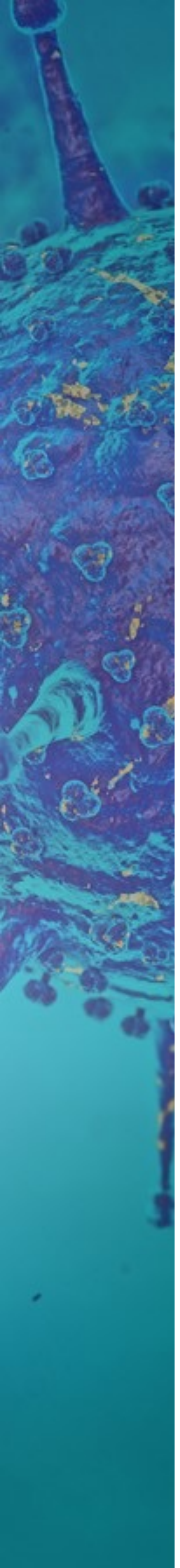
Application of SparkNLP for Development of Multi-Modal Prediction Model from EHR



Sutanay Choudhury

Advanced Computing Mathematics and Data Division

March 29, 2021

Khushbu Agarwal, Colby Ham (PNNL), Pritam Mukherjee, Siyi Tang, Suzanne Tamang (Stanford University), Sindhu Tipirneni, Chandan Reddy (Virginia Tech.), Veysel Kocaman (John Snow Labs)

PNNL is operated by Battelle for the U.S. Department of Energy

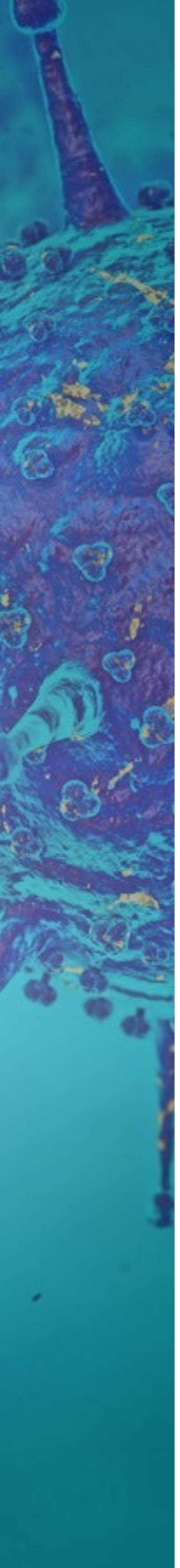
Outline

- ❖Problem Statement and Approach
- ❖Description of the dataset
- ❖Details of SparkNLP Extraction
- ❖Motivation for predicting Length-of-Stay
- ❖Prediction model performance
- ❖Conclusion

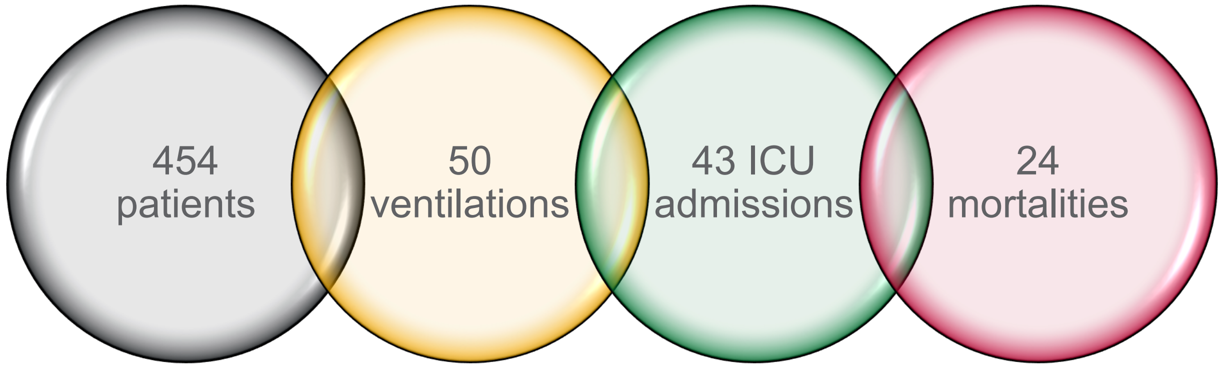
Pacific Northwest NATIONAL LABORATORY

Problem Statement and Approach

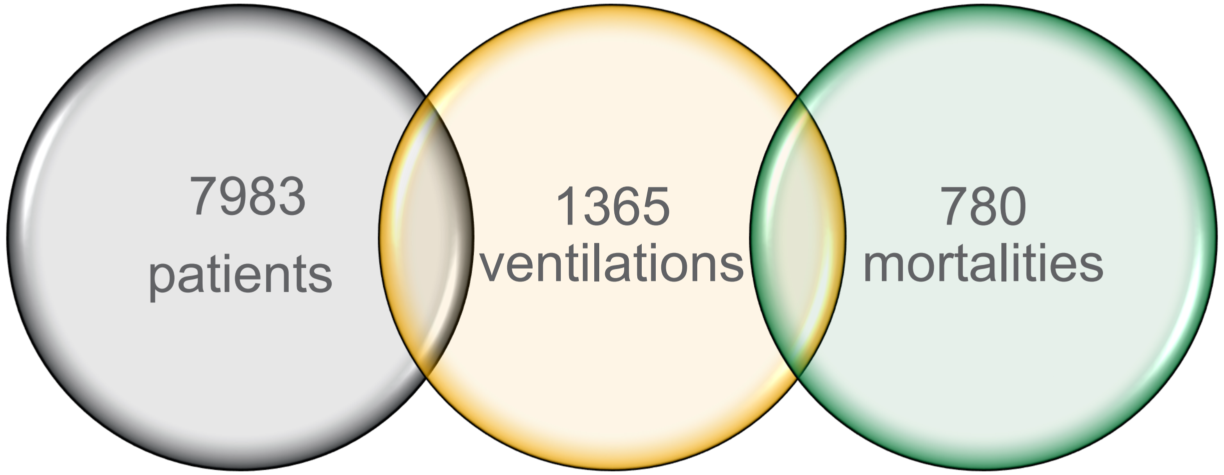
- **Need**: Given a patient's social, clinical and mental health (history) can we predict severity risk
  - Predict from baseline admission + hospital stay for first 'k' days
  - Predict severity as a function of look ahead (1 day, 3 days, 7 days)
- **Uniqueness of our Approach**:
  - Transforming multi-model data sources into a unified representation learning space
  - Develop a self-supervised learning prediction on that unified representation
- **Case study**:
  - A 9-month COVID-19 dataset from Stanford University Medical Center



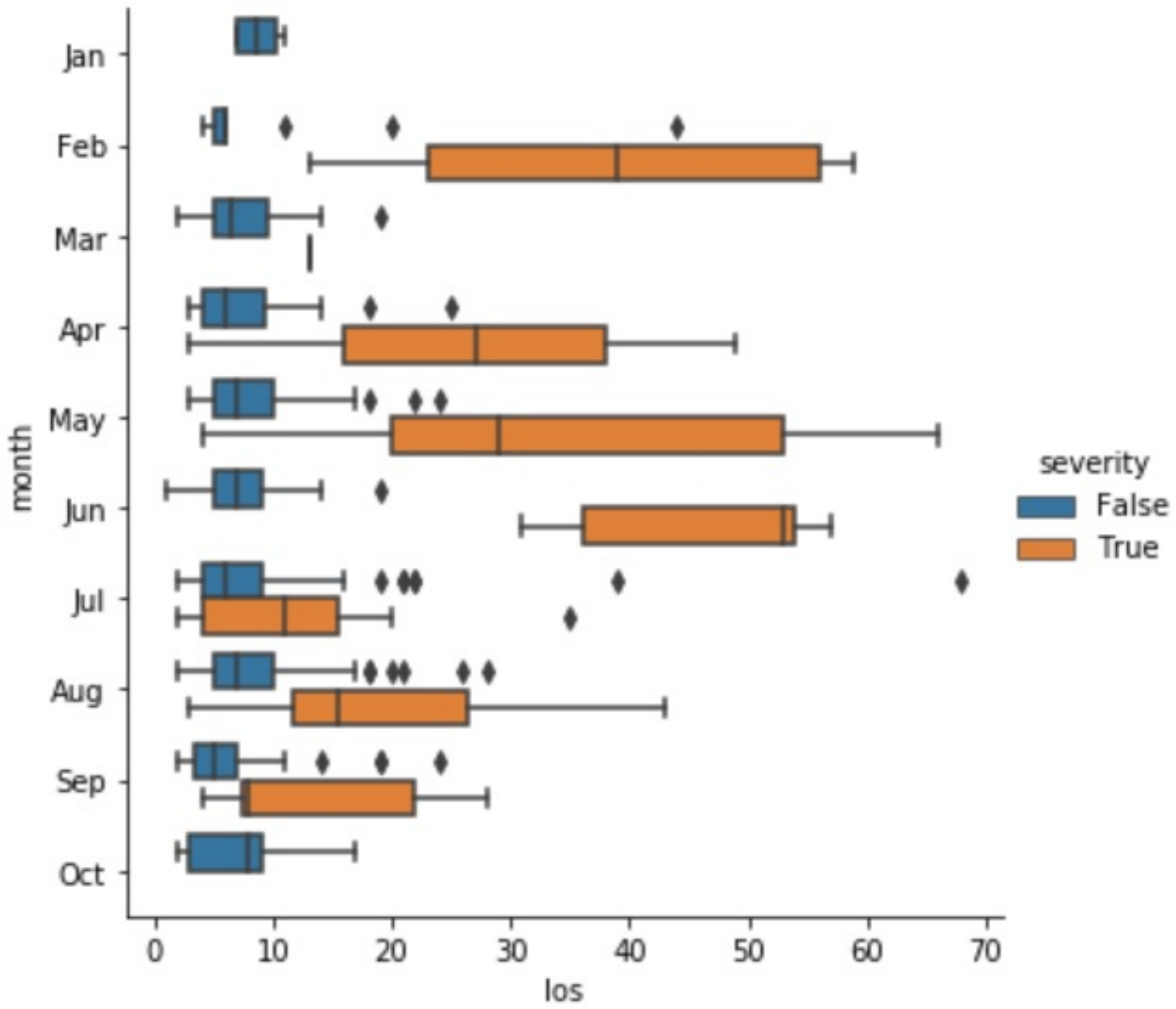
Our Data



COVID-19



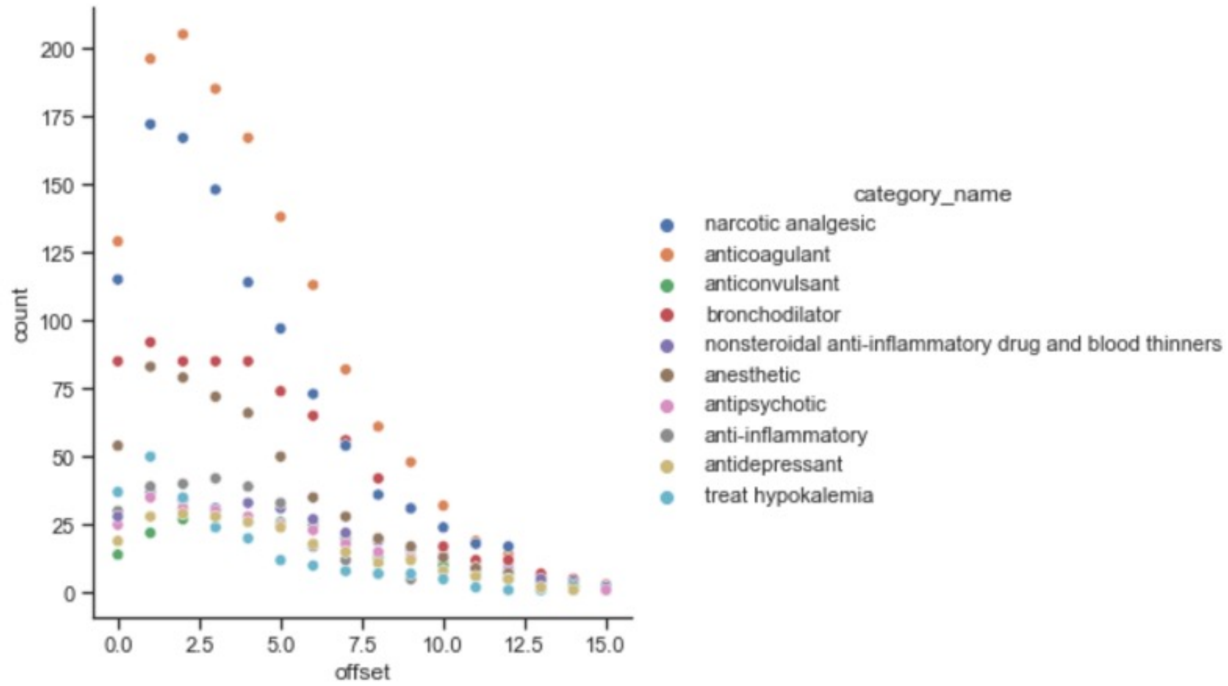
Acute Respiratory Distress Syndrome (ARDS)



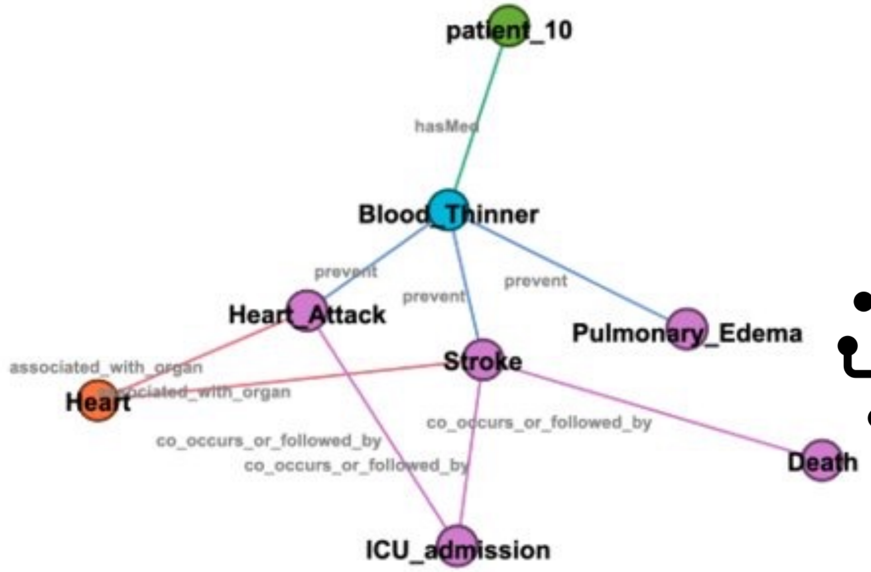
Distribution of Length of Stay for COVID-19 (LOS)

Heterogeneous Datasets

Diagnosis codes and drug codes over time



Clinical Knowledge Graph



Natural language clinical notes

82yo female with hx cad,chf,htn who was recently at [**Hospital1 **] with PE presented to ew with **fever/hypoxia/sob**. Pt was being tx at rehab for PNAx3 days. See admission fhpa for details pmh/hpi.

R.O.S.
Resp- **Chest xray with b/l lower lobe infiltrates**. Admitted on 100%nr with sats 94-98%. **Pt will desat to 80's very quickly when 02 off**. Pt becoming sob with minimal activity with rr 30's. **Lungs with crackles half up bilaterally**. To recieve daily lasix in am. Abg on 100% nr 92/29/7.40.
[**Name (NI) **] Pt recieving 2l ns in ew. Bp and hr stable with adequate uo. Pt denies cp. Does c/o back pain. Ekg done without change.
[**Name (NI) **] Pt **alert and orientedx**3. Cooperative with care.
Id- T-max 102.6 in ew. Now down to 100.1. Cont on zosyn/vanco. Cultures pending.
Gi- Taking liquids without problem. **Abdomen soft with good bowel sounds**. No s/s active bleeding. Pt with elevated inr on coumadine.
[**Name (NI) **] Pt had lived alone. Has been at rehab for past month. Daughters [**First Name8 (NamePattern2) **] [**Last Name (NamePattern1) 9173**] and [**First Name4 (NamePattern1) 6626**] [**Last Name (NamePattern1) **] very involved and are health care proxys. **Although pt had been dnr in past is now full code and would be intubated.**

SparkNLP: Extracting High-Confidence Symbolic Representation of Clinical Notes

- We parse each clinical notes into a set of blocks and invoke SparkNLP pipeline (*entity extraction, assertion, relation extraction*) on selected blocks
- Eventual goal is to go beyond extracting a set of entities from text

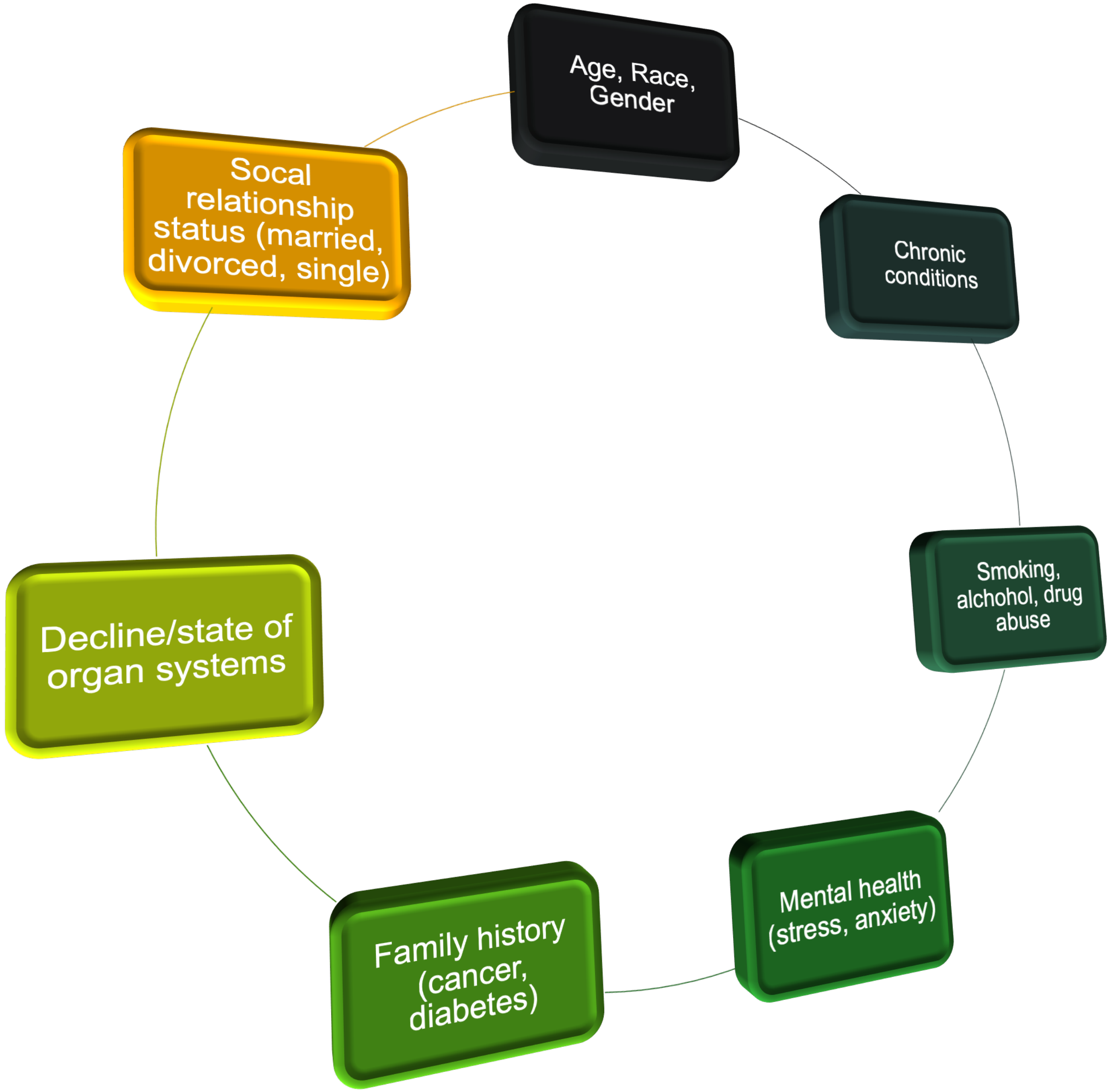
```
clinical_text = """
Patient with severe fever and sore throat.
He shows no stomach pain and he maintained on an epidural and PCA for pain control.
He also became short of breath with climbing a flight of stairs.
After CT, lung tumor located at the right lower lobe. Father with Alzheimer.
"""

light_model = LightPipeline(model)

get_clinical_assertion_light (light_model, clinical_text)
```

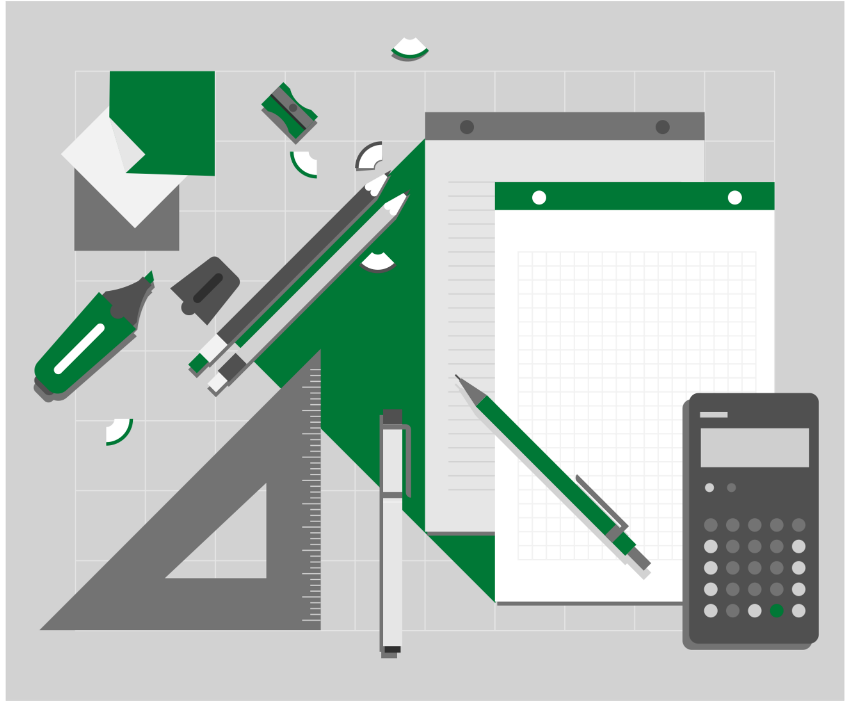
	chunks	entities	assertion
0	severe fever	PROBLEM	present
1	sore throat	PROBLEM	present
2	stomach pain	PROBLEM	absent
3	an epidural	TREATMENT	present
4	PCA	TREATMENT	present
5	pain control	PROBLEM	present
6	short of breath	PROBLEM	conditional
7	CT	TEST	present
8	lung tumor	PROBLEM	present
9	Alzheimer	PROBLEM	associated_with_someone_else

After

Before

Important Admission Profile Factors

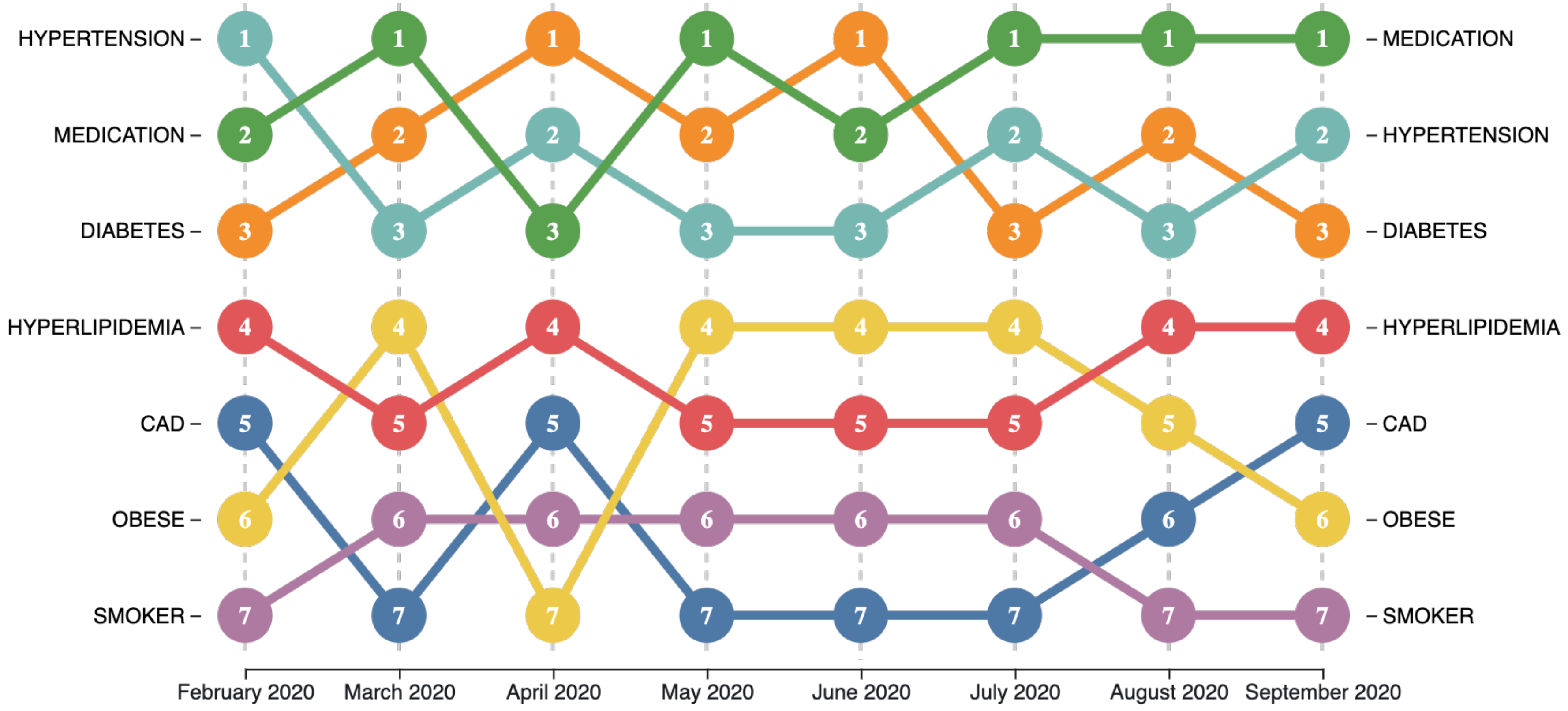




Clinical notes are particularly valuable for such information

Patient Risk Factors: Extracted using SparkNLP

- We looked at most frequent risk factors each month of hospital admission
- Top-3 remain consistent over time



Turning towards Analysis with Multiple Factors

Example of multiple factors: comorbidities, set of concomitant drugs, demographics

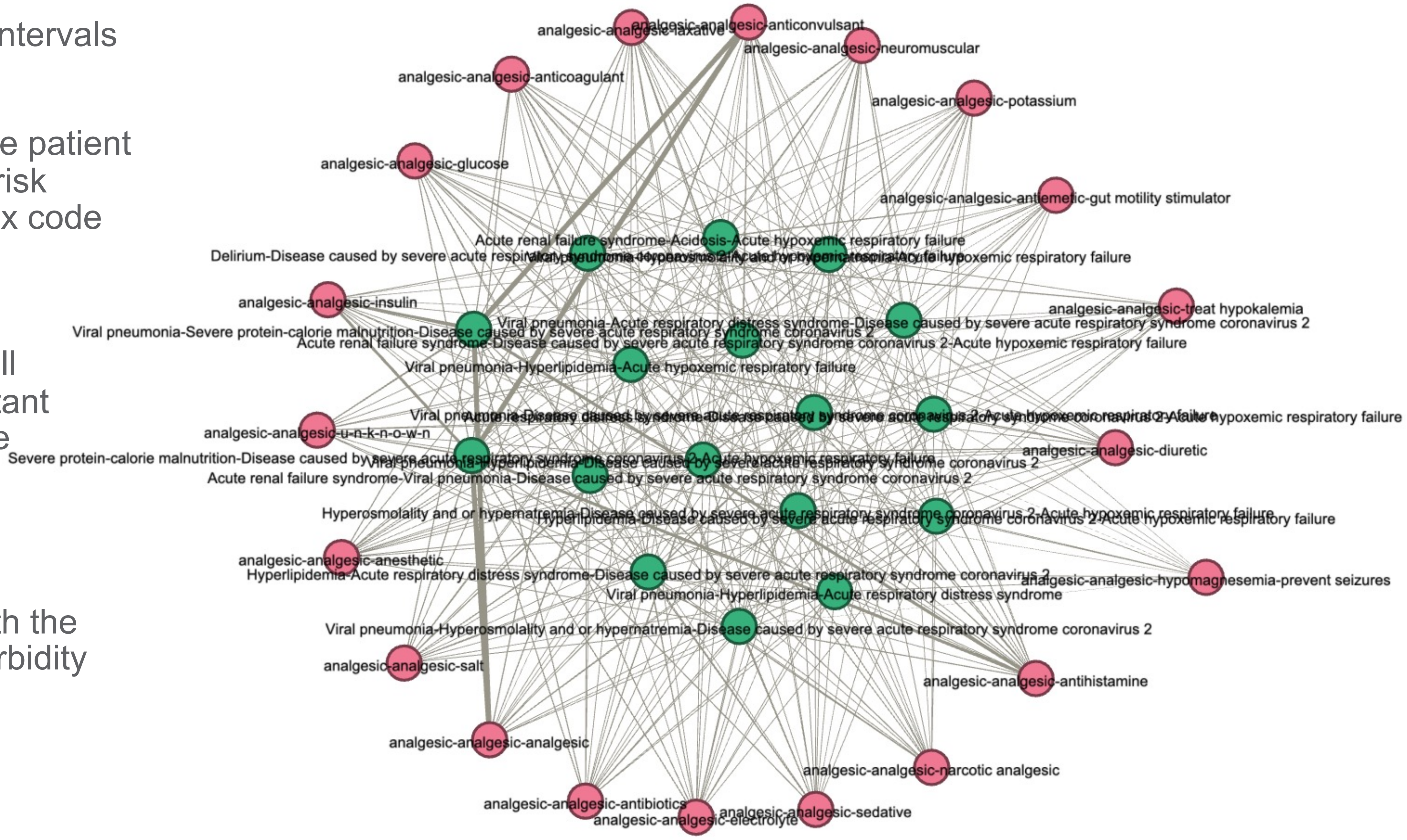
Studying relationships between co-morbidity and concomitant drugs are an obvious step

Reality of data:

- Sparse coverage of condition codes (maybe logged only during change)
- High-resolution coverage of drugs

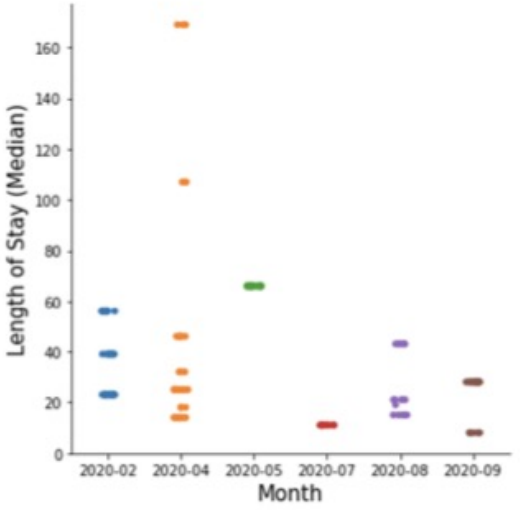
Building frequent comorbidity and concomitant drug Interaction graph

- Map patient data into time intervals
- For each time interval define patient state using combination of risk factors and observed Dx, Rx code categories
- The graph edge indicates all comorbidities and concomitant drugs that occurred in same interval.
- The edge weight indicates median LOS associated with the patients who had the comorbidity pattern and the treatment

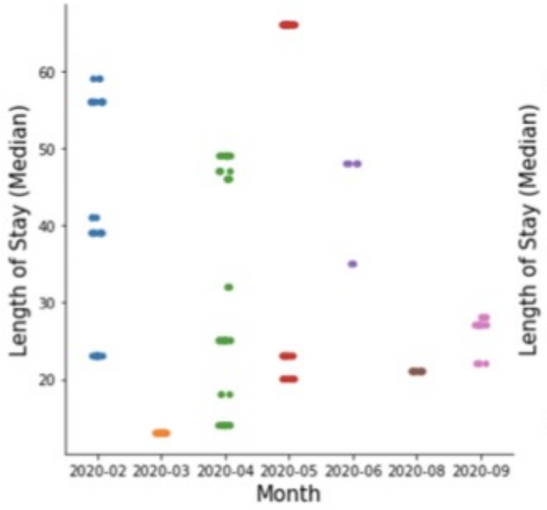


Study Treatment Effectiveness via Comorbidity Patterns

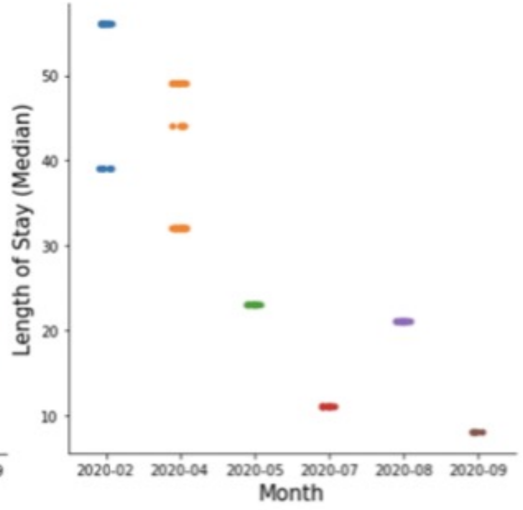
- Comorbidity analysis can shed light on where the medical community has learnt to treat COVID-19 patients better (or as a mix of population adapting to COVID-19)



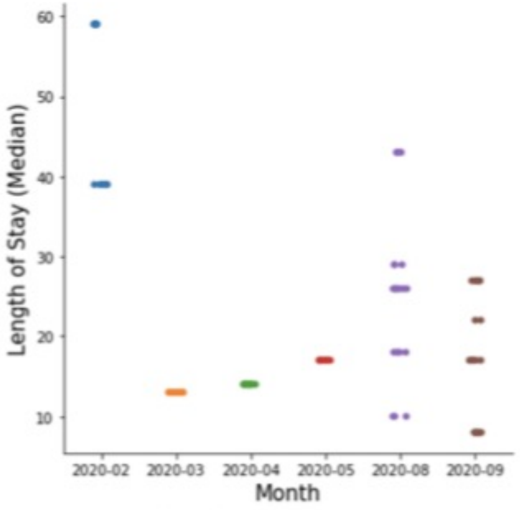
Acute renal failure; acute hypoxemic respiratory failure; disease caused by COVID



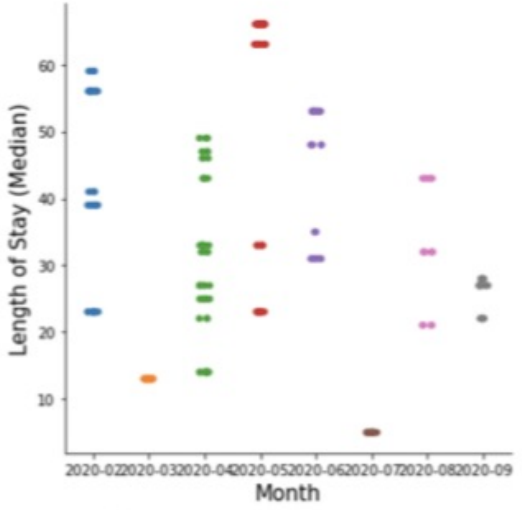
Acute respiratory distress Syndrome; COVID; Acute Hypoxemic respiratory failure



Delirium; COVID; Acute Hypoxemic respiratory failure



Hyperlipidemia; COVID; Acute Hypoxemic respiratory failure



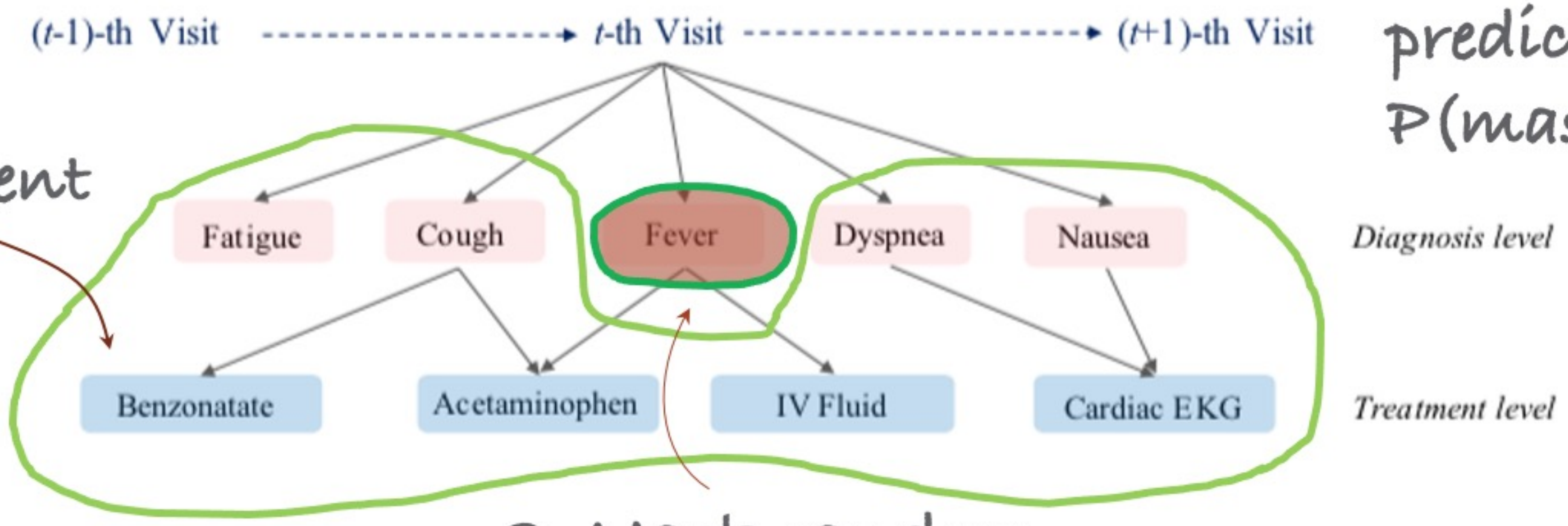
Viral pneumonia; Acute respiratory distress Syndrome; COVID;

Choudhury S., K. Agarwal, C.M. Ham, P. Mukherjee, S. Tang, S. Tipirneni, C. Reddy, S. Tamang, R. Rallo, V. Kocaman, 2021. "Tracking the Evolution of COVID-19 via Temporal Comorbidity Analysis from Multi-Modal Data." In *AMIA 2021 Annual Symposium (under review)*

Patient State Representation Learning and Prediction

1. Sample "patient context"

3. Learn model to predict P(mask|context)



2. Mask random observations

Choi, E., Xu, Z., Li, Y., Dusenberry, M.W., Flores, G., Xue, Y. and Dai, A.M., 2019. Graph convolutional transformer: Learning the graphical structure of electronic health records. *arXiv preprint arXiv:1906.04716*

Pacific Northwest
NATIONAL LABORATORY

Self supervised learning for patient outcome prediction

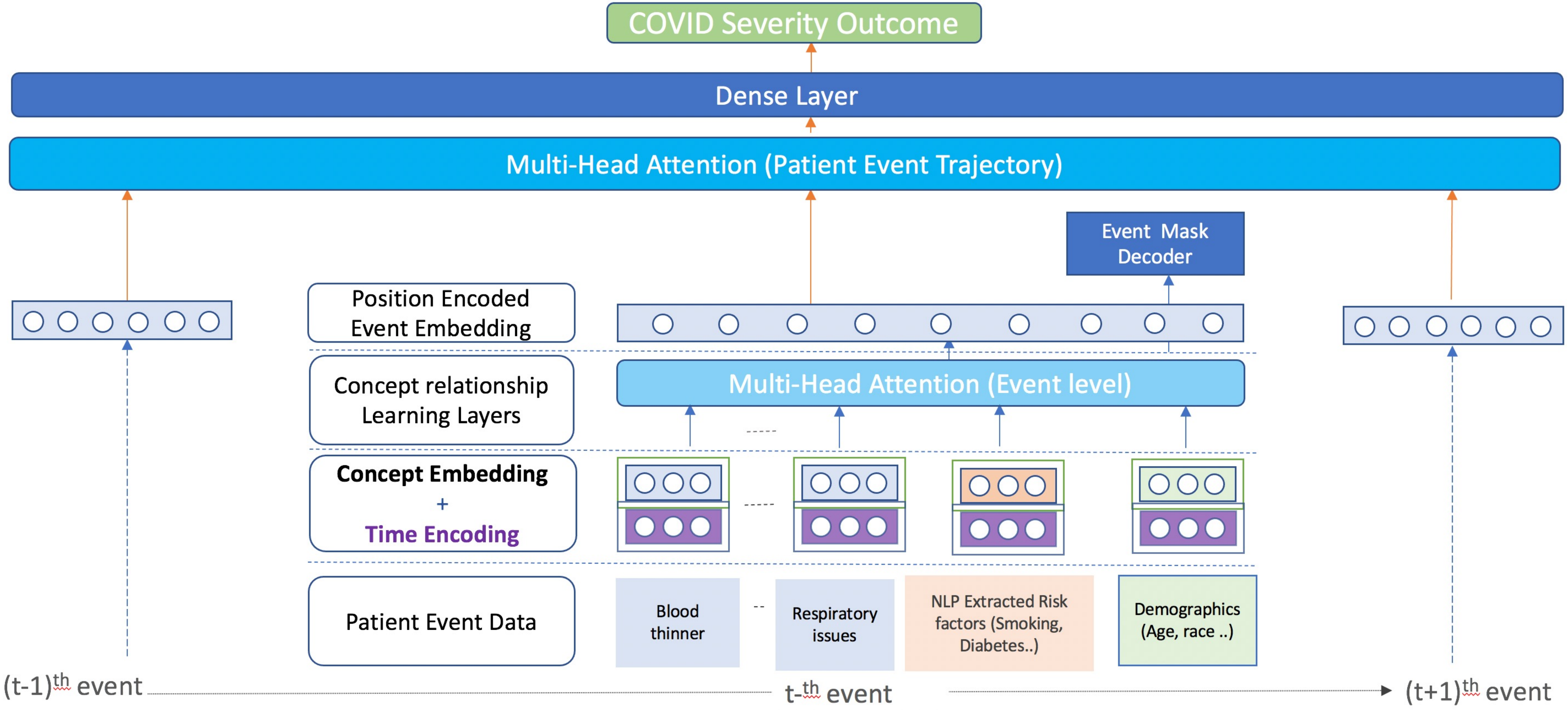
Step 1: Learn Event Representation

- Build event representation
  - Encode Time
- Self supervised event masking approach to generate multiple samples for each patient's temporal event chain.

Step 2: Finetuning for outcome

- Generate event embeddings using layer output from (1)
- Train on COVID severity as an outcome

Multi-modal COVID-19 Severity Prediction Model

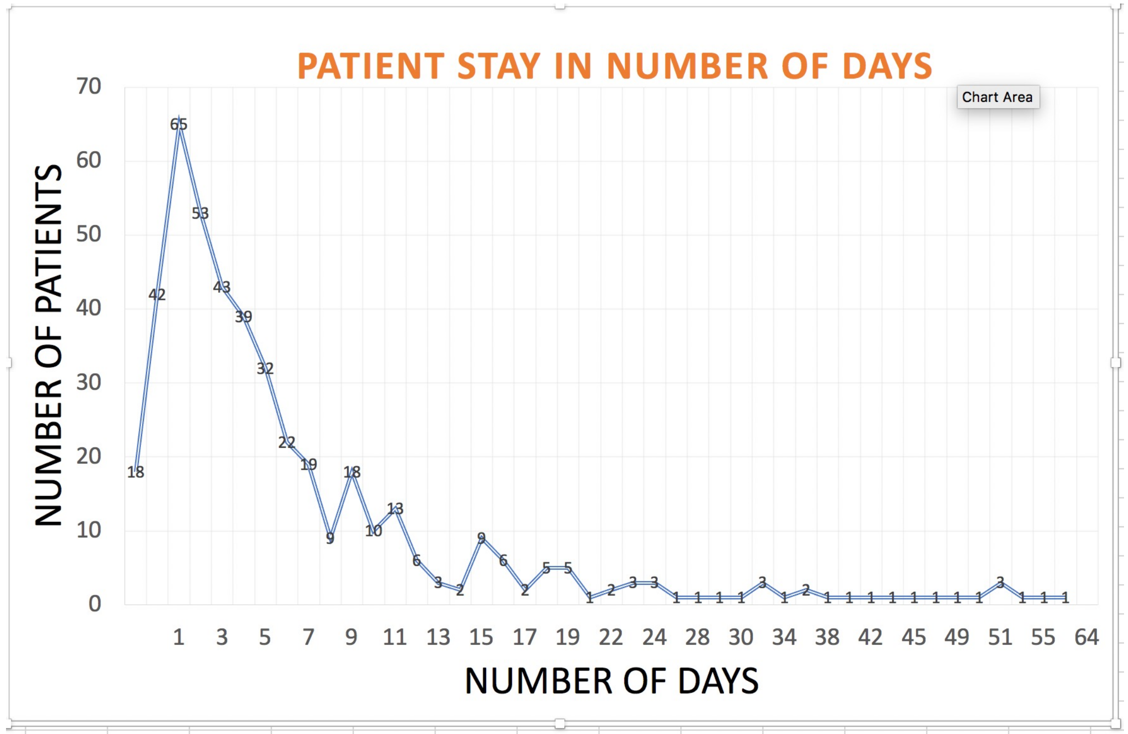


Model Training Details

- **ARDS cohort (used for event representation learning):**
  - Number of patients : 7983
  - Total unique codes observed : 7235
  - Average stay : 16.19 days (max : 536 days)
  - Max num codes in 1 day : 237

- **COVID cohort**
  - Number of patients : 454
  - Outcome : Length of stay, binned to <=3 days

- **Prediction target** : Current Outcome Variable: Will patient be discharged in next 3 days



Model Performance : Impact of aggregation interval

Predicting Masked Feature:

- 24 hours : 7%
- 12 hours : 9%
- 6 hours : 12%

Smaller aggregation intervals leading to more samples and better training

Predicting LOS (using Dx and Rx codes):

- 24 hours : 0.67
- 12 hours : 0.619
- 6 hours : 0.575

Larger aggregation intervals lead to better trajectory learning

Model Performance : Impact of multi-modal features

Aggregation Interval: 24 hours

Without pretraining : (average f1 score, std-dev)

- Dx_Rx codes: 0.53, 0.073
- All structured (Dx, Rx, labs, procedures): 0.54, 0.057
- All structured + demographics : 0.55, 0.044

With pretraining :

- **Dx_Rx only : 0.677, 0.05**
- Dx_Rx_labs_procedures : 0.645, 0.02
- Dx_Rx_labs_procedures + demographics : 0.597, 0.02
- **All structured + demographics + NLP risk factors : 0.69**

Conclusion

Our big idea: Neural-Symbolic Reasoning on Unified Multi-Modal Data

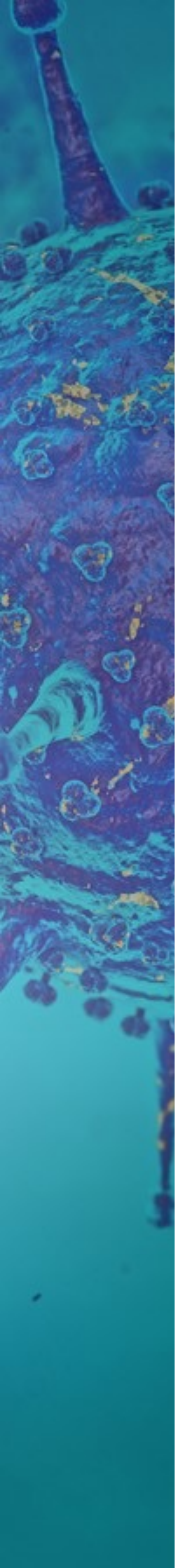
- Integration of clinical notes and structured electronic healthcare records data into a unified symbolic representation and develop neural-symbolic methods on top of it
- We use SparkNLP for transforming clinical notes into this realm

Discover unique insights by working on this Unified Representation

- Allowed us to discover where COVID-19 treatments are being more effective

Multi-Modal prediction models can offer superior performance

- Performance of NLP-integrated model for predicting length of stay for COVID-19 patients is better than model using only structured data
- Bigger value is the interpretability that comes from not turning a document to a vector



Pacific Northwest
NATIONAL LABORATORY

Thank you