL SCHOOL

Marinka Zitnik
marinka@hms.harvard.edu

Problem set 2 is released

BMI702_Hw2_2023.pdf
Page 2 of 7



Questions:

1. (20 points) For the first task, we want to identify whether "Gene A" and "Gene B" are associated with autoimmune disease. In the PPI network shown above, nodes with GWAS evidence are associated with the disease. With this in mind, we want to see if nodes "Gene A" and "Gene B" are either disease-associated nodes or not disease-associated nodes.
   (a) (5 points) What is the name of the machine learning task we are solving in this exercise?
   (b) (5 points) Using the local hypothesis principle, assign labels to nodes "Gene A" and "Gene B", using the direct neighbor scoring approach. Please provide reasoning for your assignments.
   (c) (5 points) Using the local hypothesis principle, assign labels to nodes "Gene A" and "Gene B", using the direct weighted neighbor scoring approach. Please provide reasoning for your assignments.
   (d) (5 points) A colleague suggests examining 5-hop neighbors of node "Gene A " to predict biological functions of "Gene A"'. Given the local hypothesis principle, do you expect that considering a 5-hop neighborhood is necessary? Note that the diameter (i.e., the shortest distance between the two most distant nodes in the network) of the entire PPI network is less than 9. Why or why not?
2. (5 points) Shown in the PPI network are two drugs and their interactions with genes in the network. Based on your analysis in the previous question and Barrio-Hernandez et al. (2023), which of the two drugs would you investigate further as a more promising therapeutic candidate? Please answer in 3-4 sentences.

2 Introduction to Multimodal Graph Learning [25 points]

You are a recently hired ML scientist at Genentech. You are fresh out of your master's degree, where you had a capstone project exploring techniques to improve the training of graph neural networks (GNNs) in protein-protein interaction networks. On your first day, your manager calls you into their office and tells you they hired you because of your expertise in training GNNs. Some higher-ups at Genentech read a recent review in Nature Machine Intelligence and were inspired to

2

BMI702_Hw2_2023.pdf
Page 4 of 7

5. (5 points) In your opinion, why is MGL relevant to biomedicine? Please answer in 4-5 sentences.

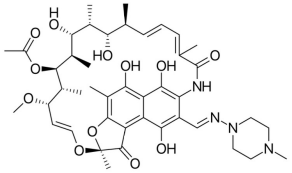
3 MGL for Antibiotic Discovery for Tuberculosis [50 points]

The WHO has tasked Genentech to discover the next generation of antibiotics to address the threat of antibiotic resistance. You are an ML scientist at Genentech, and your manager has tasked you to assist in the effort.

3.1 In-Silico Compound Screening Using Deep Learning [25 points]

Your manager asks you first to use MGL to create a model that, given the structure of a small molecule, predicts whether or not the small molecule can inhibit the growth of tuberculosis. Inspired by reading the 2020 paper in Cell by Stokes et al., you set out on your task. For the coding section of this problem please use this Google Colab notebook. As in the last problem set, create and modify a copy of this notebook. This notebook will be used for questions 3.1.2, 3.1.3, and 3.1.4.

Note: For this question, you will use the same train and test data from the Cell paper highlighted in the introduction. You will create a machine learning model to predict growth inhibition in E. coli because data on tuberculosis is limited.



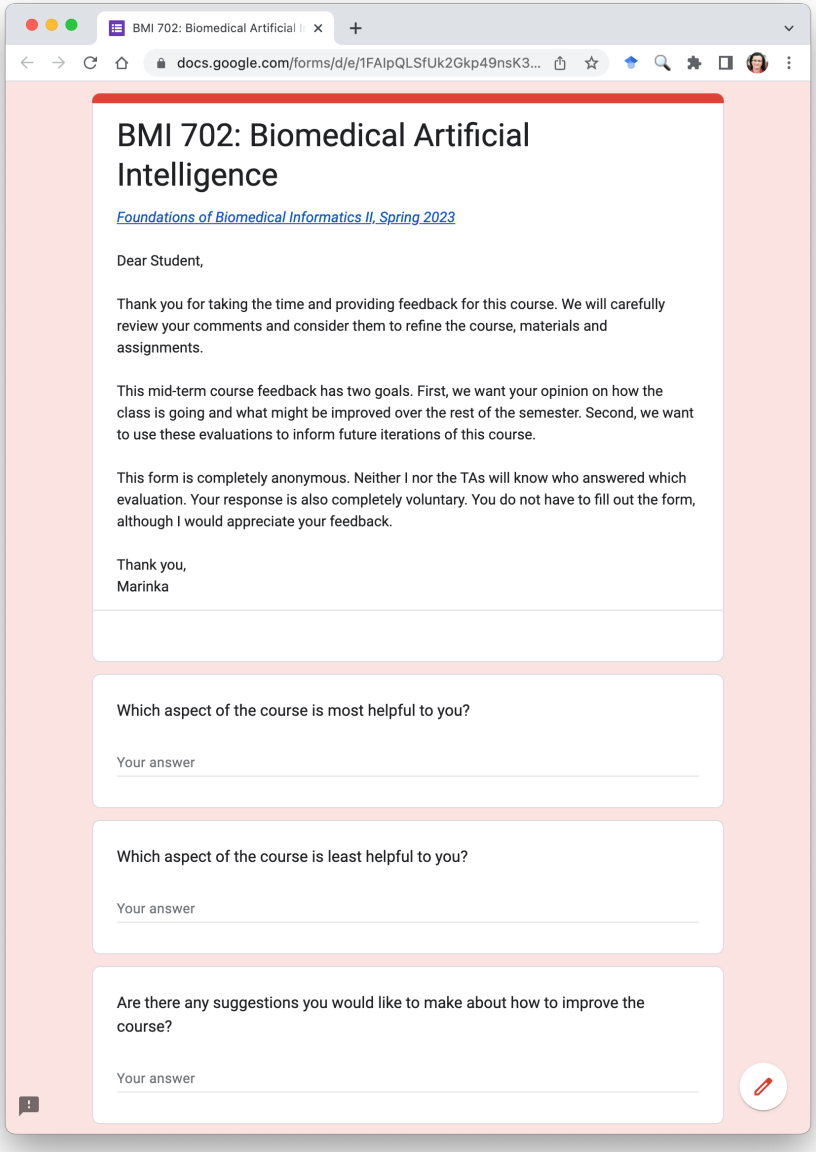
Questions:

1. (5 points) The following questions concern the above-shown molecule of Rifampicin and how to construct a graph that can be used in a GNN from the molecular structure of RIF. Answer the following questions, each in 1-2 sentences.
   a) What would the nodes be in a graph representing the molecular structure of Rifampicin?
   b) What would the edges be?
   c) What would some potential node and edge features be?
2. (2 points) As a first step for model development, you run a simple model to establish a baseline in performance or control. This is a classification task where given a small molecule (represented by the SMILES chemical statistics fingerprint), the model should predict 1 (growth inhibition in E. coli) or 0 (no growth inhibition). After reading this technical report, you

4

Thematic focus is on networks, molecular graphs, geometric deep learning, and graph representation learning

Mid-term class feedback



Thank you for taking the time and providing feedback for the course!

https://forms.gle/NEtazbg5neysMu7WA

Outline for today's class

a. Clinical trial recruitment
   - Doctor selection
   - Site selection

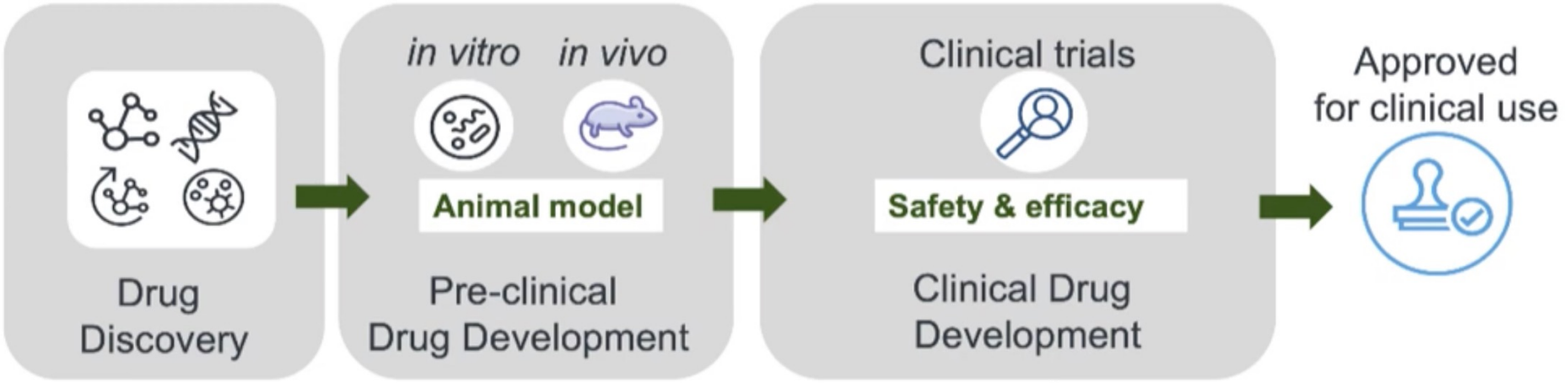
b. Patient-trial matching

c. Trial outcome prediction

Clinical trials

- Global clinical trial market has reached $44.3 billion in 2020. It is expected to grow to $69.3 billion by 2028
  - Costs of conducting clinical trials are high (up to hundreds of millions of dollars)
- Trials can take multiple years, with a low success probability
  - Many factors can lead to failure:
    - Poor efficacy of the drug in development
    - Drug safety issues and adverse events
    - Poor trial protocol design
    - Failure to recruit patients

Drug discovery & development

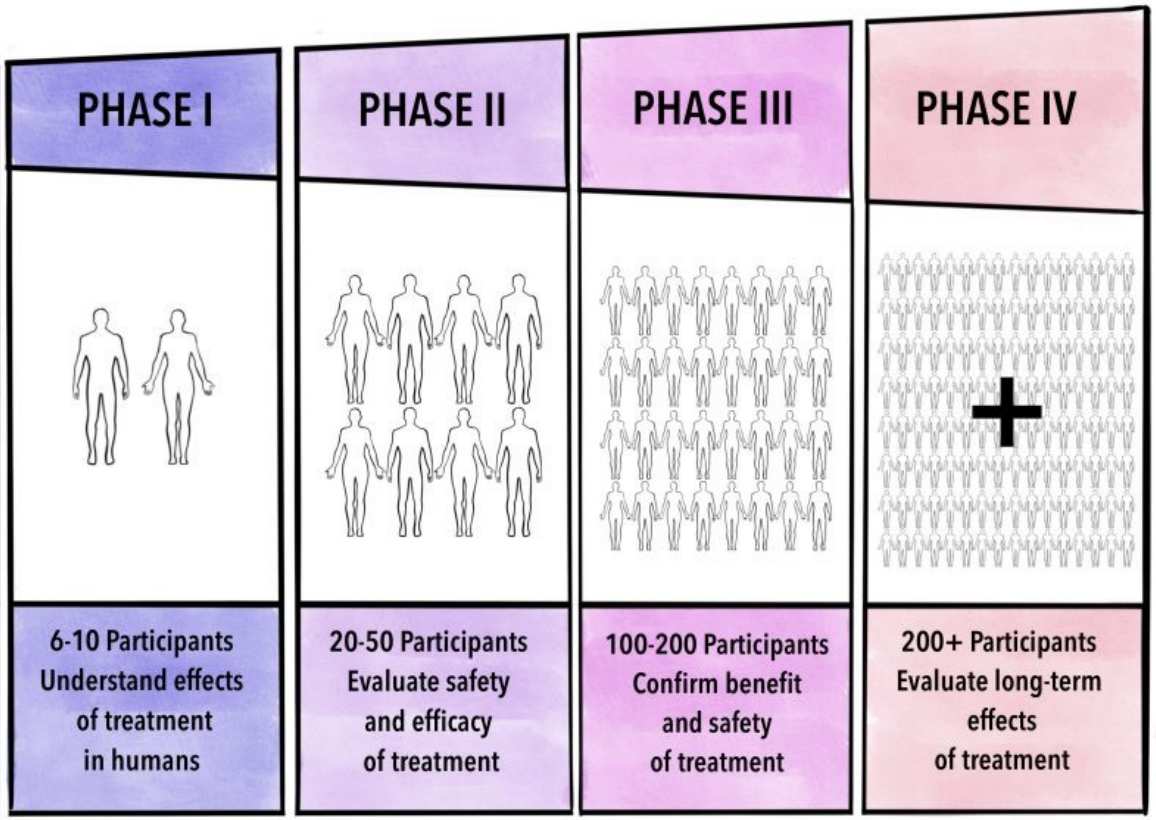


	Drug discovery	Pre-clinical	Phase 1	Phase 2	Phase 3
Time spent	4-5 years	1-2 years	1-2 years	1-2 years	2-3 years
$ spent	$550M	$125M	$225M	$250M	$250M
Output	5,000 - 10,000 compounds	10-20 candidates	5-10 candidates	2-5 candidates	1-2 candidates

huge discovery cost

huge enrollment cost

Clinical phases and their costs



Extremely Low Success Rate!

** Gene Vision*
+ American Pharmaceutical Review

How can AI help improve effectiveness of clinical trials?

a **Site or Doctor Selection**

b **Patient Trial Matching**

c **Trial Outcome Prediction**

Part I

Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment

Biswal et al, Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment, AAAI 2020

Motivation

- Massive EHRs enable learning patient representations to support various predictive health applications
- In contrast, doctor representations are not well studied despite doctors having key role in healthcare:
  - How to create the "right" doctor representations?
  - How to use doctor representations for health analytics?
- Doctors play key role in recruiting patients into clinical trials for drug development
  - This study is about identifying the right doctors to help conduct trials based on the trial description and patient EHR data of those doctors
  - Thus, effective doctor representations can better support a wider range of health analytic tasks

Biswal et al, Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment, AAAI 2020

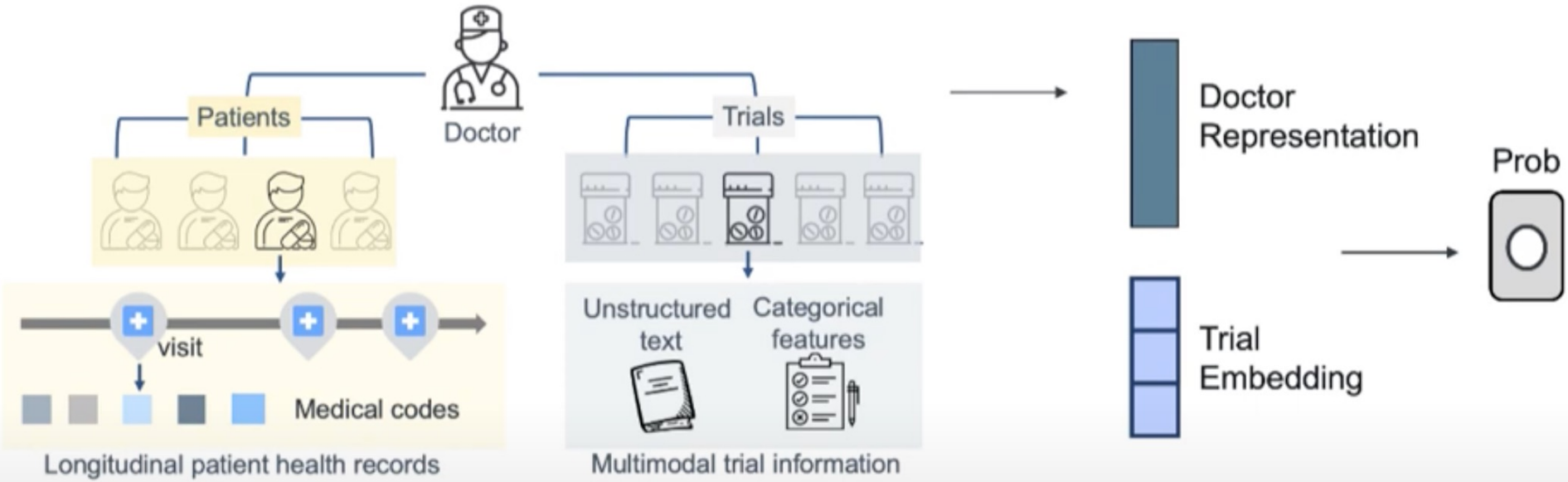
Motivation

1. Current practice determines median enrollment rate of 1 as predicted enrollment for every participating doctor:
   - It can be inaccurate
2. Multi-step manual matching process for site selection:
   - It can be labor-intensive
3. ML methods applied to site selection tasks via static medical concept embeddings using frequent medical codes and simple matching of keywords to trial descr:
   - No sense of evolving doctor's experience and expertise encoded in EHR data of patients the doctor has previously seen
   - Given a trial for a particular disease, the doctor's specific expertise in that disease is important
   - Doctor's representation should change based on specifics of a given trial

Biswal et al, Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment, AAAI 2020

Machine learning for doctor selection

Which doctors to recommend to conduct a clinical trial?

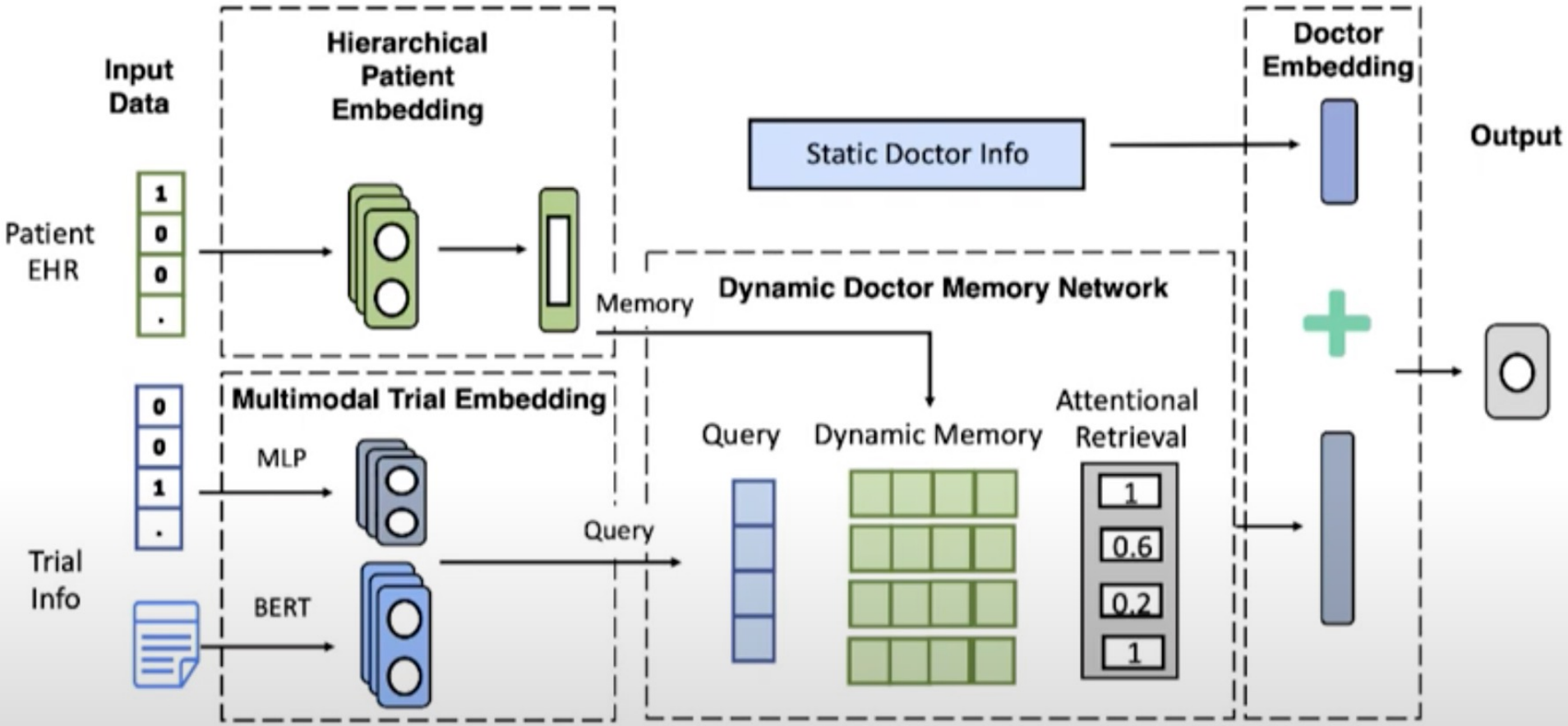


Challenges

1. How to capture the time-evolving patterns of doctors experience/expertise?
2. How to learn a dynamic representation based on the corresponding trial?

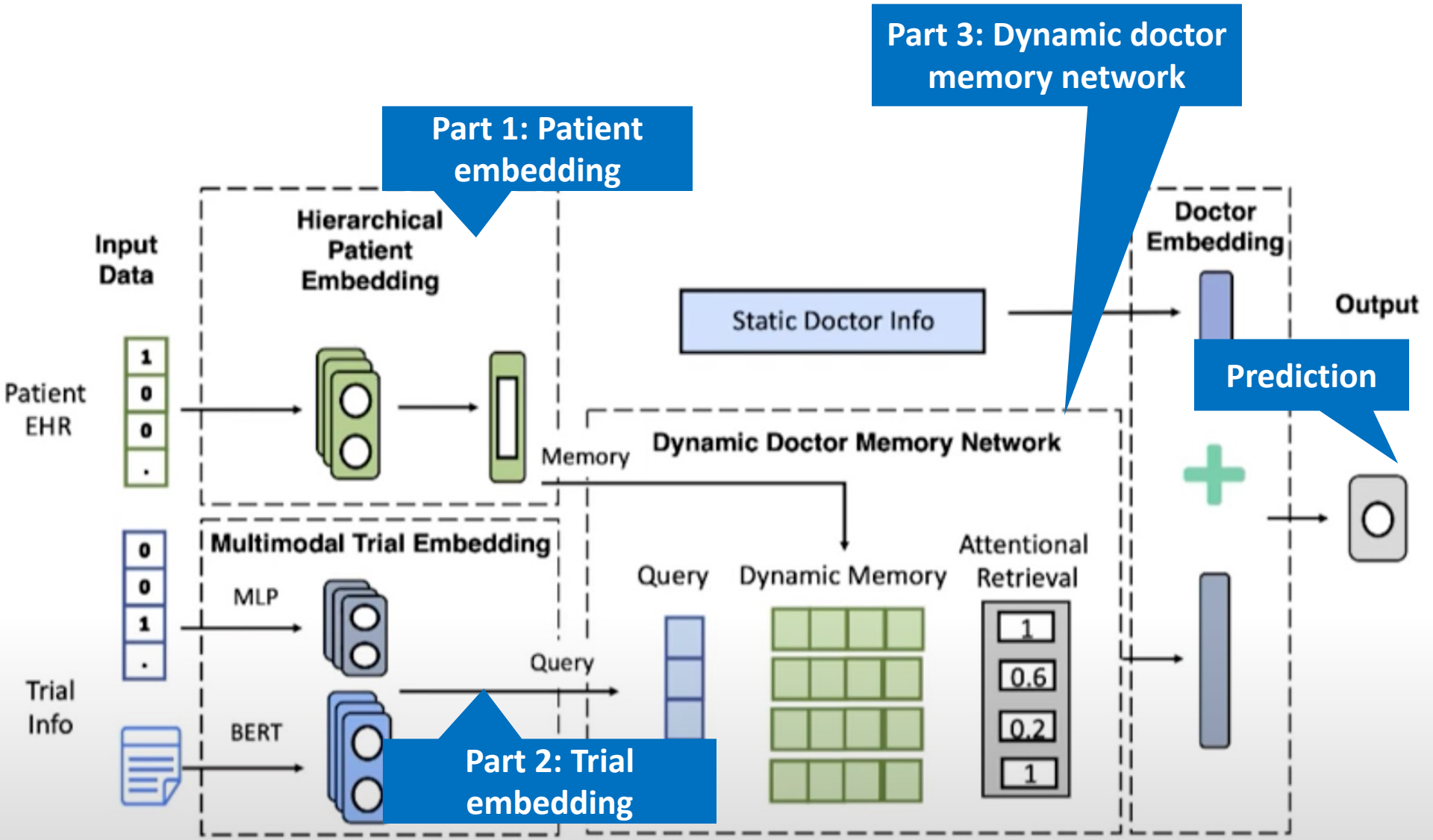
Biswal et al, Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment, AAAI 2020

Doctor2vec: Overview



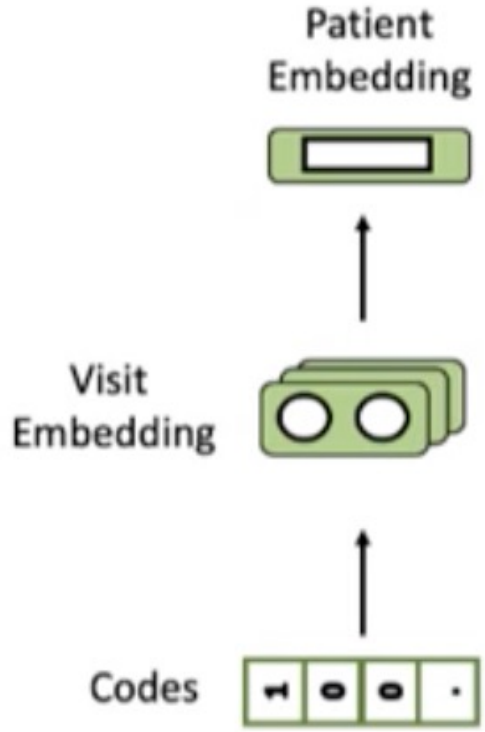
Biswal et al, Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment, AAAI 2020

Doctor2vec: Overview



Biswal et al, Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment, AAAI 2020

Part 1: Hierarchical patient embedding



$$\mathbf{I(k)} = \sum \alpha_t \cdot \mathbf{h_t}$$

$$\mathbf{g_1}, \mathbf{g_2}, ...\mathbf{g_t} = \text{bi-LSTM}(\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_t})$$

$$\mathbf{e_t} = \mathbf{w}_\alpha^\mathbf{T} * \mathbf{g_t} + \mathbf{b}_\alpha$$

$$\alpha_1, \alpha_2, ..\alpha_t = \text{softmax}(\mathbf{e_1}, \mathbf{e_2}, ..\mathbf{e_t})$$

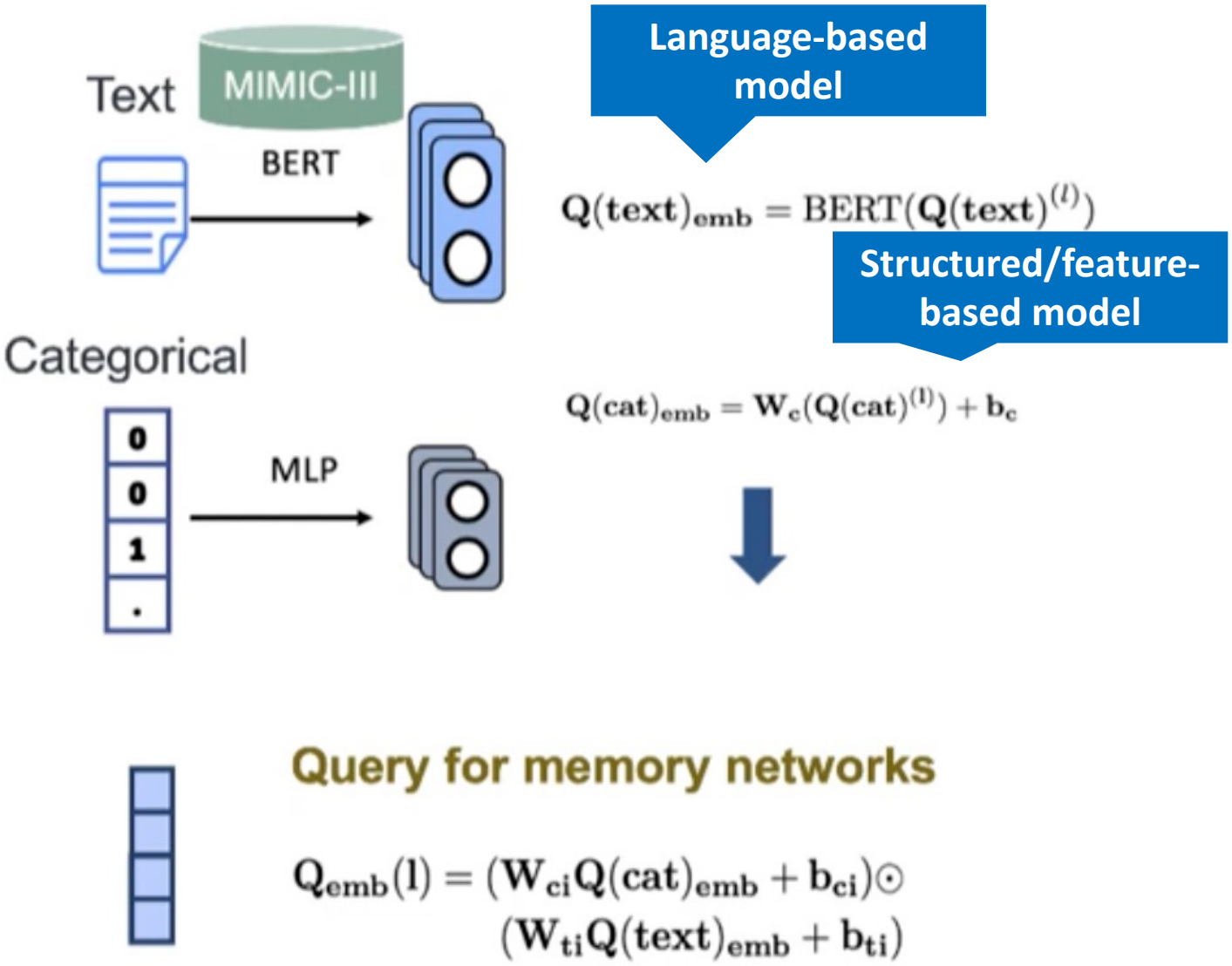
$$\mathbf{h_t(k)} = \mathbf{W_{emb}} * \mathbf{v_t(k)}$$

ML model attends (i.e., appropriately up-weights, down-weights) past visits for each patient

Memory for memory networks

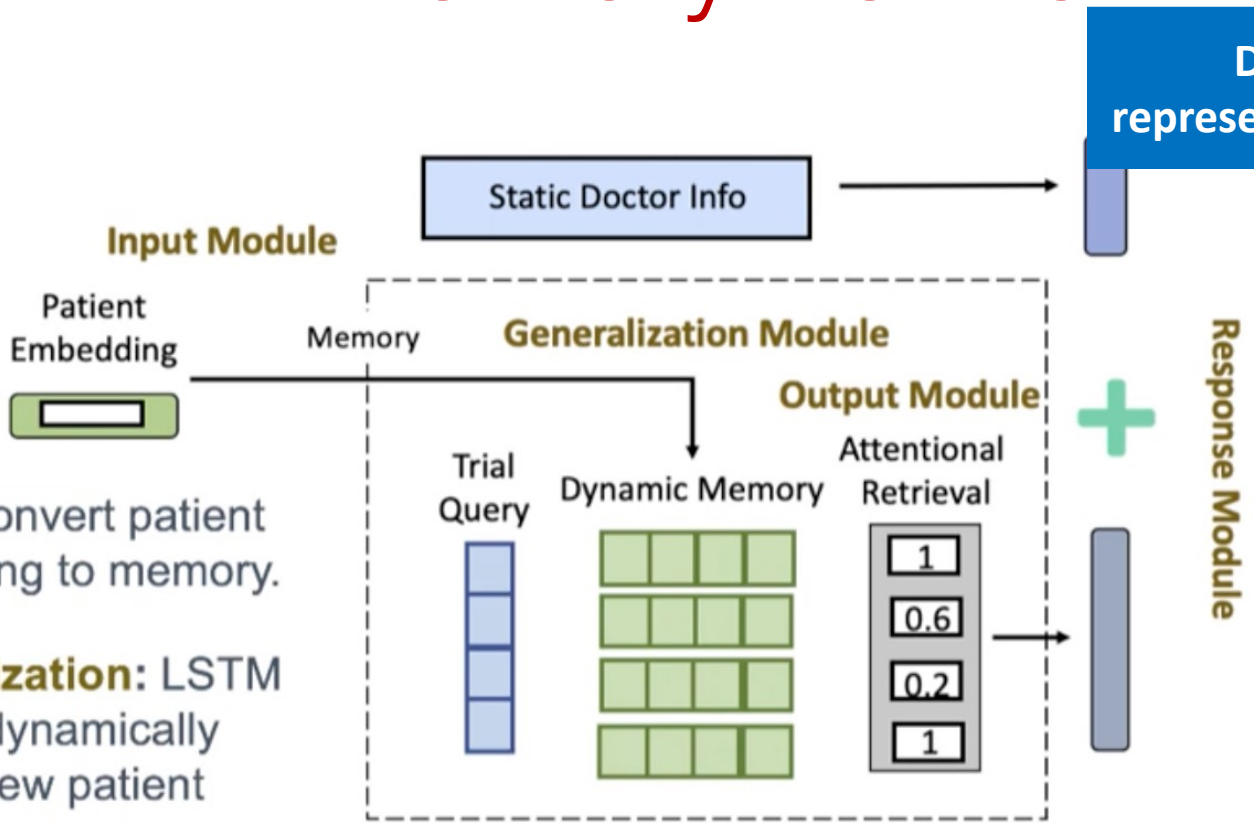
Biswal et al, Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment, AAAI 2020

Part 2: Multimodal trial embedding



Biswal et al, Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment, AAAI 2020

Part 3: Dynamic doctor memory network



Input: Convert patient embedding to memory.

Generalization: LSTM layer to dynamically update new patient memory.

Output: Attentional memory retrieval to learn doctor embedding only based on trial-specific memory.

Response: Concat. to get doctor embedding

Output predicted enrollment rate between 0 and 1

Both regression and classification

Biswal et al, Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment, AAAI 2020

Experimental setup: Data

a) IQVIA trial data about trials formed during 2014 and 2019 across 28 countries.
b) clinical trial description from clinicaltrials.gov, matched with IQVIA trial data on NCT ID
c) IQVIA claims data

Table 2: Data Statistics

# of clinical trials	2609
# of doctors	25894
# of doctor-trial pair(samples)	102487
# of patients	430,239
Avg # of Dx codes per visit	4.23
Max # of Dx codes per visit	56
Avg # of Procedure codes per visit	1.23
Max # of Procedure codes per visit	18
Avg # of Med codes per visit	9.36

Biswal et al, Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment, AAAI 2020

Setup: Metrics

Classification Task

The precision recall area under curve (PR-AUC) is the area under the PR curve. A good metric for data imbalanced setting. The higher the better.

Regression Task

The coefficient of determination (R-squared) is the square of the correlation between predicted scores and actual scores. The higher the better.

Biswal et al, Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment, AAAI 2020

Setup: Baseline ML methods

- **Median Enrollment (Median):** considers the median enrollment rate for each therapeutic area as estimated rate for all trials in that area.
- **Logistic Regression (LR):** Combine all features and then apply LR.
- **Random Forest (RF):** Combine all features and then apply RF.
- **AdaBoost:** Combine all features and then apply AdaBoost.
- **Multi-layer Perceptron (MLP):** Convert codes to count vectors, convert categorical information of clinical trials to multi-hot vectors and obtain TF-IDF features from text information of clinical trials. Then apply MLP.
- **Long Short-Term Memory Networks (LSTM):** process all temporal data using LSTM and then concatenate with other features.
- **DeepMatch:** Features for the doctors are obtained from the top 50 most frequent medical codes and passed through an MLP layer to obtain an embedding vector.

Biswal et al, Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment, AAAI 2020

Results: Across all trials

Doctor2vec has 8.7% relative improvement in PR-AUC over the best baseline method, LSTM

	PR-AUC	R^2 Score
Median	0.571 ± 0.014	0.54 ± 0.072
LR	0.672 ± 0.041	0.314 ± 0.082
RF	0.731 ± 0.034	0.618 ± 0.034
AdaBoost	0.747 ± 0.002	0.684 ± 0.146
MLP	0.761 ± 0.019	0.762 ± 0.049
LSTM	0.792 ± 0.034	0.780 ± 0.621
DeepMatch	0.735 ± 0.068	0.821 ± 0.073
Doctor2Vec	$\mathbf{0.861 \pm 0.021}$	$\mathbf{0.841 \pm 0.072}$

Biswal et al, Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment, AAAI 2020

Results: Transferring ML model to new country or new disease

Transfer to a less populated or newly explored country

13.7% better PR-AUC than LSTM and 8.1% R2 than DeepMatch.

	PR-AUC	R^2 Score
Median	0.524 ± 0.032	0.420 ± 0.039
LR	0.601 ± 0.023	0.279 ± 0.014
RF	0.661 ± 0.038	0.552 ± 0.048
AdaBoost	0.672 ± 0.01	0.581 ± 0.039
LSTM	0.758 ± 0.013	0.721 ± 0.025
DeepMatch	0.703 ± 0.087	0.756 ± 0.031
Doctor2Vec	$\mathbf{0.862 \pm 0.003}$	$\mathbf{0.817 \pm 0.025}$

Transfer to rare or low prevalence diseases

8.1% better PR-AUC than LSTM and 5.2% R2 than DeepMatch.

	PR-AUC	R^2 Score
Median	0.413 ± 0.013	0.387 ± 0.001
LR	0.521 ± 0.021	0.225 ± 0.028
RF	0.610 ± 0.019	0.517 ± 0.032
AdaBoost	0.623 ± 0.002	0.548 ± 0.046
LSTM	0.725 ± 0.002	0.623 ± 0.038
DeepMatch	0.638 ± 0.021	0.678 ± 0.049
Doctor2Vec	$\mathbf{0.784 \pm 0.032}$	$\mathbf{0.716 \pm 0.014}$

Biswal et al, Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment, AAAI 2020

Part II

STAN: Spatio-temporal attention network for pandemic prediction using real world evidence

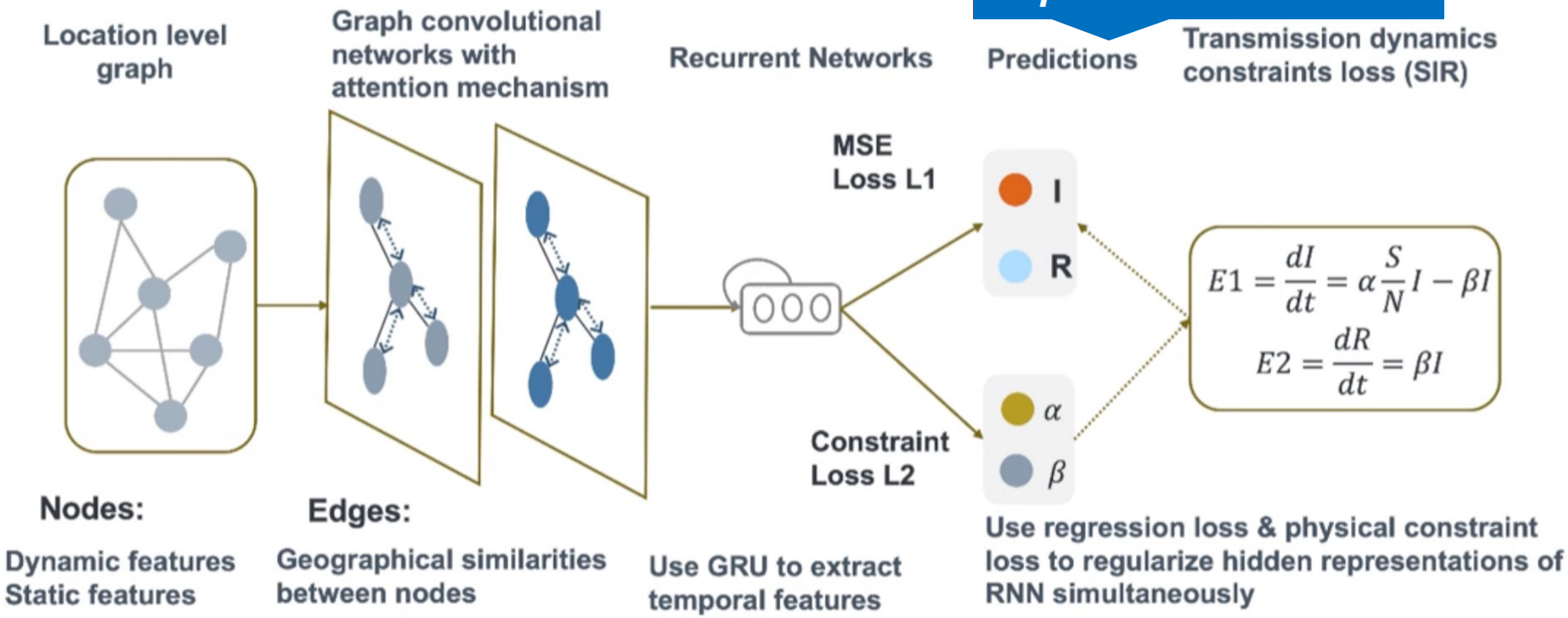
Clinical site selection

- **Task:** Predict locations where there will be a surge in Covid-19 cases in the near future
- Challenges:
  - Data: Diseases can be seasonal, acute and infectious. Longitudinal patient health records do not encode useful patterns to capture these effects
  - Tools: Existing epidemic models (e.g., Susceptible → Infectious → Recovered (SIR)) are designed to evaluate the impact of public health interventions. They do not provide accurate predictions for individual patients

STAN: Spatio-Temporal Attention Network for Pandemic Prediction Using Real World Evidence, JAMIA 2021

STAN: Overview (1/5)

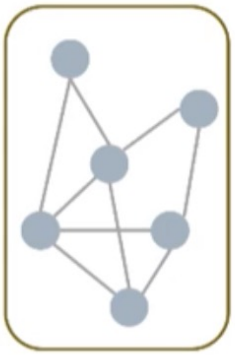
Predictions (outputs):
- I: # of infected cases
- R: # of recovered cases
- α: Transmission rate
- β: Recover rate



STAN: Spatio-Temporal Attention Network for Pandemic Prediction Using Real World Evidence, JAMIA 2021

STAN: Overview (2/5)

Location (e.g., county or state) level graph



Selected Dynamic features: counts of related medical codes →

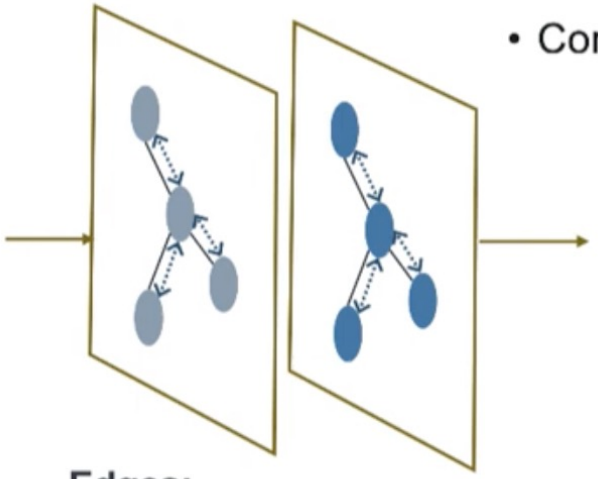
Selected Diagnosis Codes	
ICD-10 Code	**Description**
U071	COVID-19
J09 - J118	Influenza related
J120 – J189	Pneumonia related
R05	Cough
R0602	Shortness of breath
R509	Fever, unspecified
Z20828	Communicable diseases
Selected Procedure Codes	
CPT Code	**Description**
99291 - 99292	Critical care, evaluation and management of the critically ill or critically injured patient
99221 - 99239	Hospital inpatient services

Can incorporate demographics information and intervention policy. Can be used in regions where claims data are not available (reduced model).

STAN: Spatio-Temporal Attention Network for Pandemic Prediction Using Real World Evidence, JAMIA 2021

STAN: Overview (3/5)

Graph convolutional networks with attention mechanism



Edges:

Geographical similarities between nodes, can also incorporate transportation volume.

- Given feature matrix $\boldsymbol{X}$, graph $\boldsymbol{G}$ with N nodes, graph adjacency matrix $\boldsymbol{A}$ and graph degree matrix $\boldsymbol{D}$
- Compute *i*-th node embedding $\widetilde{\boldsymbol{z}}_i^t$ using GCN with K-head attention:

$$\widehat{\boldsymbol{A}} = \boldsymbol{X}^{-\frac{1}{2}}(\boldsymbol{X} + \boldsymbol{I}_N)\boldsymbol{D}^{-\frac{1}{2}}$$

$$\boldsymbol{z}_t = \widehat{\boldsymbol{A}}\mathrm{Relu}(\widehat{\boldsymbol{A}}\boldsymbol{X}_t\boldsymbol{W}_0)\boldsymbol{W}_1$$

$$e_{ij} = a(\boldsymbol{W}_a\boldsymbol{z}_i^t, \boldsymbol{W}_a\boldsymbol{z}_j^t)$$

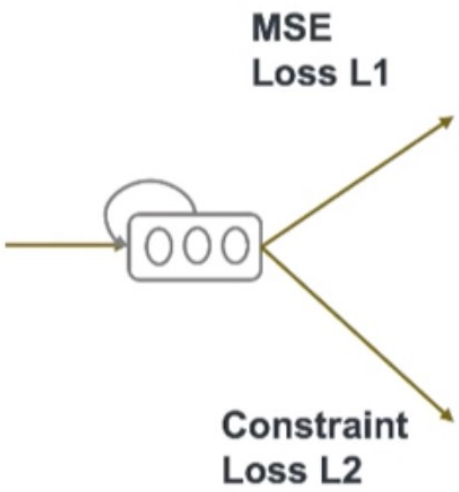
$$a_{ij} = softmax(e_{ij})$$

$$\widetilde{\boldsymbol{z}}_i^t = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{N} a_{ij}^k \boldsymbol{W}^k \boldsymbol{z}_i^t\right)$$

STAN: Spatio-Temporal Attention Network for Pandemic Prediction Using Real World Evidence, JAMIA 2021

STAN: Overview (4/5)

Recurrent Networks



Use RNN to extract temporal features

- For the i-th node embedding, using GRU to extract temporal patterns :

$$\boldsymbol{h}_i = GRU(\widetilde{\boldsymbol{z}_i^1}, \widetilde{\boldsymbol{z}_i^2}, \ldots, \widetilde{\boldsymbol{z}_i^t})$$

- Predict future P-step infected and recovered cases:

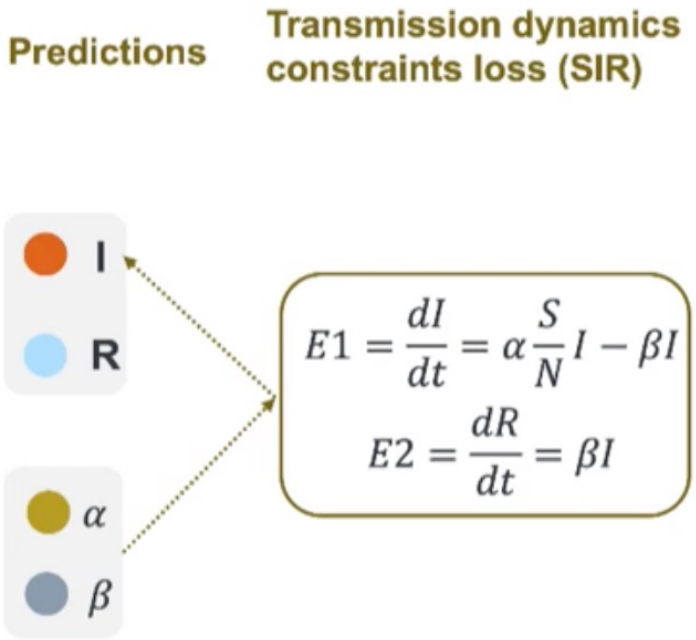
$$\widehat{I_1}, \widehat{R_1}, \ldots, \widehat{I_P}, \widehat{I_P} = MLP(\boldsymbol{h}_i)$$

- Predict future P-step transmission/recover rate:

$$\alpha, \beta = MLP(\boldsymbol{h}_i)$$

STAN: Spatio-Temporal Attention Network for Pandemic Prediction Using Real World Evidence, JAMIA 2021

STAN: Overview (5/5)



Use regression loss & physical constraint loss to regularize hidden representations of RNN simultaneously

- Compute loss function using physical constraints $E1$ and $E2$ across P days.

$$L_{pred} = \sum_{t=1}^{P} MSE(\hat{I}_t, I_t) + MSE(\hat{R}_t, R_t)$$

$$E1 = \frac{dI}{dt} = \alpha \frac{S}{N} I - \beta I$$

$$E2 = \frac{dR}{dt} = \beta I$$

$$L_{phy} = \sum_{t=1}^{P} MSE(\hat{I}_{t-1} + E1, I_t) + MSE(\hat{R}_{t-1} + E2, R_t)$$

$$L = L_{pred} + L_{phy}$$

STAN: Spatio-Temporal Attention Network for Pandemic Prediction Using Real World Evidence, JAMIA 2021

Results: County-level and state-level predictions

Name	Count
# of States	45
# of Counties (Cases > 1000)	193
# ICD codes	48
Start date	03-22-20
End date	06-10-20
Train & Val size	55 days
Test size	5 days

Mean Squared Error (MSE); Mean Absolute Error (MAE) --- lower the better
The Concordance Correlation Coefficient (CCC) --- higher the better

- Predictions on 45 states
- Ours vs. Baselines (Average)

Model	MSE	MAE	CCC
SIR	2,968,711	921.06	0.41
	(814,014-4,152,617)	(776.93-1209.22)	(0.37-0.45)
SEIR	1,890,708	679.64	0.49
	(612,049-3,562,890)	(681.57-1197.38)	(0.44-0.54)
GRU	925,701	582.43	0.55
	(501,309-1,792,855)	(479.50-842.38)	(0.50-0.60)
ColaGNN	601,840	440.26	0.66
	(381,907-982,354)	(323.57-568.44)	(0.59-0.72)
CovidGNN	830,517	500.11	0.58
	(430,127-1,109,311)	(367.55-645.72)	(0.53-0.64)
STAN-PC	323,325	313.72	0.75
	(213,702-450,314)	(280.39-392.01)	(0.70-0.79)
STAN-Graph	472,245	362.04	0.67
	(276,391-612,099)	(310.08-452.39)	(0.63-0.71)
STAN	**237,412**	**220.50**	**0.84**
	(159,995-290,801)	**(172.71-272.03)**	**(0.81-0.87)**

Significantly better than SEIR

- **87%** lower MSE, **55%** lower MAE, **60%** higher CCC

- Conduct predictions on 193 counties
- Ours vs. baselines (Average)

SIR	93,512	151.33	0.40
	(44,864-159,117)	(125.49-177.86)	(0.38-0.44)
SEIR	134,494	165.14	0.35
	(50,223-251,893)	(136.94-194.09)	(0.32-0.38)
GRU	79,982	121.76	0.47
	(39,820-136,096)	(100.96-143.10)	(0.43-0.51)
ColaGNN	61,627	110.91	0.53
	(36,176-104,864)	(91.97-130.36)	(0.49-0.58)
CovidGNN	71,664	120.01	0.47
	(37,718-121,941)	(99.16-140.55)	(0.43-0.51)
STAN-PC	53,194	107.69	0.58
	(32,961-103,211)	(87.63-123.12)	(0.53-0.63)
STAN-Graph	50,331	104.99	0.57
	(29,023-97,304)	(85.09-117.33)	(0.53-0.61)
STAN	**44,177**	**79.80**	**0.66**
	(13,028- 79,916)	**(66.17-93.79)**	**(0.60-0.71)**

Significantly better than SEIR/SIR

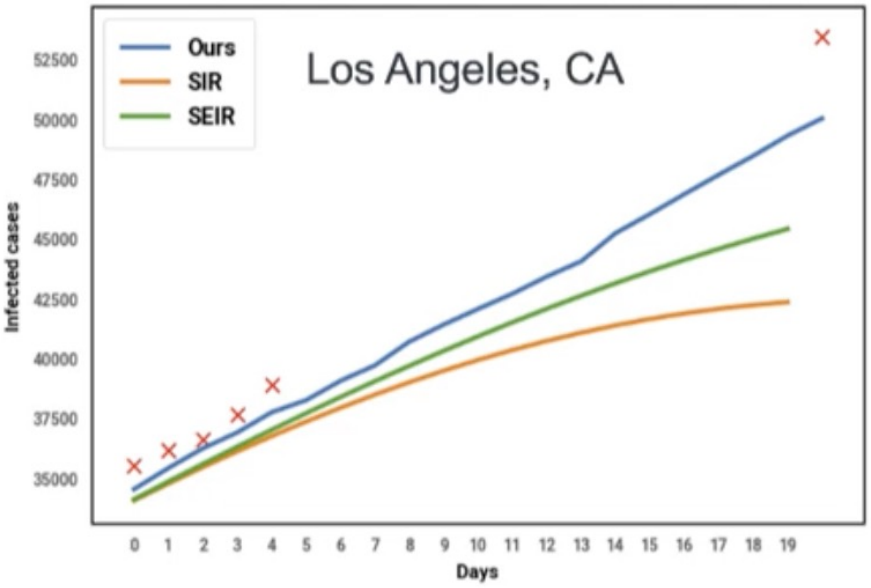
- **61%** lower MSE, **44%** lower MAE
- **57%** higher CCC

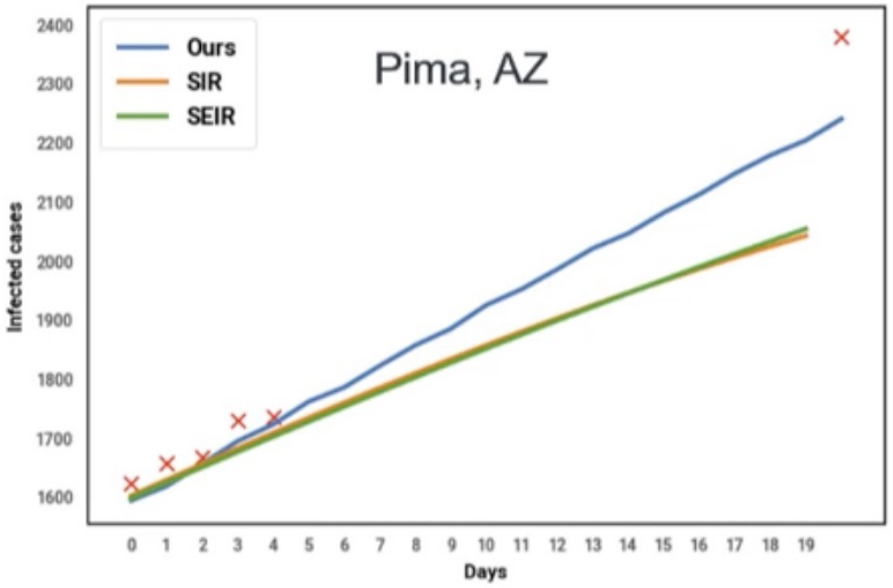
STAN: Spatio-Temporal Attention Network for Pandemic Prediction Using Real World Evidence, JAMIA 2021

Results: County-level predictions

Predictions for the next 20 days (May 16 – June 5)

The model shows less overfitting and performs better than SIR and SEIR (x indicates ground truth)





STAN: Spatio-Temporal Attention Network for Pandemic Prediction Using Real World Evidence, JAMIA 2021

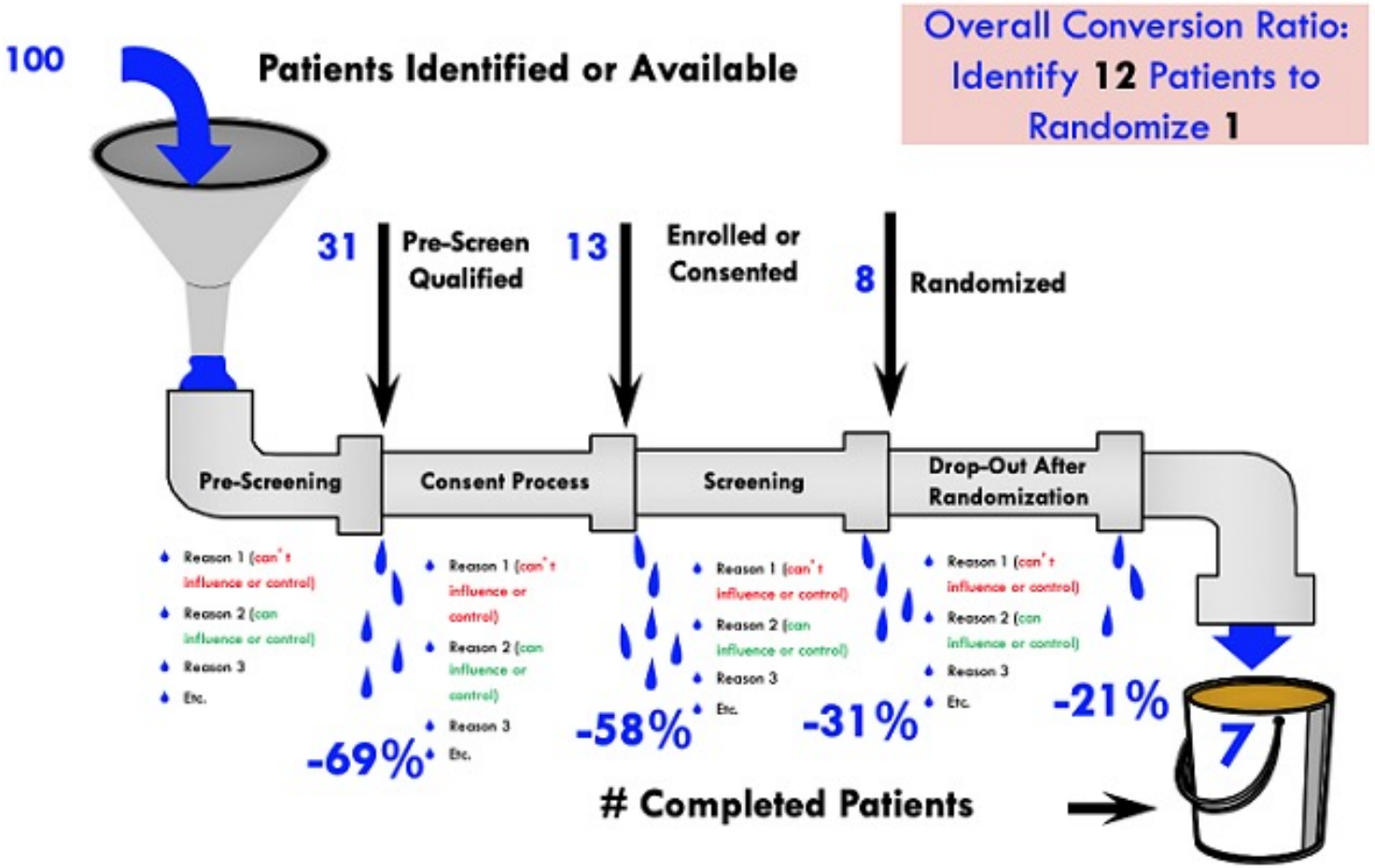
Part III

COMPOSE: Cross-modal Siamese network for patient trial matching

Understanding patient recruitment

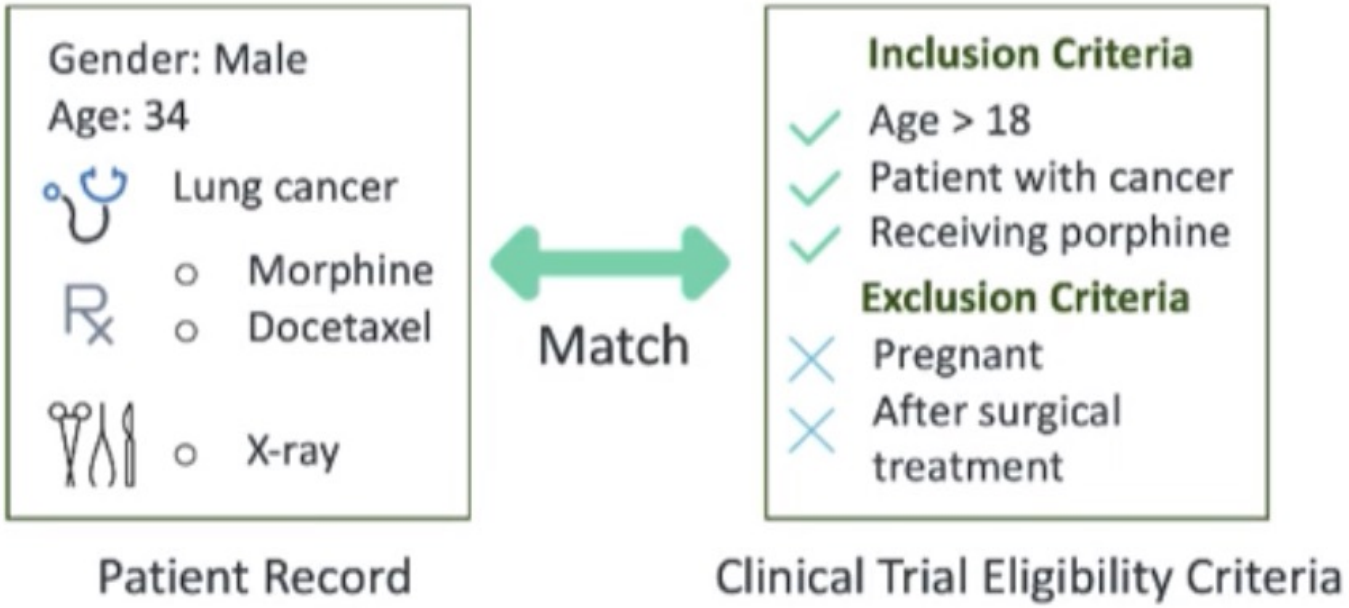
- Nearly 80% of all clinical studies fail to finish on time, and 20% of those delayed are for six months or more
- 85% of clinical trials fail to retain enough patients
- The average dropout rate across all clinical trials is around 30%
- Over two-thirds of sites fail to meet original patient enrollment for a given trial
- Up to 50% of sites enroll one or no patients in their studies

"Leaky pipe" framework for understanding patient recruitment



What is patient-trial matching?

Goal: Find qualified patients for a clinical trial given patient data and trial eligibility criteria (EC) described as both inclusion and exclusion criteria



Patient data can come from longitudinal EHRs or screening or surveys

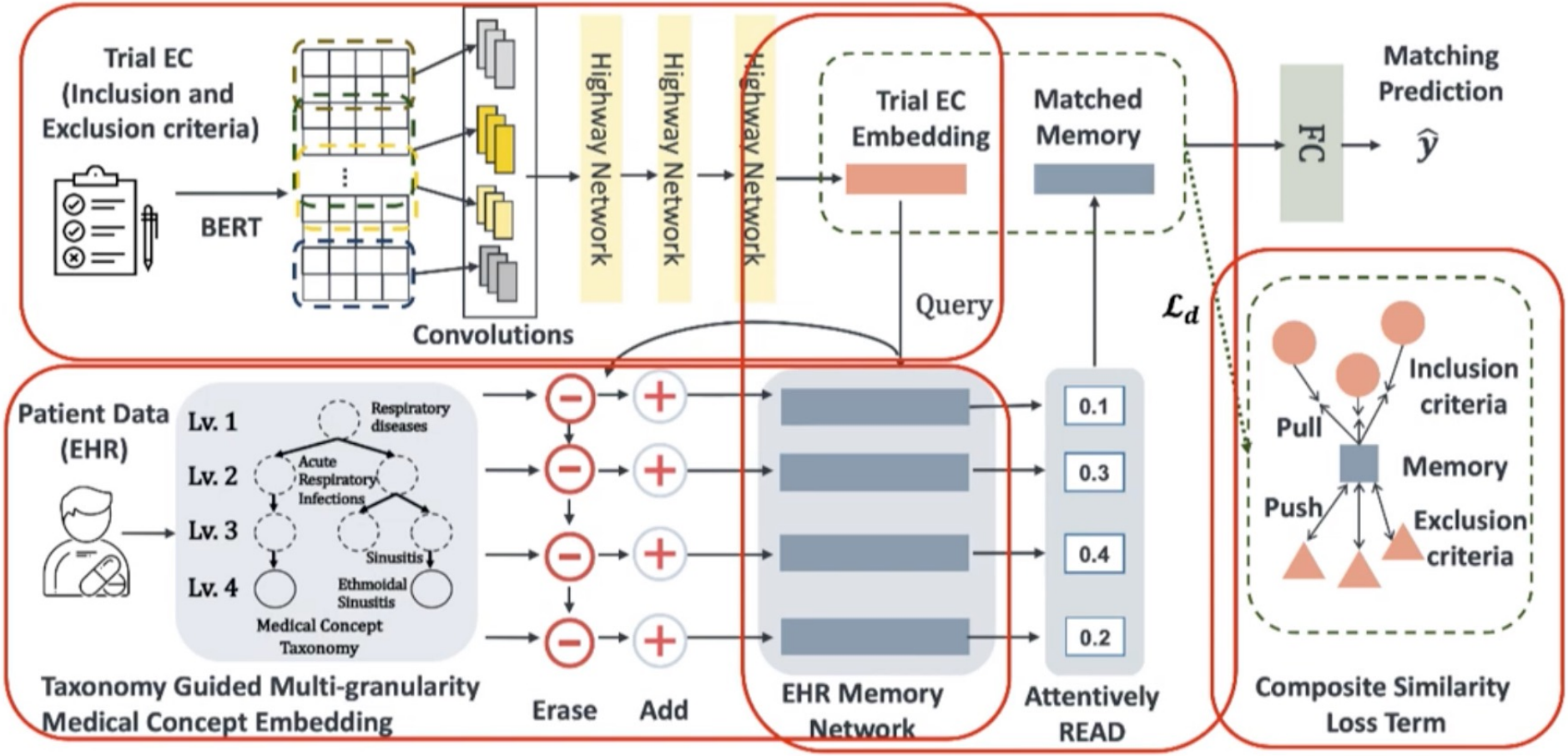
COMPOSE: Cross-Modal Pseudo-Siamese Network for Patient Trial Matching, KDD 2020

Challenges of patient-trial matching

1. **Varying concept granularity**
   - Eligibility criteria encode general diseases
   - EHRs use specific medical codes
2. **Many-to-many matching**
   - Every patient might enroll in more than one trial and vice versa
   - Aligning patient embeddings to multiple trial embeddings can confuse the embedder
3. **Handling explicit inclusion/exclusion criteria**
   - Criteria describe desired and unwanted characteristics of target patients

COMPOSE: Cross-Modal Pseudo-Siamese Network for Patient Trial Matching, KDD 2020

COMPOSE: Method overview (1/6)



COMPOSE: Cross-Modal Pseudo-Siamese Network for Patient Trial Matching, KDD 2020

COMPOSE: Method overview (2/6)

- Use BERT to learn contextual embeddings for EC sentence $[w_1, \ldots, w_N]$

$$\widetilde{c} = [\widetilde{w}_1, \ldots, \widetilde{w}_N] = BERT([w_1, \ldots, w_N])$$

- Use different kernel sizes to capture different granularity semantics

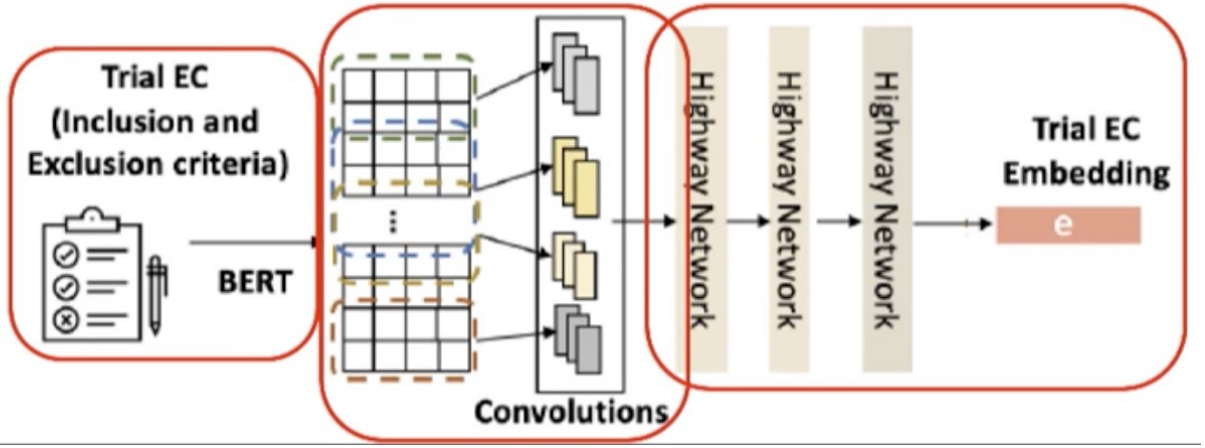
$$x = [Conv(\widetilde{c}, k_1), Conv(\widetilde{c}, k_2), Conv(\widetilde{c}, k_3), Conv(\widetilde{c}, k_4)]$$

- Use highway network and max pooling to obtain the final EC embedidng

$$u = \sigma(Conv(x, k))$$

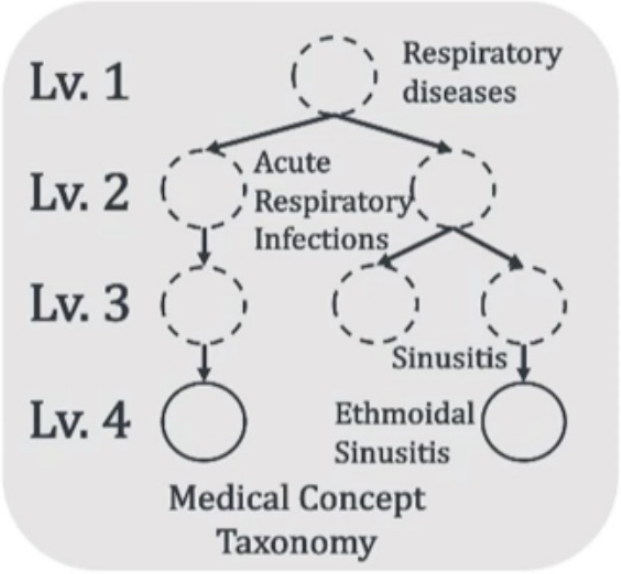
$$v = u \cdot Conv(x, k) + x \cdot (1 - u)$$

$$e = MaxPool(v)$$

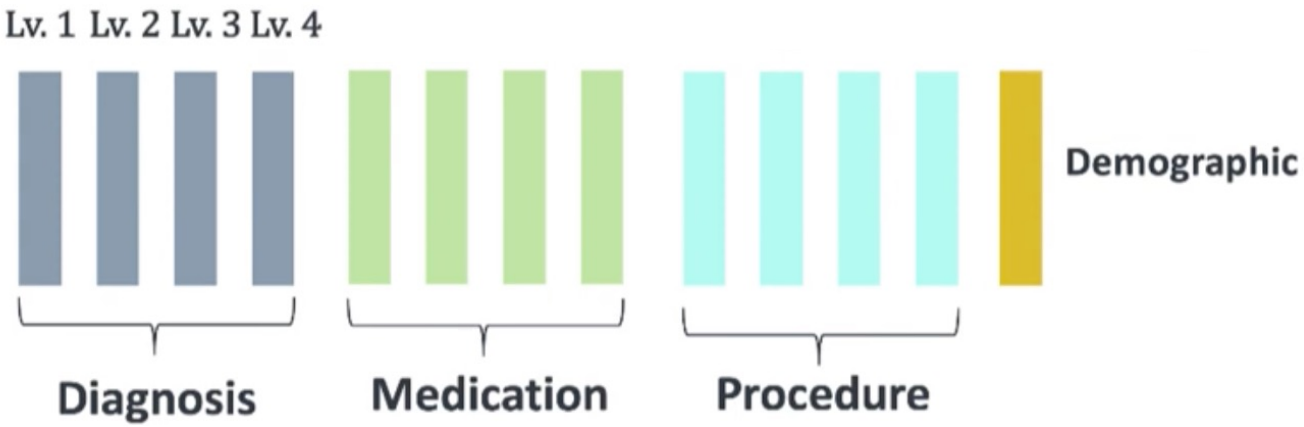


Taxonomy guided patient embeddings (3/6)

- Use medical concept taxonomy to divide each concept into four levels
  - the Uniform System of Classification (USC)
- Three memory networks to store diagnosis, medications and procedures



$$\mathbf{m} = [\mathbf{m}_{\mathcal{D}}, \mathbf{m}_{\mathcal{O}}, \mathbf{m}_{\mathcal{P}}] = [\mathbf{m}^1_{\mathcal{D}}, \ldots, \mathbf{m}^4_{\mathcal{D}}, \mathbf{m}^1_{\mathcal{O}}, \ldots, \mathbf{m}^4_{\mathcal{O}}, \mathbf{m}^1_{\mathcal{P}}, \ldots, \mathbf{m}^4_{\mathcal{P}}]$$



COMPOSE: Cross-Modal Pseudo-Siamese Network for Patient Trial Matching, KDD 2020

Taxonomy guided patient embeddings (4/6)

- Augment medical codes with textual description:
  - Code 692.9 -> "Contact dermatitis and other eczema"

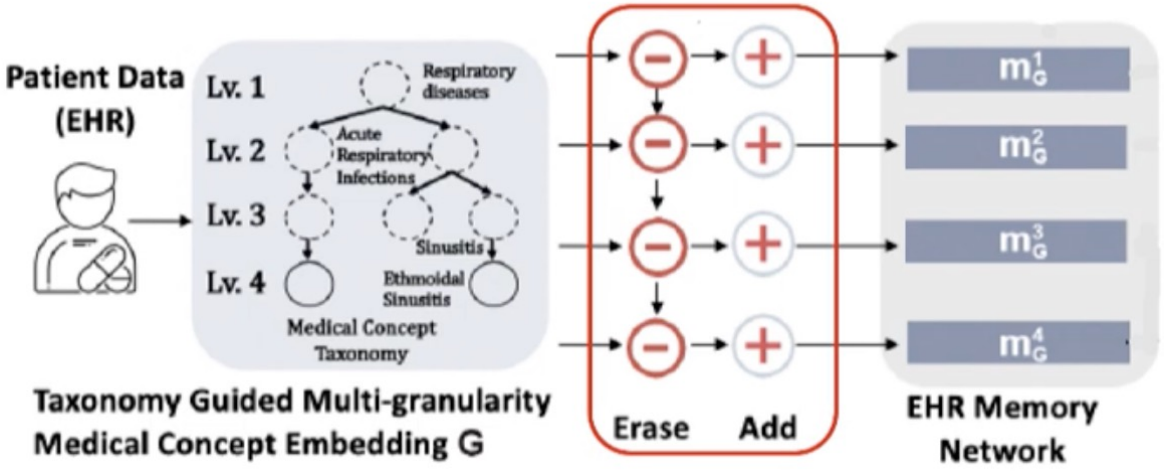
$$\tilde{g}_t = MaxPool(BERT([w_1, ..., w_L]))$$

- Update memories at each visit
  - Erase-followed-by-add:

$$\mathbf{erase}_t = \sigma(\mathbf{W}_e \tilde{g}_t^k + \mathbf{b}_e),$$

$$\mathbf{add}_t = tanh(\mathbf{W}_a \tilde{g}_t^k + \mathbf{b}_a)$$

$$m_G^k \leftarrow m_G^k \odot (1 - \mathbf{erase}_t) + \mathbf{add}_t$$

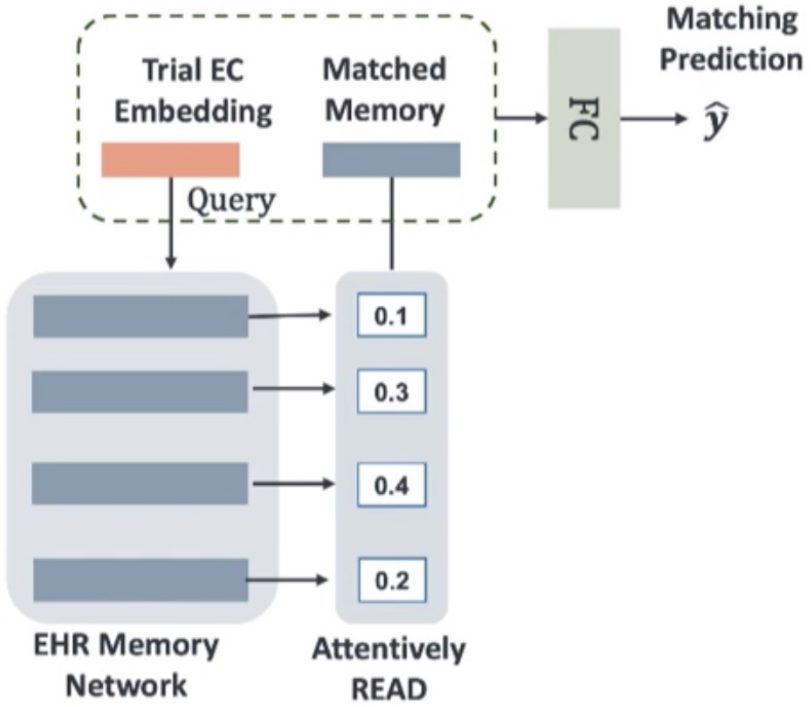


COMPOSE: Method overview (5/6)

- Let each EC correspond to the sub-memories
- Attentional matching
  - Trial EC embedding -> Query
  - Matched memory -> Response

$$a_{k,G} = \frac{exp(\boldsymbol{m}_G^{k\,\mathrm{T}} MLP(\boldsymbol{e}))}{\sum_{x\in\{\mathcal{D},\mathcal{O},\mathcal{P}\}} \sum_{i=1}^{4} exp(\boldsymbol{m}_x^{i\,\mathrm{T}} MLP(\boldsymbol{e}))}$$

$$\widetilde{\boldsymbol{m}} = \sum_{x\in\{\mathcal{D},\mathcal{O},\mathcal{P}\}} \sum_{i=1}^{4} a_{i,x} \boldsymbol{m}_x^i$$



COMPOSE: Cross-Modal Pseudo-Siamese Network for Patient Trial Matching, KDD 2020

COMPOSE: Method overview (6/6)

- Classification loss:

$$\mathcal{L}_c = -(\boldsymbol{y}^{\mathrm{T}} log(\hat{\boldsymbol{y}}) + (1-\boldsymbol{y})^{\mathrm{T}} log(1-\hat{\boldsymbol{y}}))$$

- Inclusion/Exclusion loss:

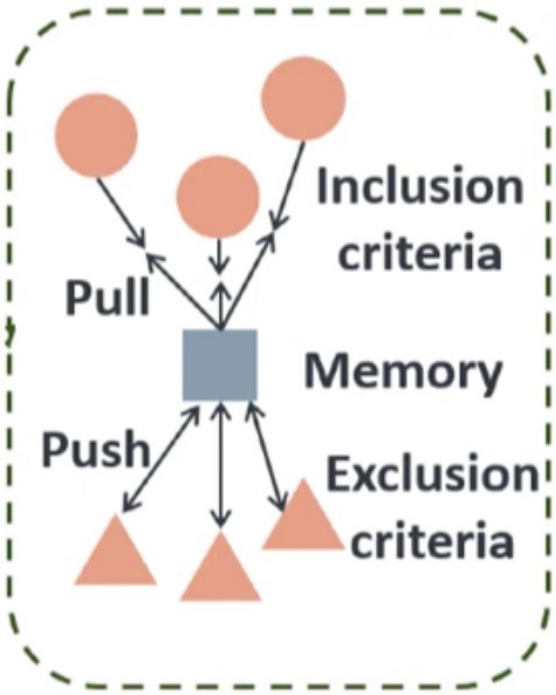
$$\mathcal{L}_d = \begin{cases} 1 - d(\boldsymbol{e}, \widetilde{\boldsymbol{m}}_I)), & if\ \boldsymbol{e}\ is\ \boldsymbol{e}_I \\ max(0, d(\boldsymbol{e}, \widetilde{\boldsymbol{m}}_E) - \alpha), & if\ \boldsymbol{e}\ is\ \boldsymbol{e}_E \end{cases}$$

-> 0

>= α

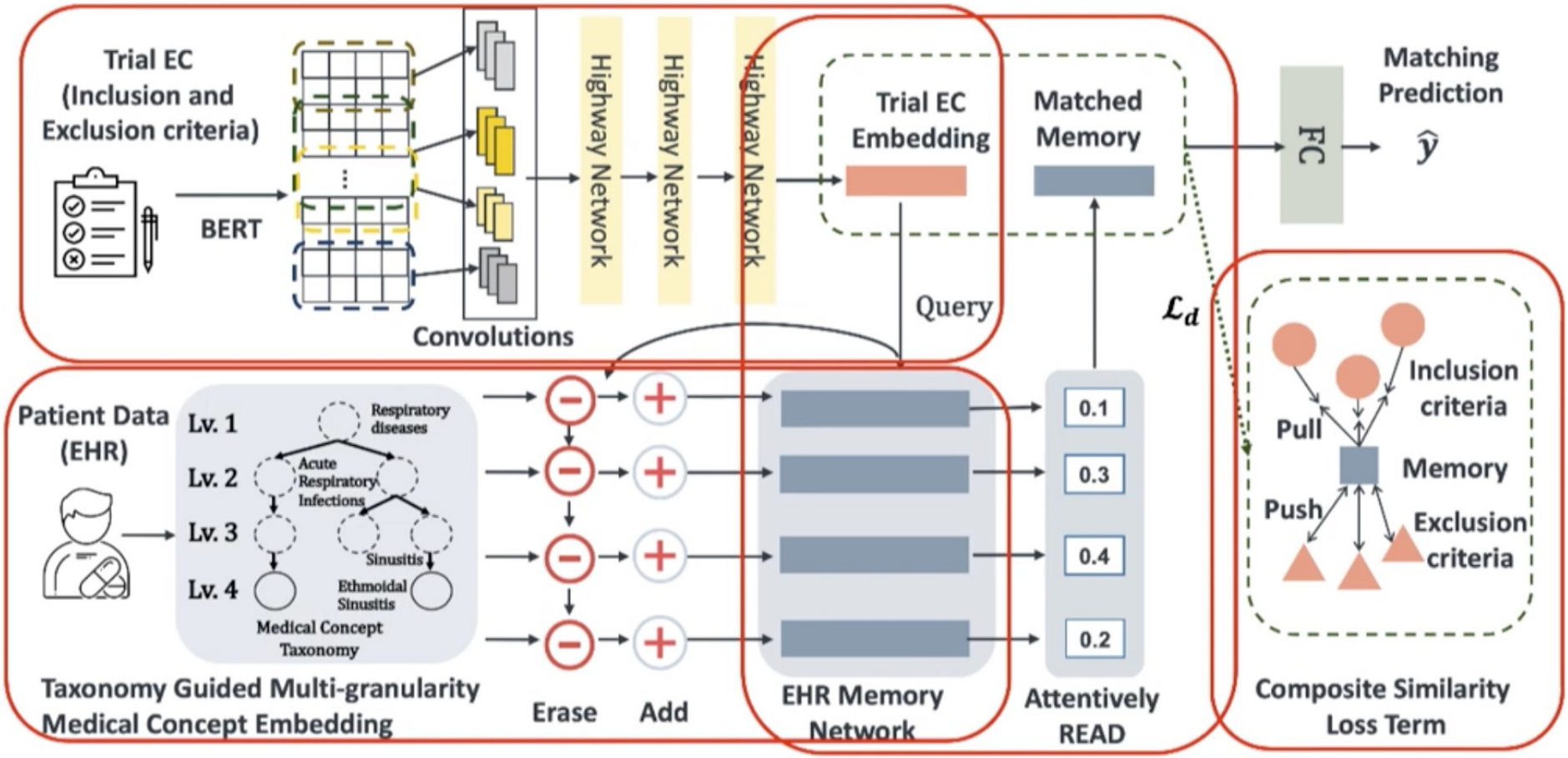
- Final loss:

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_d$$



Composite Similarity Loss Term

COMPOSE: Patient-trial matching



COMPOSE: Cross-Modal Pseudo-Siamese Network for Patient Trial Matching, KDD 2020

Experimental setup: Data

- **Clinical trials:**
  - 590 trials from publicly available data source (clinicaltrials.gov)
  - 12,445 criteria-level EC statements
- **Patient EHR dataset:**
  - 83,731 patients from 2012 to 2018

Results: Criteria-level matching

	Model	Accuracy	AUROC	AUPRC
Baselines	LSTM+GloVe	0.722±0.010	0.789±0.009	0.784±0.009
	LSTM+BERT	0.834±0.008	0.845±0.007	0.840±0.007
	DeepEnroll	0.869±0.012	0.936±0.013	0.947±0.011
Reduced	COMPOSE-MN	0.899±0.012	0.955±0.013	0.960±0.010
	COMPOSE-Highway	0.912±0.007	0.965±0.007	0.967±0.009
	COMPOSE-$\mathcal{L}_d$	0.939±0.010	0.976±0.009	0.973±0.007
Proposed	COMPOSE	**0.945±0.008**	**0.980±0.007**	**0.979±0.008**

Model	Phase I	Phase II	Phase III
LSTM+GloVe	0.0008	0.5865	0.3743
LSTM+BERT	0.0025	0.6045	0.4862
Criteria2Query	0.3025	0.6433	0.5870
DeepEnroll	0.2034	0.7493	0.6329
COMPOSE	**0.5189**	**0.8939**	**0.8005**

COMPOSE: Cross-Modal Pseudo-Siamese Network for Patient Trial Matching, KDD 2020

Results: Trial-level matching

	Model	Accuracy
Baselines	LSTM+GloVe	0.4294±0.010
	LSTM+BERT	0.5460±0.008
	Criteria2Query	0.6147±-
	DeepEnroll	0.6737±0.021
Reduced	COMPOSE-MN	0.7833±0.011
	COMPOSE-Highway	0.8102±0.009
	COMPOSE-$\mathcal{L}_d$	0.8212±0.010
Proposed	COMPOSE	**0.8373±0.012**

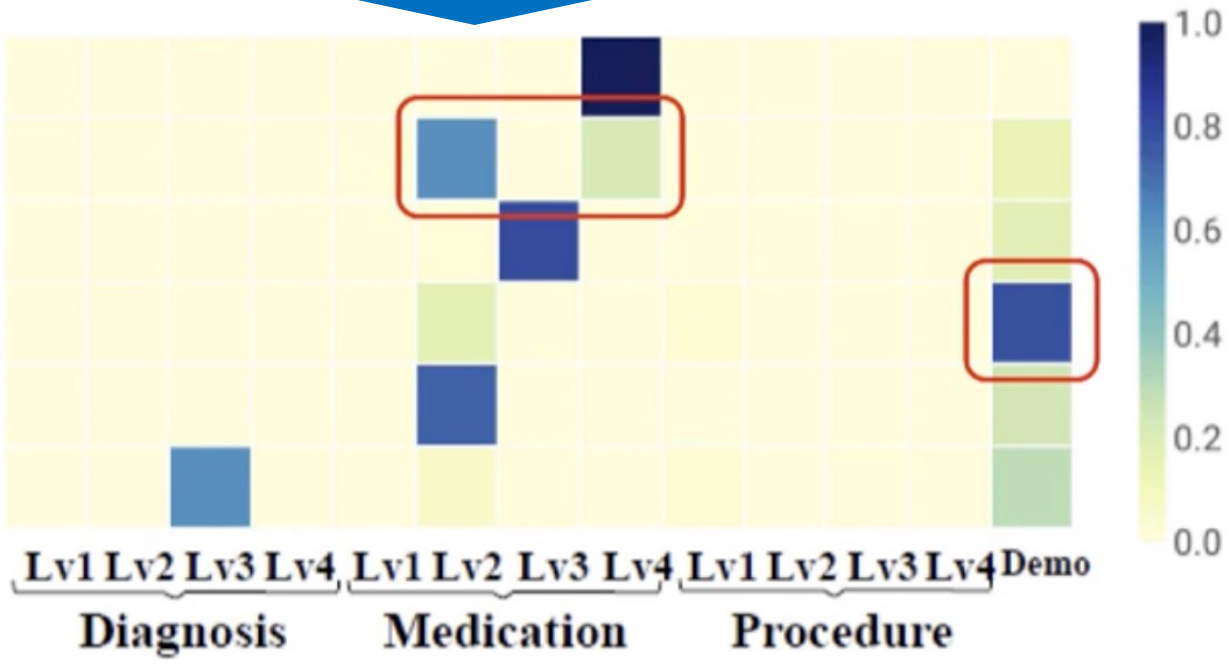
Model	Chronic Diseases	Oncology	Rare Diseases
LSTM+GloVe	0.1793	0.0000	0.0000
LSTM+BERT	0.2062	0.0000	0.0000
Criteria2Query	0.5103	0.2722	0.2292
DeepEnroll	0.3345	0.0000	0.0000
COMPOSE	**0.5931**	**0.6370**	**0.6875**

COMPOSE: Cross-Modal Pseudo-Siamese Network for Patient Trial Matching, KDD 2020

Trial on Cabozantinib, which treats grade IV astrocytic tumors

Attention weights on the memory slots for the Cabozantinib trial for treating grade IV astrocytic tumors

1. received temozolomide therapy
2. receiving warfarin (or other coumarin derivatives)
3. acute intracranial/ intratumoral hemorrhage.
4. pregnant or breast-feeding
5. serious intercurrent illness
6. inherited bleeding diathesis or coagulopathy

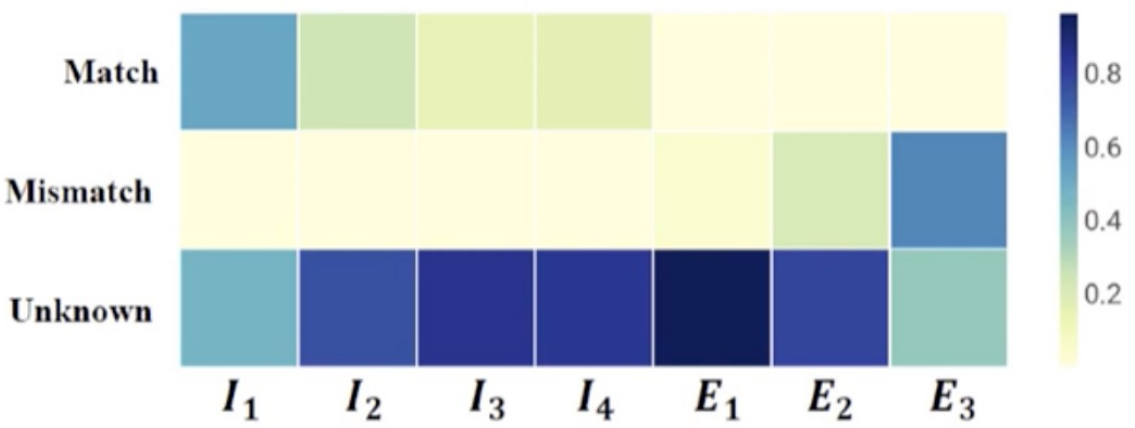


COMPOSE successfully matches this trial (94% matching) while all baselines fail (< 50% matching)

Trial for early-stage non-small cell lung cancer

#NCT02998528

- I_1: Early stage IB-IIIA, operable non-small cell lung cancer, confirmed in tissue
- I_2: Lung function capacity capable of tolerating the proposed lung surgery
- I_3: Eastern Cooperative Oncology Group (ECOG) Performance Status of 0-1
- I_4: Available tissue of primary lung tumor
- E_1: Presence of locally advanced, inoperable or metastatic disease
- E_2: Participants with active, known or suspected autoimmune disease
- E_3: Prior treatment with any drug that targets T cell co-stimulations pathways (such as checkpoint inhibitors)



An example of a trial for which it is difficult to find matching patients. All models achieve a lower than 50% accuracy score for this trial. Shown are prediction results for COMPOSE and a case patient. The results show that COMPOSE successfully matches I_1 and E_3 to the patient but classifies other ECs to unknown

Inclusion criteria are denoted as I_i and exclusion criteria as E_j

Quick Check

https://forms.gle/GCZzrgXjwuEyoJPK8

BMI 702: Biomedical Artificial Intelligence

Foundations of Biomedical Informatics II, Spring 2023

Quick check quiz for lecture 9: Clinical trial site identification, patient trial matching, clinical trial recruitment.

Course website and slides: **https://zitniklab.hms.harvard.edu/BMI702**

Sign in to Google to save your progress. Learn more

* Required

First and last name *

Your answer

Harvard email address *

Your answer

Go to https://clinicaltrials.gov and find a completed clinical trial. Describe the trial eligibility criteria (EC), including inclusion and exclusion criteria. *

Your answer

Submit Clear form

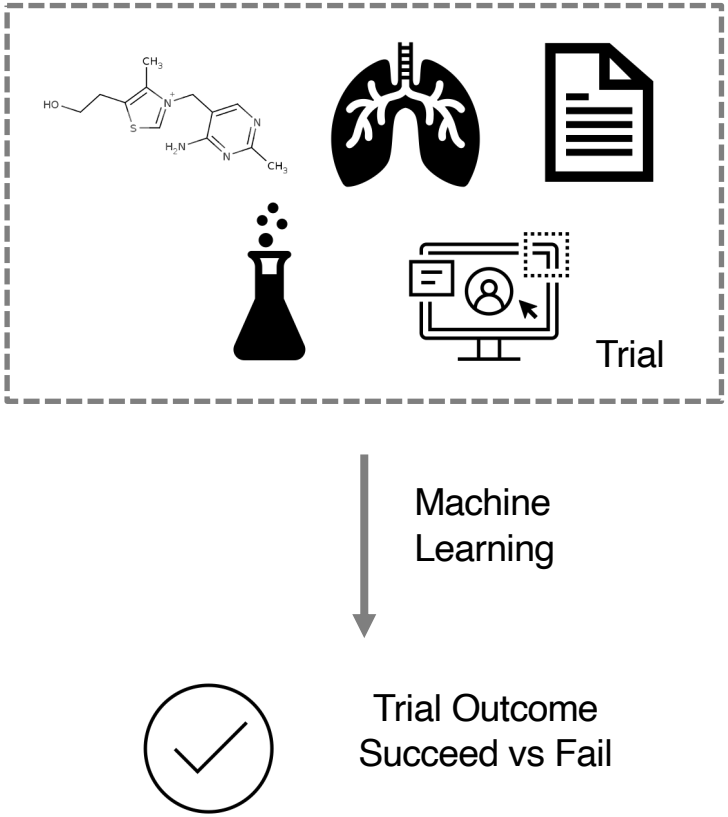
Part IV

HINT: Hierarchical interaction network for clinical trial-outcome predictions

Can we predict the trial outcome before the trial starts?

- Save patient time
- Avoid skyrocketing cost
- Better resource allocation

- **Trial outcome:** binary success indicator whether the trial is completed to meet their primary endpoints



Existing work

- Over the years, there have been attempts to predict individual components in clinical trials to improve trial results:
  - Using EEG measurements to predict the effect of antidepressant treatments in improving depressive symptoms
  - Optimizing drug toxicity based on drug- and target-property features
  - Leveraging phase II results to predict phase III trial results
  - Predicting drug approvals for disease areas based on drug and clinical trial features using ML methods

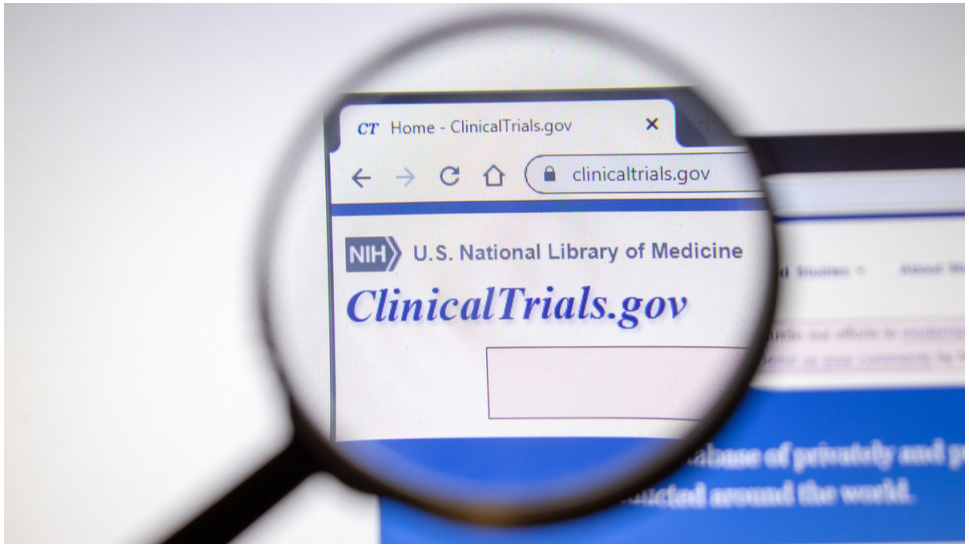
What are challenges for using ML methods?

- **Lack of benchmark data:** Data science progress in any domain needs to be measured on large and accessible benchmark data. Such datasets in clinical trial domains are not available, which severely affects data science efforts on clinical-trial-related research
- **Limited task definition and study scope:** Existing work either focuses on predicting individual components of trials, such as patient-trial matching or only a subset of disease groups
- **Limited features used for prediction:** Using only restricted-disease-specific features is not sufficient; trial outcome is determined by various factors, including drug safety, treatment efficiency, and eligibility criteria

But there are no data…

- ClinicalTrials.gov is comprehensive but messy
- Trial outcome label is usually proprietary

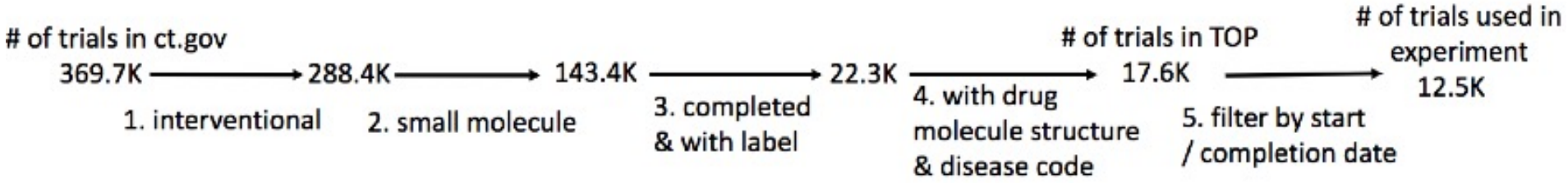
Barrier to creating ML approaches

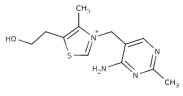


HINT: Hierarchical Interaction Network for Clinical Trial Outcome Prediction, Patterns 2022

Introducing TOP: Clinical Trial Outcome Prediction Benchmark

Workflow of data curation





SMILES (Structure)

Eligibility Criteria

ICD-10, Description

Trial Outcome From IQVIA

Trial outcome prediction: Problem formulation (1/4)

- A trial is designed to evaluate safety or efficacy of a treatment set toward a target disease set on a patient group defined by the trial eligibility criteria
- Treatment set: Treatment set includes one or multiple drug candidates, denoted by:

$$M = \{m_1, \ldots m_{N_m}\}$$

  - m_i is a drug molecule involved in a trial
  - focus is on trials that aim at discovering new drug indications (i.e., not new surgery techniques and medical devices)

Trial outcome prediction: Problem formulation (2/4)

- A trial is designed to evaluate safety or efficacy of a treatment set toward a target disease set on a patient group defined by the trial eligibility criteria
- Target disease set: Each trial targets one or more diseases. Suppose there are N_d diseases in a trial, we represent the target disease set as:

$$D = \{d_1, \ldots d_{N_d}\}$$

  - d_i is a target disease represent raw information associated with the disease including disease name, description (text data), and the corresponding diagnosis code (e.g., International Classification of Diseases [ICD])

Trial outcome prediction: Problem formulation (3/4)

- A trial is designed to evaluate safety or efficacy of a treatment set toward a target disease set on a patient group defined by the trial eligibility criteria
- Trial eligibility criteria: Patient group is specified by the trial eligibility criteria. Formally, eligibility criteria consist of a set of inclusion and exclusion criteria for recruiting patients represented as:

$$C = \{c_1^I, c_2^I, \dots c_M^I, c_1^E, c_2^E \dots c_N^E\}$$

  - c_i^I is i-th inclusion criterion, c_i^E is i-th exclusion criterion
  - Every criterion is a sentence given in unstructured natural language

Trial outcome prediction: Problem formulation (4/4)

- A trial is designed to evaluate safety or efficacy of a treatment set toward a target disease set on a patient group defined by the trial eligibility criteria
- Trial outcome: Trial outcome is a binary label $y \in \{0,1\}$, where $y = 1$ indicates the trial met their primary endpoints, while $y = 0$ means the trial failed to meet primary endpoints
  - Primary endpoints are the statistical measures to indicate whether the drug candidate works or not
  - For example, for an antihypertensive drug trial, primary endpoint can be the percentage of patients with well controlled blood pressure, e.g., systolic BP<140 mm Hg

Trial outcome prediction: Goal

- Our goal is to learn a deep neural network model f_θ for predicting trial success status $\hat{y}$ as:

$$\hat{y} = f_\theta(M, D, C)$$

  - M – treatment set
  - D – target disease set
  - C – eligibility criteria
- In general, there are three trial phases:
  - Phase I tests the toxicity and side effects of the drug
  - Phase II determines efficacy of the drug
  - Phase III focuses on the effectiveness of the drug (i.e., whether the drug is better than the current standard practice)
- Phase-level prediction determines whether a specific study successfully completes in a given phase

Trial outcome prediction: Statistics

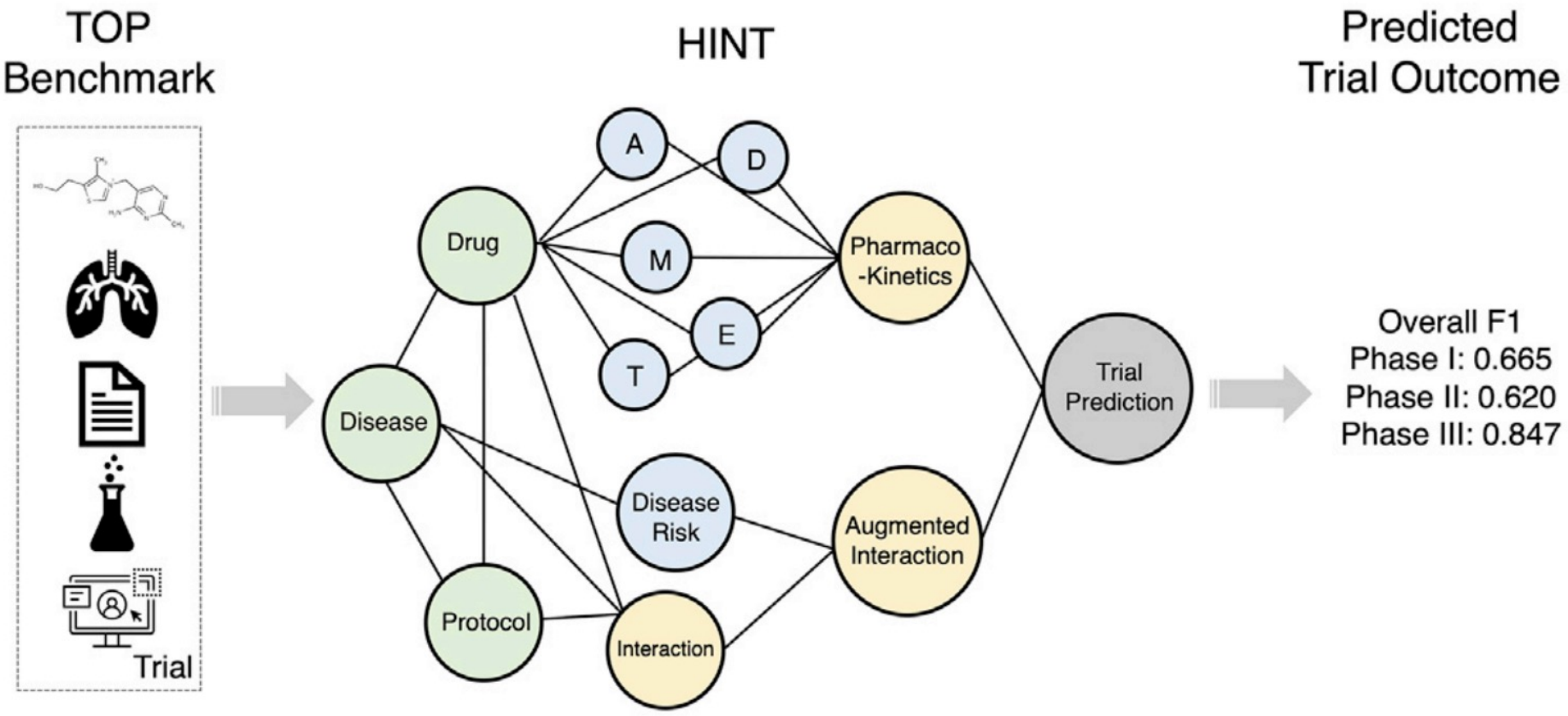
	# trials	# drugs	# diseases	# successes	# failures
All	17,538	13,880	5,335	9,999	7,539
All (filtered by start/completion date)	12,465	10,026	3,893	7,149	5,316
Neoplasm	4,246	2,456	2,008	1,752	2,494
Respiratory system	1,299	1,736	968	868	431
Digestive system	1,844	1,990	1,558	1,072	772
Nervous system	1,975	2037	1,369	1,171	804
Others	8,174	9,778	4,090	5,136	3,038
start before 2000	179	144	193	43	136
start between 2000-2004	1,753	1,092	317	771	982
start between 2005-2009	6,211	2,358	1,267	3,472	2,739
start between 2010-2014	6,846	3,277	1,613	4,185	2,661
start between 2015-2021	2,549	2,987	1,710	1,528	1,021
Phase I	1,787	2,020	1,392	582/77/347	462/39/280
Phase II	6,102	5,610	2,824	1,925/196/918	2,079/249/735
Phase III	4,576	4,727	1,619	2,042/208/854	1,050/136/286

All clinical trial records available at ClinicalTrials.gov on February 20, 2021

For phases I-III, we show the #train/#validation/#test for successes and failures

Train, validation and test are time-split according to January 1, 2014, i.e., start dates of trials in the test set are after January 1, 2014, while completion dates of trials in train and validation set are before January 1, 2014

Predicting trial outcome with hierarchical attention network



HINT: Hierarchical Interaction Network for Clinical Trial Outcome Prediction, Patterns 2022

Overview of HINT

- Deep learning models have shown many successes in modeling biomedical data
- Existing models handle one type of data, while real-world data science applications often have multi-modal datasets with various characteristics and qualities along with domain-specific knowledge.
- Hierarchical Interaction Network (HINT):
  - Graph neural network
  - It can handle complex interaction patterns from multi-modal data for clinical-trial-outcome prediction:
    - Different types of input data (e.g., graphs, text, and categorical variables)
    - Missing values

Results: Phase I outcome prediction

Phase I Trials.

train: 1,044; # valid: 116; # test: 627; # patients/trial: 45.

Method	PR-AUC	F1	ROC-AUC
LR	0.500±0.005	0.604±0.005	0.520±0.006
RF	0.518±0.005	0.621±0.005	0.525±0.006
XGBoost	0.513±0.06	0.621±0.007	0.518±0.006
AdaBoost	0.519±0.005	0.622±0.007	0.526±0.006
kNN+RF[10]	0.531±0.006	0.625±0.007	0.538±0.005
FFNN[15]	0.547±0.010	0.634±0.015	0.550±0.010
DeepEnroll[26]	**0.568±0.007**	0.648±0.011	0.575±0.013
COMPOSE[16]	0.564±0.007	0.658±0.009	0.571±0.011
HINT	0.567±0.010	**0.665±0.010**	**0.576±0.008**

HINT: Hierarchical Interaction Network for Clinical Trial Outcome Prediction, Patterns 2022

Results: Phase II outcome prediction

Phase II Trials # train: 4,004; # valid: 445; # test: 1,653; # patients/trial: 183.			
Method	PR-AUC	F1	ROC-AUC
LR	0.565±0.005	0.555±0.006	0.587±0.009
RF	0.578±0.008	0.563±0.009	0.588±0.009
XGBoost	0.586±0.006	0.570±0.009	0.600±0.007
AdaBoost	0.586±0.009	0.583±0.008	0.603±0.007
kNN+RF[10]	0.594±0.008	0.590±0.006	0.597±0.008
FFNN[15]	0.604±0.010	0.599±0.012	0.611±0.011
DeepEnroll[26]	0.600±0.010	0.598±0.007	0.625±0.008
COMPOSE[16]	0.604±0.007	0.597±0.006	0.628±0.009
HINT	**0.629±0.009***	**0.620±0.008***	**0.645±0.006**

HINT: Hierarchical Interaction Network for Clinical Trial Outcome Prediction, Patterns 2022

Results: Phase III outcome prediction

Phase III Trials # train: 3,092; # valid: 344; # test: 1,140; # patients/trial: 1418.			
Method	PR-AUC	F1	ROC-AUC
LR	0.687±0.005	0.698±0.005	0.650±0.007
RF	0.692±0.004	0.686±0.010	0.663±0.007
XGBoost	0.697±0.007	0.696±0.005	0.667±0.005
AdaBoost	0.701±0.005	0.695±0.005	0.670±0.004
kNN+RF[10]	0.707±0.007	0.698±0.008	0.678±0.010
FFNN[15]	0.747±0.011	0.748±0.009	0.681±0.008
DeepEnroll[26]	0.777±0.008	0.786±0.007	0.699±0.008
COMPOSE[16]	0.782±0.008	0.792±0.007	0.700±0.007
HINT	**0.811±0.007***	**0.847±0.009***	**0.723±0.006***

HINT: Hierarchical Interaction Network for Clinical Trial Outcome Prediction, Patterns 2022

Using HINT on recently completed trials: Case studies

Indication/Disease	Drug	Sponsor	Year	Outcome	Prediction
Heart Failure	Entresto	Novartis	2019	fail	0.476
Asthma	Fevipiprant	Novartis	2019	fail	0.352
Lung cancer	Pembrolizumab & Epacadostat	Incyte	2020	fail	0.498
Lupus Erythematosus	Ustekinumab	Janssen	2019	fail	0.567
Diabetes	Sitagliptin	Merck	2017	success	0.742
Rheumatoid Arthritis	Etanercept	Amgen	2019	success	0.673
Neovascular Glaucoma	Aflibercept	Bayer	2020	success	0.854
Depression	Naltrexone	U. Pitts.	2020	success	0.747
Liver cancer	cTACE Doxorubicin	Yale U.	2020	success	0.583
x-linked hypophosphatemia	phosphate supplement and vitamin d	Ultragenyx	2020	success	0.556

HINT: Hierarchical interaction network for clinical trial-outcome predictions, Patterns 2022

Using HINT on recently completed trials: Failed trials

- One of most promising drugs in 2019 was Entresto for heart failure
  - Entresto was sponsored by Novartis and expected to have a 5-billion-dollar peak sale.
  - However, in a multi-country phase III trials with 4,822 patients enrolled, results did not reduce death or meet any other endpoints
  - Trial took 5 years (2014–2019) and was estimated to cost $200 million dollars (we use the median per-patient cost multiplied by the number of patients to estimate the cost)
  - We feed the drug (Entresto), disease (heart failure), and their phase III eligibility criteria into HINT, and it predicts a low success probability of 0.476
    - HINT could have alerted scietists of a likely failure

HINT: Hierarchical interaction network for clinical trial-outcome predictions, Patterns 2022

Using HINT on recently completed trials: Failed trials

- Fevipiprant was expected to be Novartis' blockbuster drug for asthma:
  - Phase III trial of Fevipiprant took 4 years (2015–2019) and enrolled 894 patients
  - It incurred huge costs (an estimated 40 million dollars)
  - Unfortunately, the primary endpoint was not met, and Fevipiprant was retired
  - We feed the drug (Fevipiprant), disease (asthma), and eligibility criteria into HINT, and it predicts a low success probability 0.352
    - HINT could have alerted scientists of a likely failure

HINT: Hierarchical interaction network for clinical trial-outcome predictions, Patterns 2022

Case studies: Successful trials

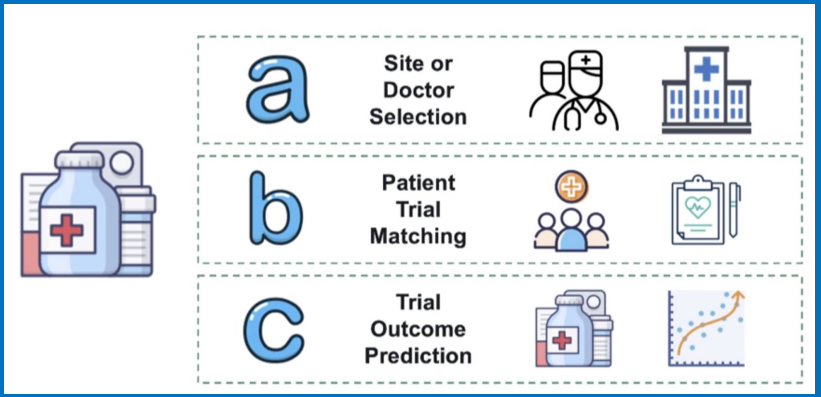
- Sitagliptin for diabetes by Merck in 2017 received a 0.742 success probability
- Etanercept for rheumatoid arthritis by Amgen in 2019 received a 0.673 success probability
- Afibercept for glaucoma by Bayer in 2020 acquired a 0.854 success probability
- Naltrexone for depression by University of Pittsburgh in 2020 received a 0.747 success probability

Indication/Disease	Drug	Sponsor	Year	Outcome	Prediction
Heart Failure	Entresto	Novartis	2019	fail	0.476
Asthma	Fevipiprant	Novartis	2019	fail	0.352
Lung cancer	Pembrolizumab & Epacadostat	Incyte	2020	fail	0.498
Lupus Erythematosus	Ustekinumab	Janssen	2019	fail	0.567
Diabetes	Sitagliptin	Merck	2017	success	0.742
Rheumatoid Arthritis	Etanercept	Amgen	2019	success	0.673
Neovascular Glaucoma	Aflibercept	Bayer	2020	success	0.854
Depression	Naltrexone	U. Pitts.	2020	success	0.747
Liver cancer	cTACE Doxorubicin	Yale U.	2020	success	0.583
x-linked hypophosphatemia	phosphate supplement and vitamin d	Ultragenyx	2020	success	0.556

HINT: Hierarchical interaction network for clinical trial-outcome predictions, Patterns 2022

Outline for today's class

a. Clinical trial recruitment
   - Doctor selection
   - Site selection



b. Patient-trial matching

c. Trial outcome prediction