CPSL
Cyber-Physical Systems Laboratory

Assisting clinical decision making: perioperative risk prediction

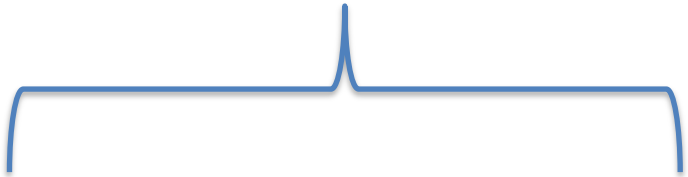
Bing Xue

CSE 521S Lecture

Washington University in St. Louis

Recap on perioperative handoff

- OR-ICU handoff: Transfer patients from operation room to ICU room
- Intraoperative handoff: Change of shifts during surgery
- Handoff lasts only for a very short time
  - must be very informative
  - concentrate on key information relevant to risk mitigation



Early risk identification

Modifiable risks

Early prediction in clinical studies

- Surgeries are highly risky:
  - More than 10% of surgical patients experience major postoperative complications
- Benefits of early prediction
  - Patients: Early identification of risk factors can be crucial to early intervention and improved outcomes
  - Hospitals: Perioperative care is also a major contributor to overall hospitalization expenses
  - Clinicians: Predictions of outcomes provide clinicians with lead time in planning and allocating resources to reduce the cost of care.

Modifiable risks

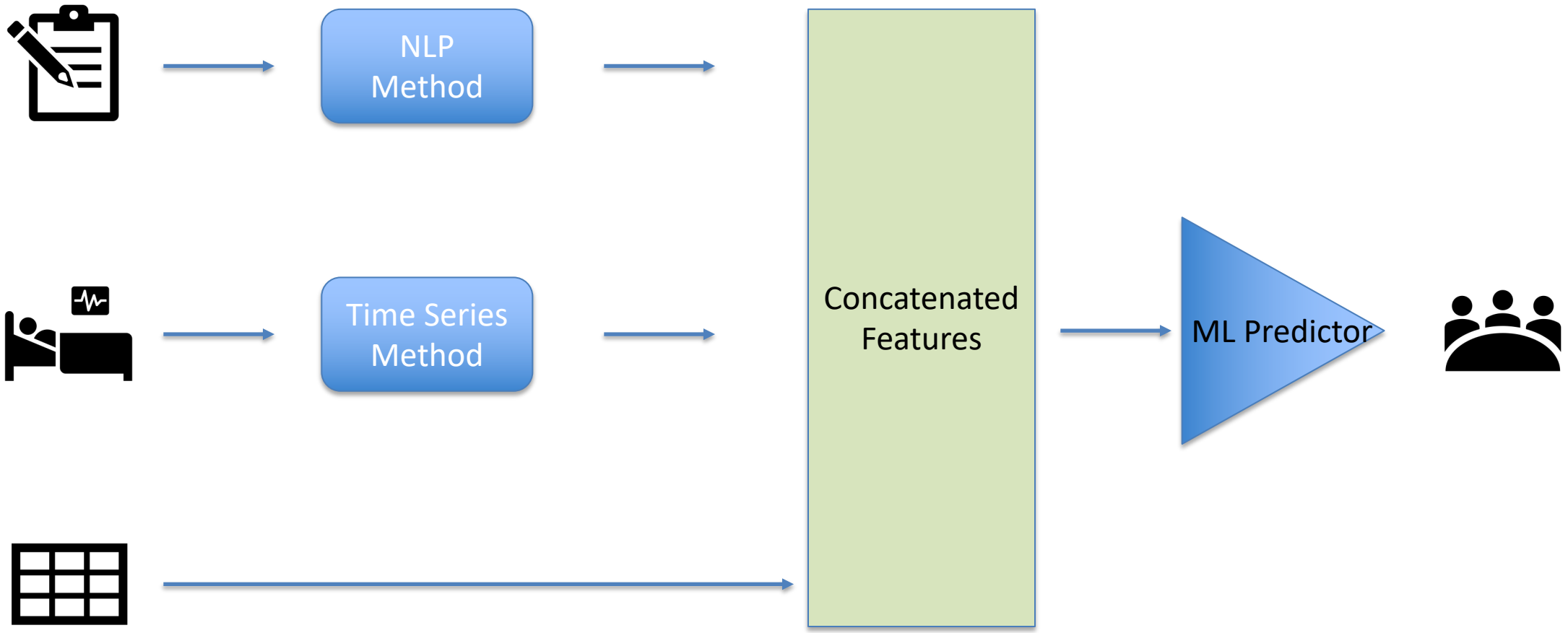
- Selection criteria:
  1. essential for postoperative care management in critical care / surgical units;
  2. potentially modifiable through early detection and mitigation (treatment, medication, etc.)
- Prediction of Postoperative complications
  - OR-ICU handoff: acute kidney injury (AKI), delirium, deep vein thrombosis (DVT), pulmonary embolism (PE), and pneumonia.
  - Intraoperative handoff:
    - Classification outcomes: 30-day Mortality, Delirium, AKI
    - Regression outcomes:
      - Blood Pressure issues (hypo/hyper -tension) (low_sbp_time/total_t, aoc_low_sbp /total_t, low_map_time, low_relmap_time)
      - Temperature issues at arrival to ICU
      - Glycemic control issues (n_glu_high if >180, n_glucose_low if <60)

How to use machine learning to enhance handoff and perioperative care?

CPSL Cyber-Physical Systems Laboratory

- Risk identify: Explore the use of machine learning:
  - Inputs - Electronic Health Records (EHR):
    - Inputs are collected different times: preoperative & intraoperative
    - Inputs are multi-modal:
      - Static Features: demographics, comorbidities, lab tests, treatment / interventions, drugs, etc.
      - Time Series: vital signs, events, measurements, etc.
      - Texts: Clinical notes.
  - Outputs: probability of complications
  - Models?

Overall Framework



NLP Method

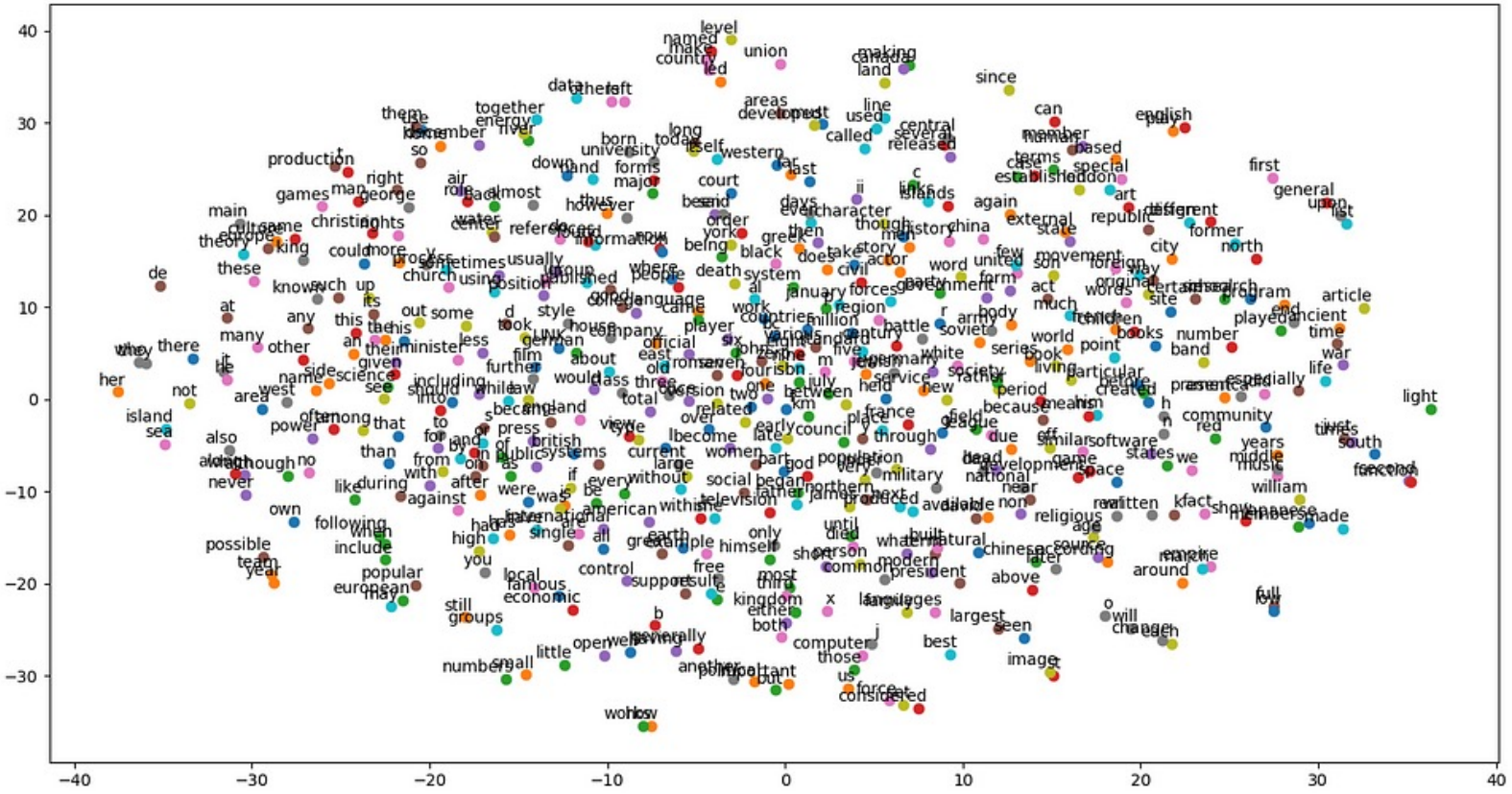
- What are clinical texts?
  - Clinical texts have complex abbreviations or medical terminologies
  - Clinical texts can be structured or unstructured

AN_PROC_NAME	AN_DATE
ESOPHAGOGASTRODUODENOSCOPY wu prep (N/A); CO...	2019-03-05 00:00:00
BILATERAL MYRINGOTOMY TUBE INSERTION (Bilatera...	2019-03-14 00:00:00
FUSION SPINAL ANTERIOR LUMBAR/THORACIC WITH IN...	2019-03-28 00:00:00
EXCHANGE STENT - URETERAL (Right Ureter)	2019-03-05 00:00:00
Robotic Assisted Laparoscopic Partial Fundopli...	2019-03-25 00:00:00

PROCED_ICD	PROCED_ICD_SHORT_TITLE	PROCED_ICD_LONG_TITLE
[9604, 9671, 9634]	['Insert endotracheal tube', 'Gastric tube irr...	Insertion of endotracheal tube Other irrigatio...
[3613, 3615, 3961, 8872, 9904, 9905, 9907]	['Aortocor bypas-3 cor art', '1 int mam-cor ar...	Aortocoronary bypass of three coronary arterie...
[151]	['Ex cereb meningeal les']	Excision of lesion or tissue of cerebral meninges
[3601, 3607, 3723, 8852, 8856, 9920]	['Ins drug-elut coronry st', 'Rt/left heart ca...	Insertion of drugeluting coronary artery stent...
[3601, 3607, 3722, 8856, 9920, 3722, 8856, 885...	['Ins drug-elut coronry st', 'Left heart cardi...	Insertion of drugeluting coronary artery stent...

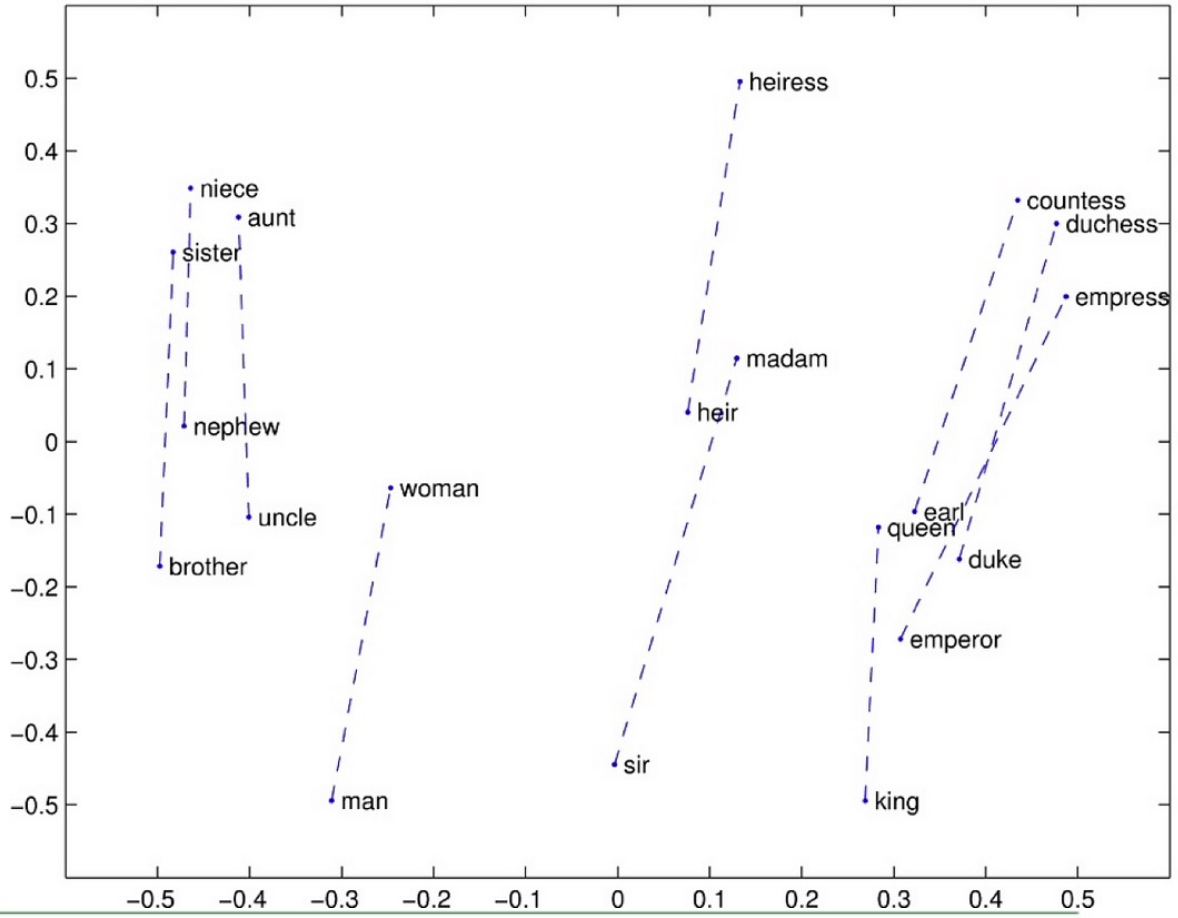
NLP Method

- How to read the texts?
  - **Word Embedding**: Convert words to embeddings



NLP Method

- How to read the texts?
  - Continuous Bag-of-Words (CBOW)
    - predicts which is the most likely word in the given context
  - GloVe
    - trained on aggregated global word to word co-occurrence matrix from a given text collection of text documents.
  - FastText
    - It is an extension to CBOW model
    - but unlike Word2Vec which feeds whole words into the neural network, FastText first breaks the words into several sub-words (or n-grams) and then feed them into the neural network.



NLP Method

- **Limitations of Word Embedding**
  - they don’t take into consideration the order of words in which they appear
    - leads to loss of syntactic and semantic understanding of the sentence.
    - For example, “You are going there to teach not play.” And “You are going there to play not teach.”
  - cannot give satisfactory results on large amount of text data, as same word may different meaning in different sentence depending on the context of the sentence.
    - “I have scuba diving in my **bucket** list.” And “There is a **bucket** filled with drinking water.” In both the sentences, the word “bucket” has different meanings.

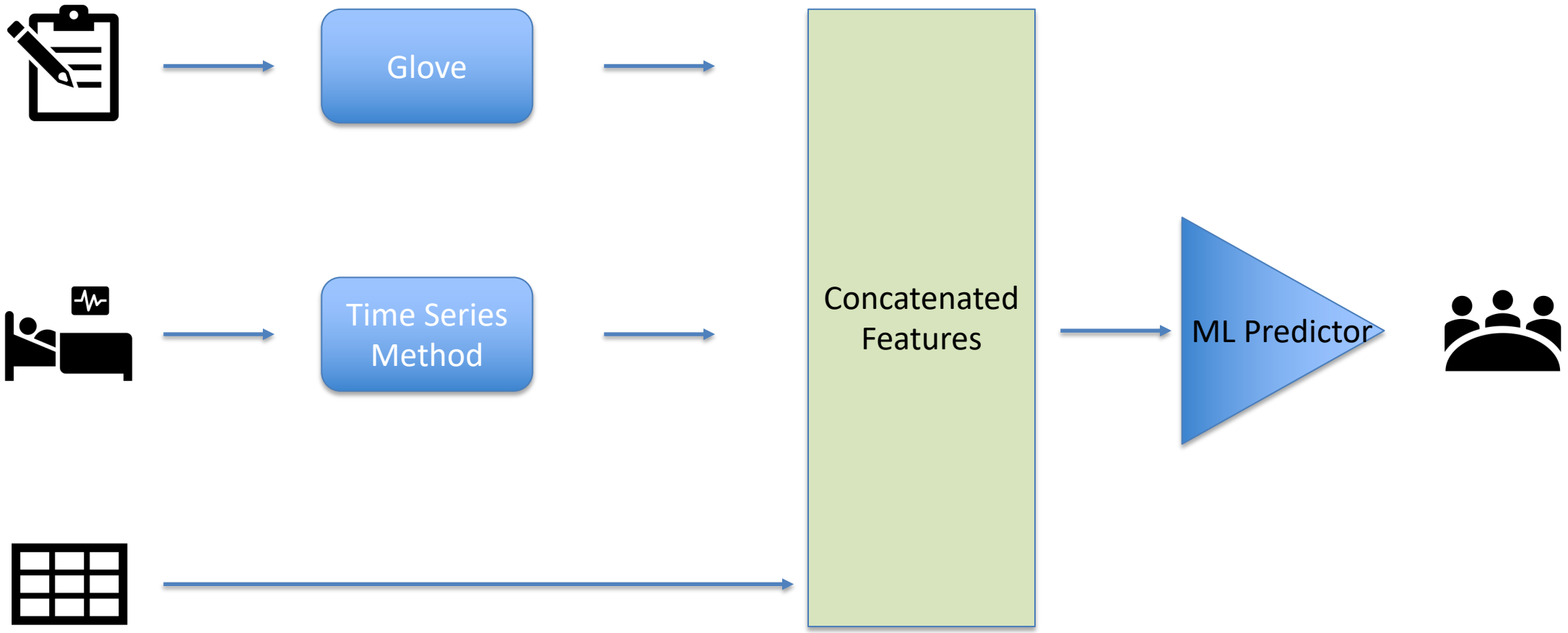
NLP Method

- **Sentence Embedding**
  - A simple way of obtaining sentence embedding is by averaging the word embeddings of the all the words present in the sentence. But they are not accurate enough.
  - ELMO (Embedding from Language Model)
    - bi-directional deep LSTM network for producing vector representation. ELMo considers words within which context they have been used rather than creating dictionary of words with its vector form.
  - Doc2Vec
    - Aka Paragraph Vector, is an extension of Word2Vec that **learns vector representations of documents rather than words**.
    - learns vector representations of documents by combining the word vectors with a document-level vector.

NLP Method

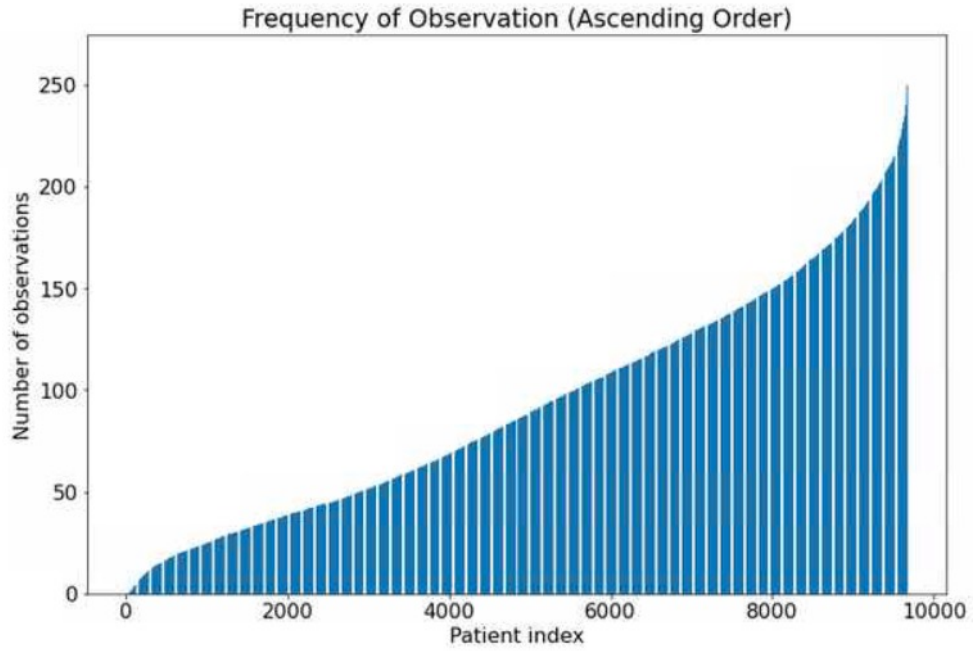
- We compared the performance between various word embedding & sentence embedding methods, and found out that GloVe performs the best
  - On both BJC dataset and MIMIC-III dataset
  - On various tasks: postoperative complications, mortality, LoS, discharge, etc.

Overall Framework



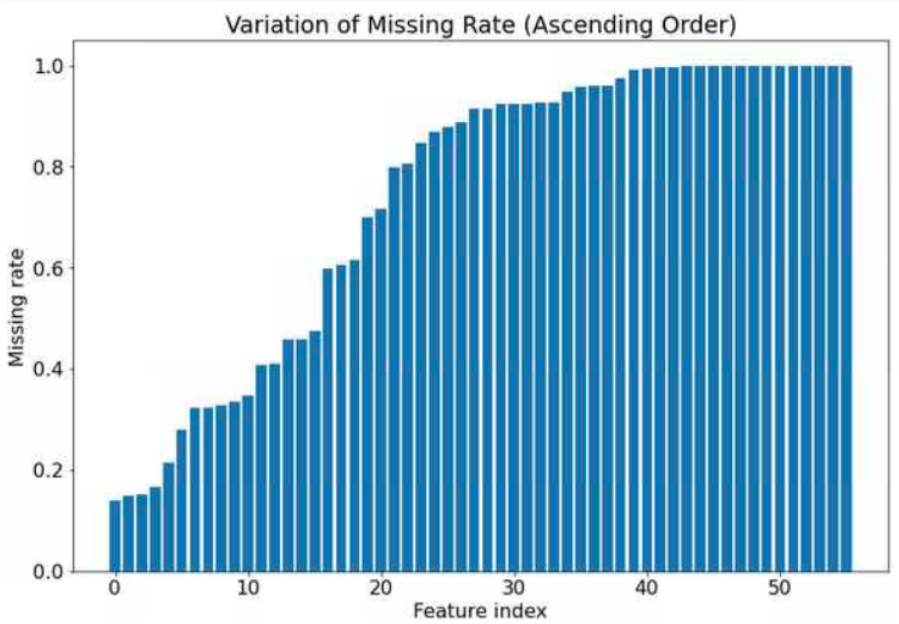
Challenges in time series

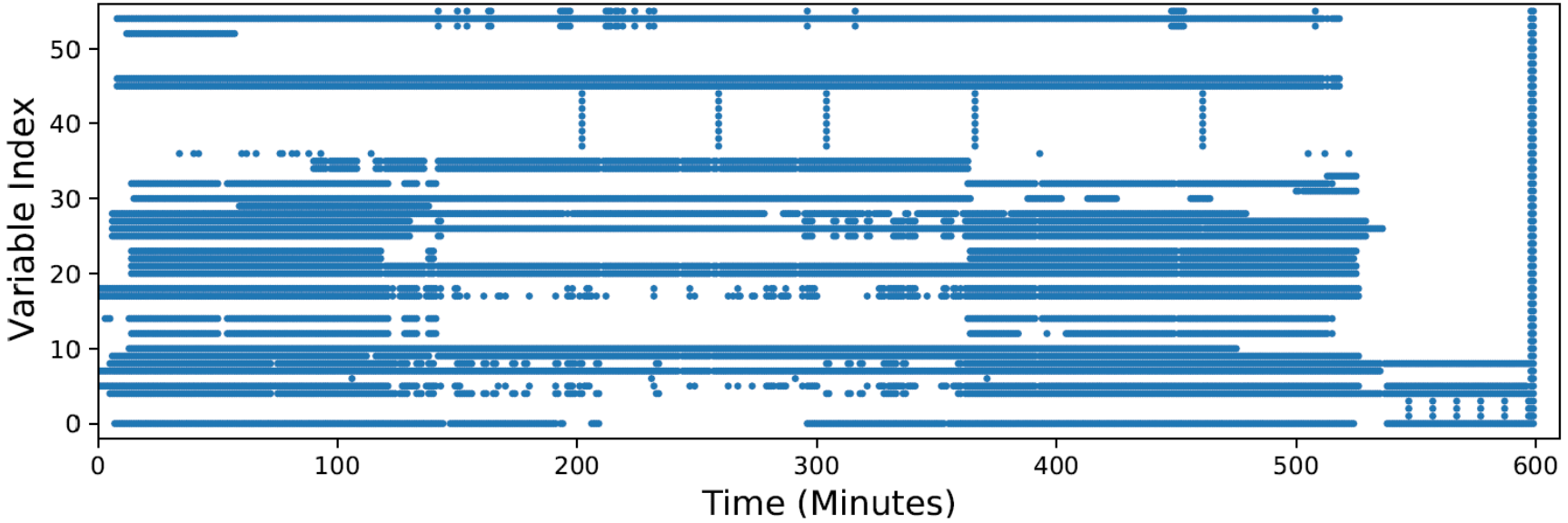
- Granularity:
  - Different vital signs are collected with different frequency (per minute, hour, day, etc)
- Sequence lengths:
  - Surgery may last very long (with more observations/data in the time series) or short (with few observations), depending on individual's surgery



Challenges in time series

- Even for the same patient, time series are irregular:
  - Each feature is collected with different irregularity



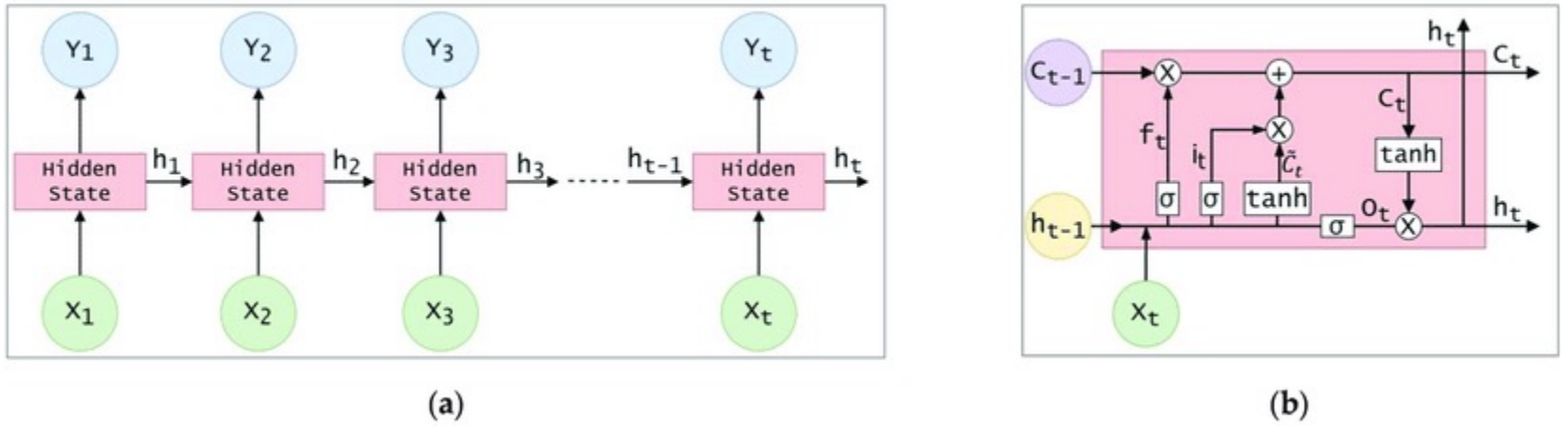


Solutions

- There are 2 problems to solve:
  - How to deal with the irregularity
  - How to capture the local as well as global patterns
- What are the candidate solutions?
  - LSTM/RNN, BRITS, GRUD, feature engineering, etc.

Solutions

- Solution 1: RNN/LSTM
  - Impute the missingness by previous observations (or average value)
  - Use RNN/LSTM to learn the final hidden states from the multivariate time series



The architectures for time-series forecasting model: (a) Recurrent Neural Network (RNN); (b) Long-Short Term Memory (LSTM).

Solutions

- Solution 2: BRITS (Bidirectional Recurrent Imputation for Time Series)
  - Use a mask vector to record the missingness
  - Use the previous hidden states of RNN to predict the next observation
  - If next observation is missing, then use the predicted value; otherwise use the actual observation

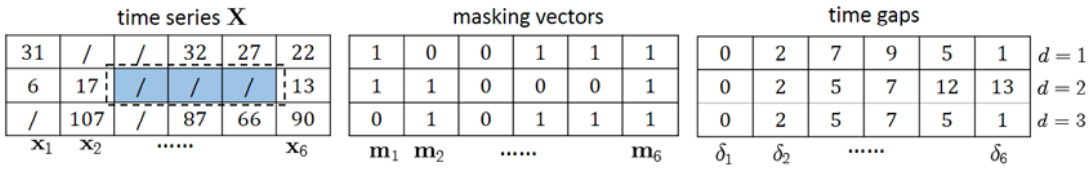


Figure 1: An example of multivariate time series with missing values. $\mathbf{x}_1$ to $\mathbf{x}_6$ are observed at $s_{1...6} = 0, 2, 7, 9, 14, 15$ respectively. Considering the 2nd feature in $\mathbf{x}_6$, the last observation of the 2nd feature took place at $s_2 = 2$, and we have that $\delta_6^2 = s_6 - s_2 = 13$.

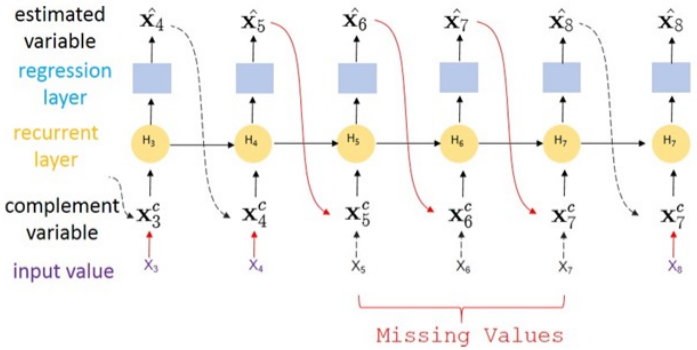


Figure 2: Imputation with unidirectional dynamics.

Solutions

- Solution 3: Gated Recurrent Unit GRU-D
  - Similar to BRITS in the setting
  - Use GRU instead of RNN
  - Instead of imputing observations at each time step, use *the time interval since last observation* to estimate the current hidden state



(a) GRU (b) GRU-D

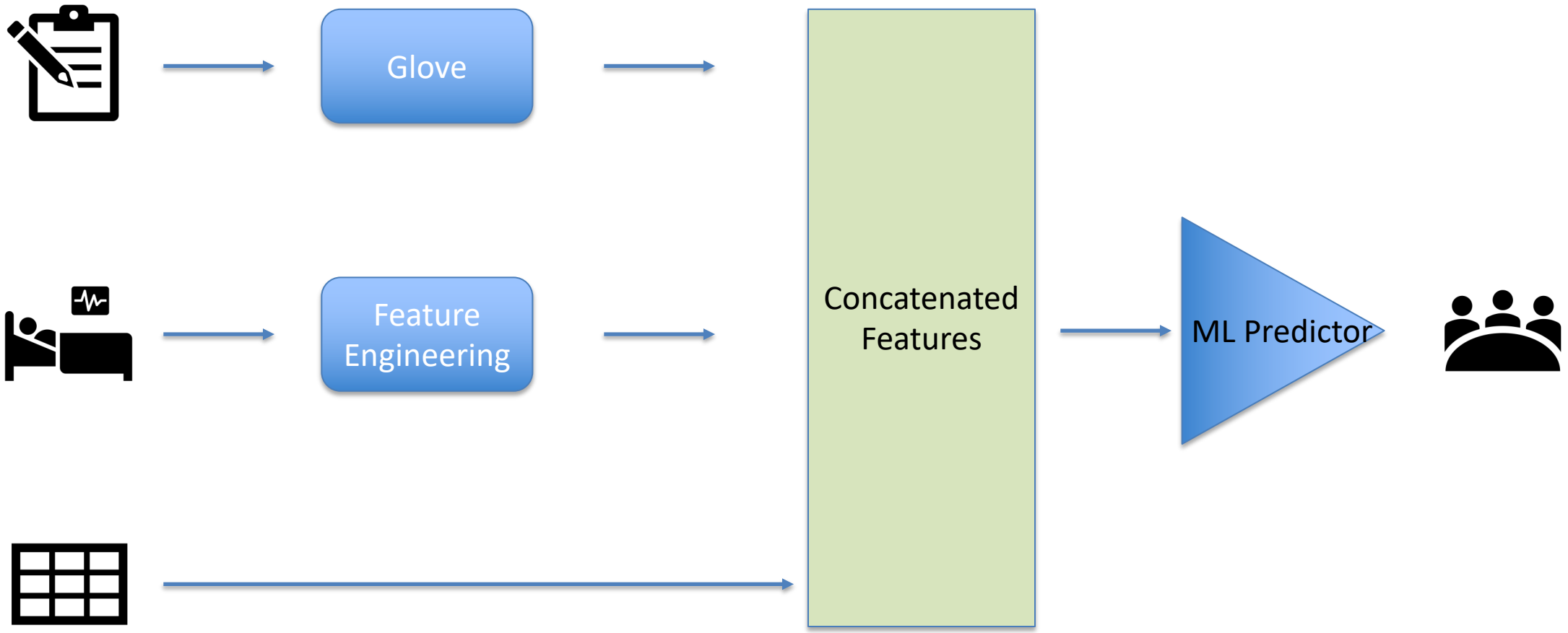
Figure 3: Graphical illustrations of the original GRU (left) and the proposed GRU-D (right) models.

Solutions

- Solution 4:
  - Feature engineering
    - Impute by average (if necessary)
    - Compute statistical features to characterize the distribution of the observations
      - E.g.: max, std, entropy, kurtosis, skewness, etc.

- Which one is better?
  - Feature engineering had similar or better results in most cases
  - Fusion of time series outputs with static features is an open problem
    - Dingwen had a paper discussing this

Overall Framework



Challenges in EHR data: prevalence of missing values

- How to characterize patients with missing values:
  - The correlation between input variables is undermined by missing values;
    - Example: various missing rates of static features in different datasets

Dataset	Delirium	OR	MIMIC-III
Missing Rate	52.65%	2.50%	11.79%

Solution

- Common imputation methods:
  - Mean Imputation
  - Zero Imputation
  - Median Imputation
  - Mode Imputation
  - K-NN Imputation
  - MissForest
    - find the "best" predictions for the missing values given some set of predictors and random forest predictor
  - MICE
    - runs multiple imputation models on bootstrapped samples of observed data and randomly selects the predicted value from one of the models.

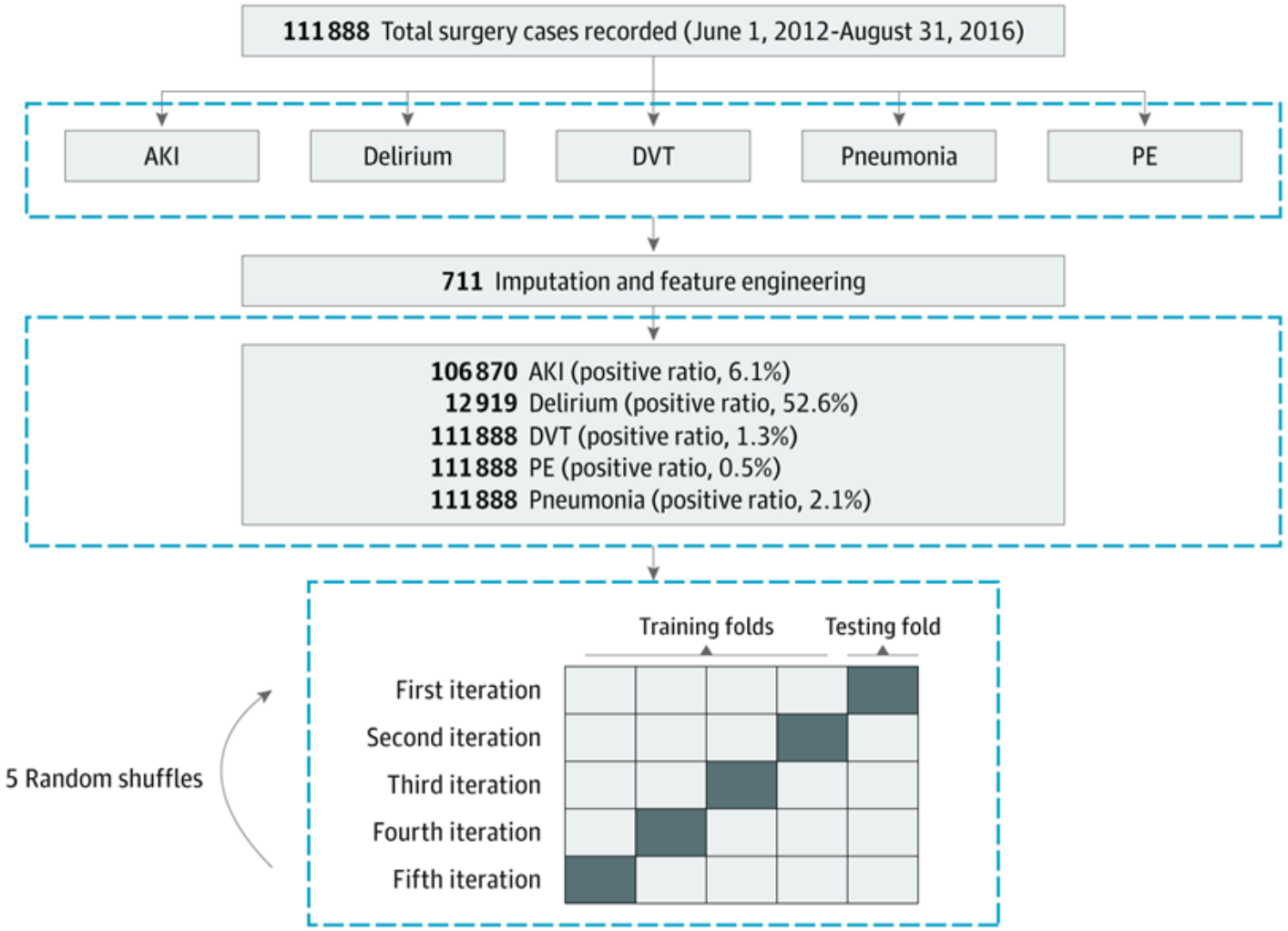
Solution

- Indicator method
  - Use mean imputation but use an extra indicator vector to highlight the missingness
  - Achieved better performance than other imputation methods
    - Indicator reflects clinicians' intention to skip/order certain lab tests, treatments, etc
    - Our findings were empirically and theoretically validated by Mike et. al (KDD'23)

eTable7: The comparison of imputation methods

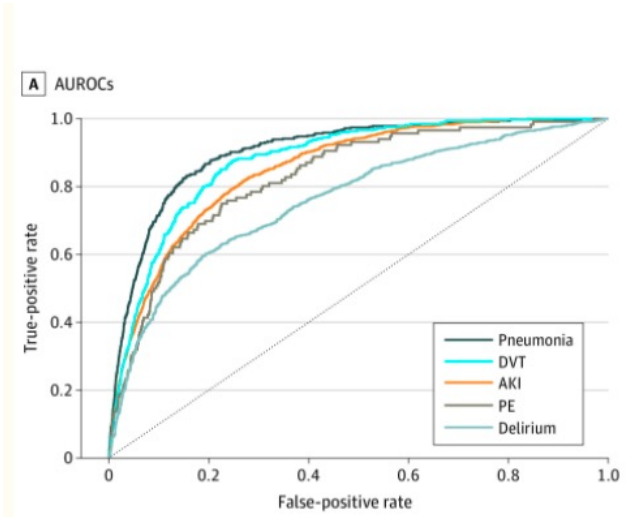
Imputation Method	AUROC	AUPRC
GBT		
Dummy Indicator	0.905 (0.903,0.907)	0.208 (0.203,0.213)
Mean	0.901 (0.898,0.904)	0.206 (0.199,0.213)
Median	0.901 (0.898,0.904)	0.204 (0.199,0.210)
Mode	0.902 (0.900,0.904)	0.203 (0.198,0.208)
kNN-3	0.884 (0.882,0.887)	0.182 (0.176,0.188)
kNN-5	0.884 (0.882,0.887)	0.182 (0.176,0.188)
MICE	0.893 (0.890,0.896)	0.193 (0.187,0.200)
MissForest	0.890 (0.888,0.893)	0.184 (0.178,0.189)

Dataset

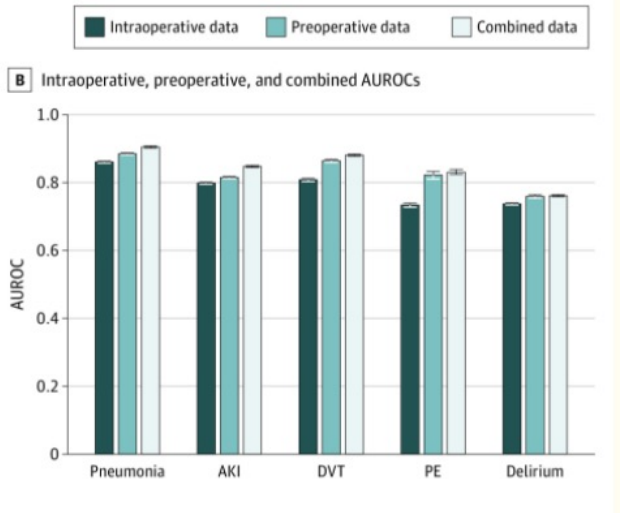

111888 Total surgery cases recorded (June 1, 2012-August 31, 2016)
AKI
Delirium
DVT
Pneumonia
PE
711 Imputation and feature engineering
106870 AKI (positive ratio, 6.1%)
12919 Delirium (positive ratio, 52.6%)
111888 DVT (positive ratio, 1.3%)
111888 PE (positive ratio, 0.5%)
111888 Pneumonia (positive ratio, 2.1%)
Training folds
Testing fold
First iteration
Second iteration
Third iteration
Fourth iteration
Fifth iteration
5 Random shuffles

Observations

- We start with some existing machine learning models : logistic regression, random forest, gradient boosting tree, deep neural network, etc.
  - Preoperative features are more informative than intraoperative time series!



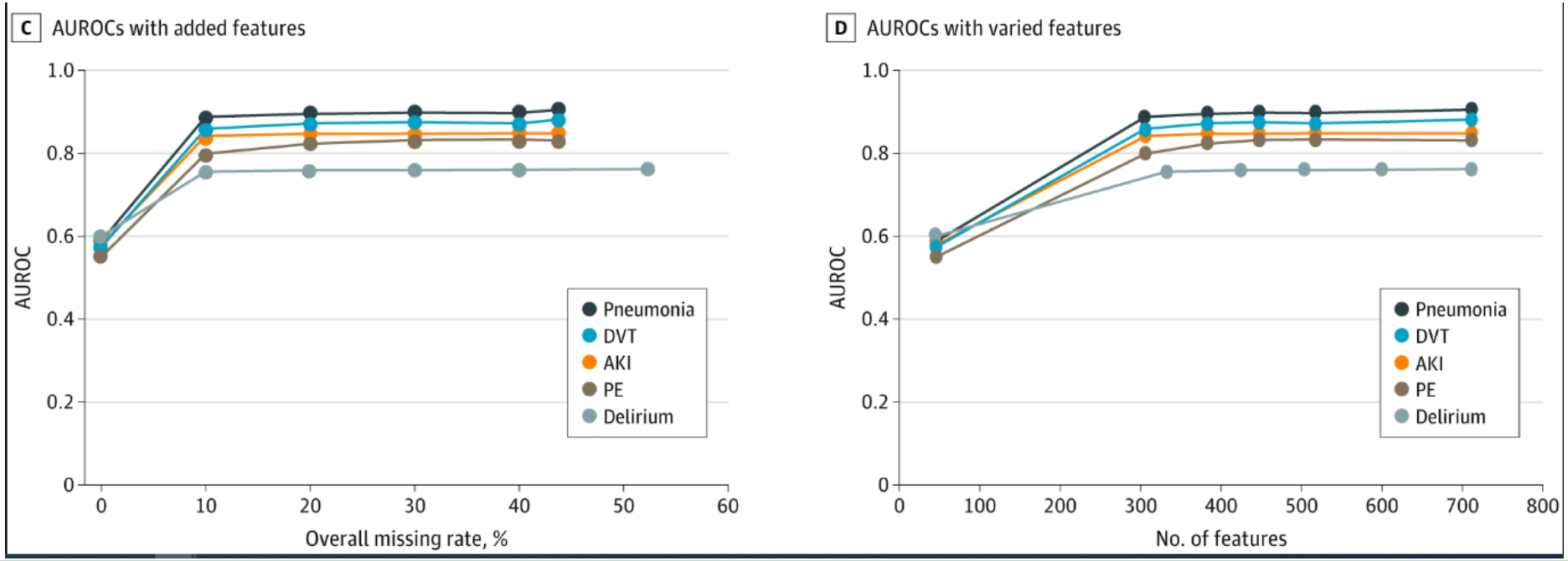
Prediction performance



Usefulness of input variables

Observations

- We start with some existing machine learning models : logistic regression, random forest, gradient boosting tree, deep neural network, etc.
  - Features with higher missing rate also (slightly) help with prediction
  - More features, (slightly) better performance



This brings up our next question

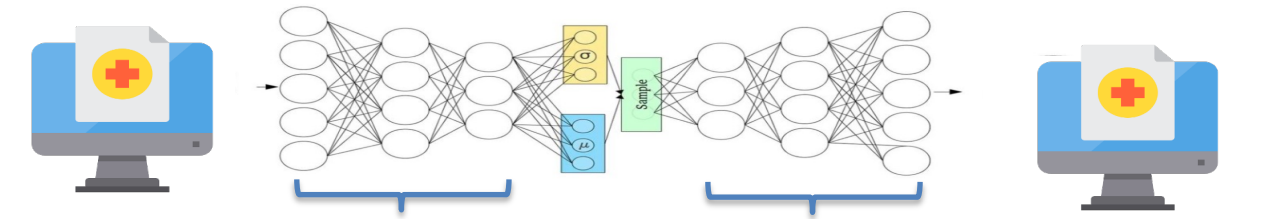
- How to build models with high-dimensional inputs:
  - Many of the input variables are correlated:
    - OR-ICU Handoff: >700 input features; National COVID Cohort Collaborative: >10k input features only in measurements;
  - ML models need to handle large data
    - Curse of dimensionality:
    - Data sparsity
    - Training stage: slow and overfitting
    - Application stage: generalization of the model
  - Objective: can we distill the intrinsic characteristics of patients into a lower dimensional representation?

Objectives

- A new representation / latent embedding:
  - Low-dimensional (easy for predictions);
  - No missingness;
  - Learns the implicit, nonlinear relationship between input features

Solution: generative modeling

- Generative modeling:
  - Encoder-Decoder Architecture:
    - Encoder transforms inputs into a latent representation
    - Decoder transforms latent representation into original/new representation
- Why Variational Autoencoder (VAE):
  - Self-supervised training:
    - VAE has a reconstruction term to self-supervise the latent representation
    - Nice properties in the latent space



Electronic Health Records (EHR)

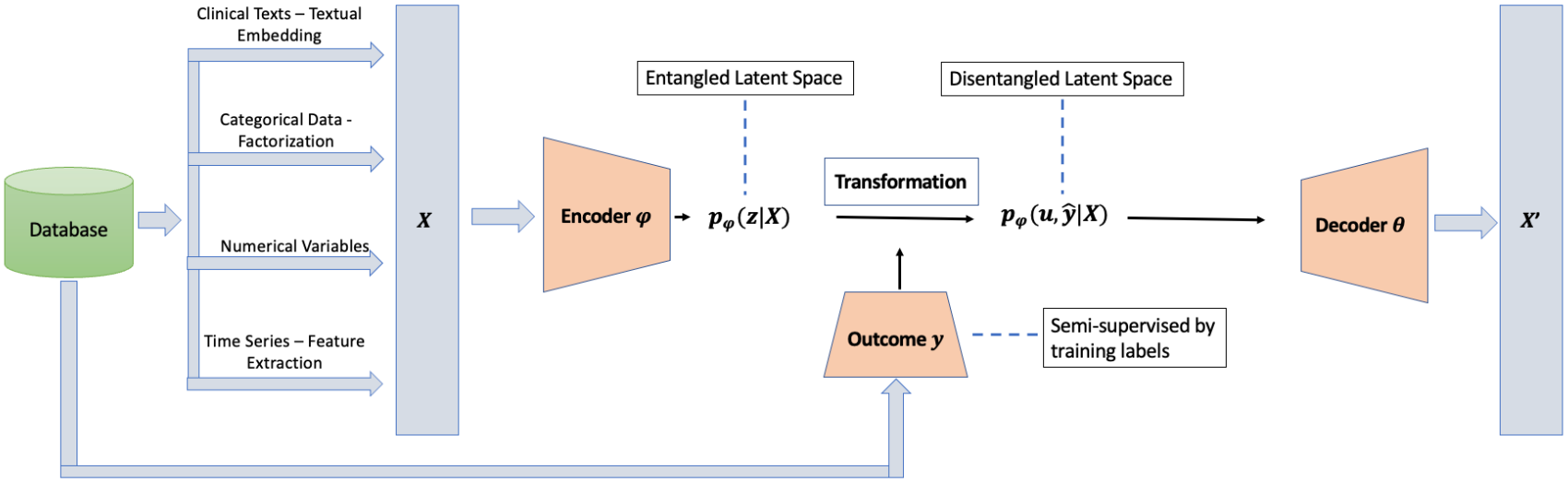
Encoder: compressing into latent representation

Decoder: Reconstructs the original representation

Electronic Health Records (EHR)

clinical VAE (cVAE): Semi-supervised and disentangled latent space

- Open challenges to address
  - Performance: latent encoding does not benefit downstream learning
  - Entangled latent space: difficult to interpret and trust the predictions
- Our proposed cVAE:
  - Prediction guided: use training labels to guide the latent encoding
  - Disentanglement: each dimension represents a single aspect of information → more "interpretable"



Evaluation: Predictive Performance

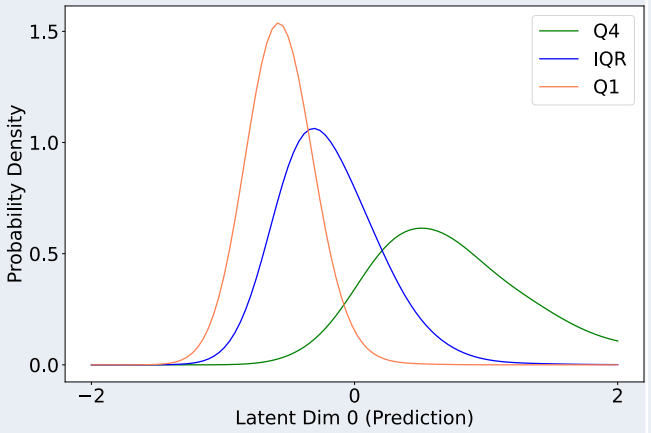
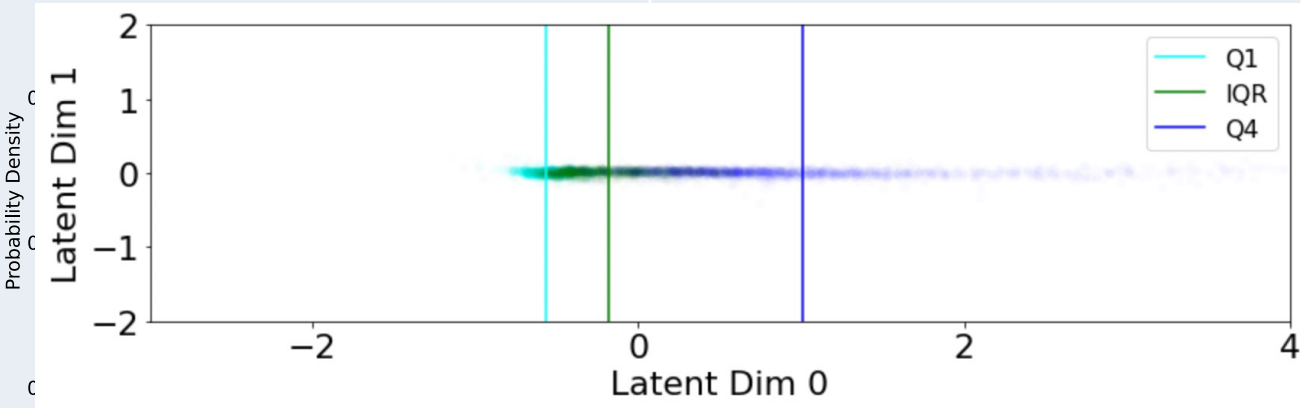
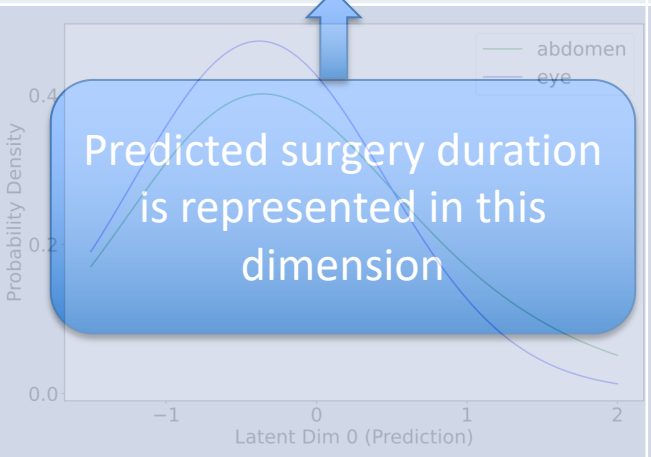
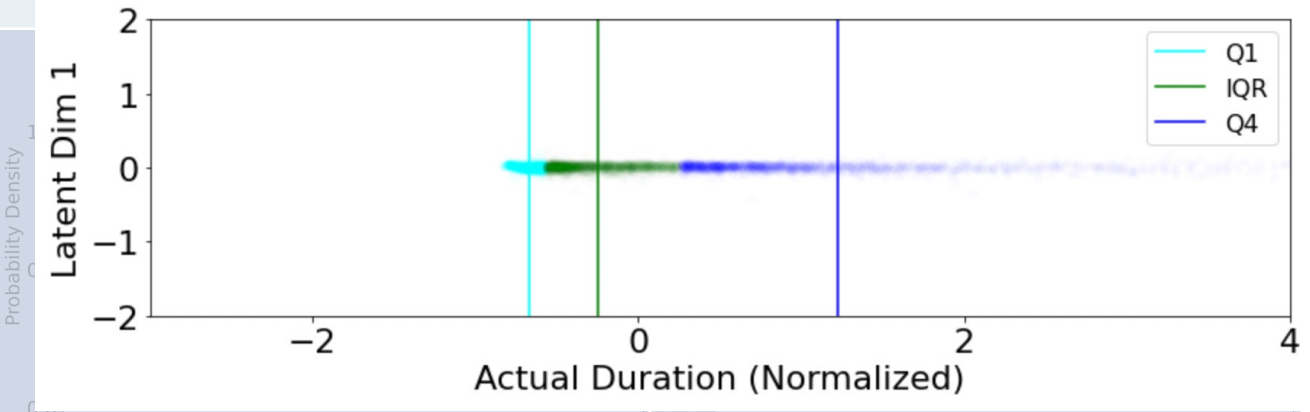
- Lower dimensionality
- Better downstream prediction
- Eliminates the needs for a predictor

Table: Prediction performance for postoperative delirium (5-fold CV)

Transformation Method (d=10)	Direct Prediction		LR		XGBoost		SVM		DNN	
	ROC AUC	Average Precision	ROC AUC	Average Precision	ROC AUC	Average Precision	ROC AUC	Average Precision	ROC AUC	Average Precision
PCA	-	-	.717(.009)	.739 (.015)	.706 (.013)	.728 (.011)	.630 (.071)	.652 (.077)	.722 (.008)	.746 (.015)
ICA	-	-	.747(.009)	.769 (.012)	.672 (.007)	.701 (.012)	.688 (.002)	.705 (.009)	.746 (.009)	.766 (.013)
GMM	.720(.007)	.732(.010)	-	-	-	-	-	-	-	-
AE	-	-	.643(.007)	.657 (.015)	.641 (.009)	.653 (.010)	.645 (.009)	.659 (.013)	.637 (.011)	.652 (.014)
VAE	-	-	.647(.007)	.667 (.009)	.650 (.010)	.668 (.011)	.649 (.006)	.668 (.015)	.635 (.009)	.654 (.012)
pi-VAE	-	-	.690(.014)	.714(.018)	.667(.011)	.681 (.016)	.687 (.017)	.710 (.022)	.668 (.012)	.688(.024)
$cVAE^{-P}$	-	-	.656(.010)	.672 (.007)	.655 (.011)	.672 (.006)	.657 (.009)	.671 (.008)	.638 (.015)	.656 (.013)
$cVAE^{-D}$	.760 (.010)	.776 (.015)	.760 (.010)	.778 (.015)	.758 (.010)	.773 (.014)	.760 (.010)	.777 (.013)	.760 (.010)	.778 (.014)
cVAE	**.776 (.009)**	**.794 (.015)**	**.773 (.010)**	**.790(.017)**	**.774 (.009)**	**.787(.012)**	**.769 (.010)**	**.786(.016)**	**.775 (.009)**	**.793(.016)**
Raw Data (d=562)	-	-	.758(.009)	.780 (.015)	.737 (.010)	.764 (.013)	.746 (.007)	.769 (.015)	.742 (.011)	.763 (.015)

Case study: Disentanglement

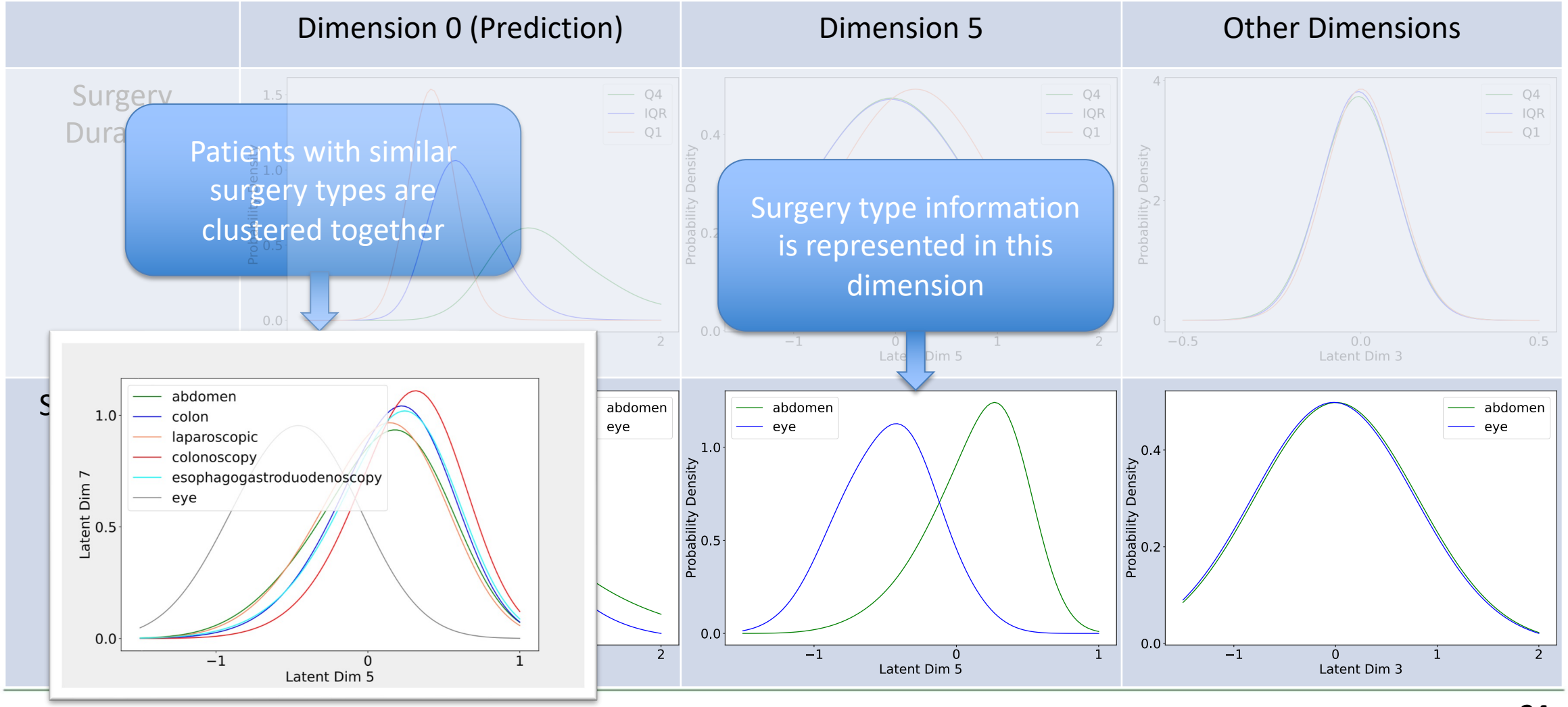
- Latent space: clustering (phenotyping) and disentanglement:

	Dimension 0 (Prediction)	Dimension 5	Other Dimensions
Surgery Duration			
Surgery type			

Predicted surgery duration is represented in this dimension

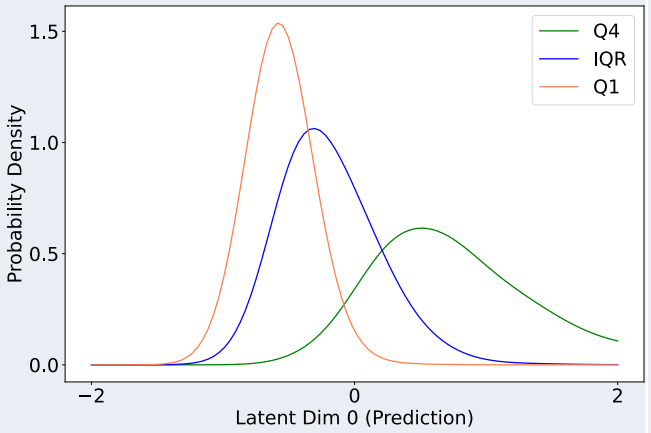
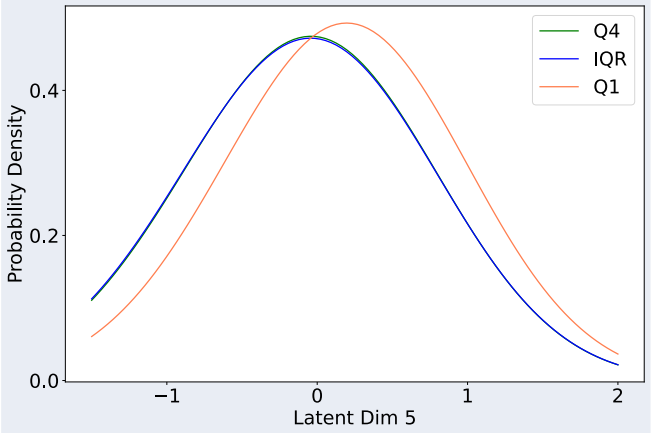
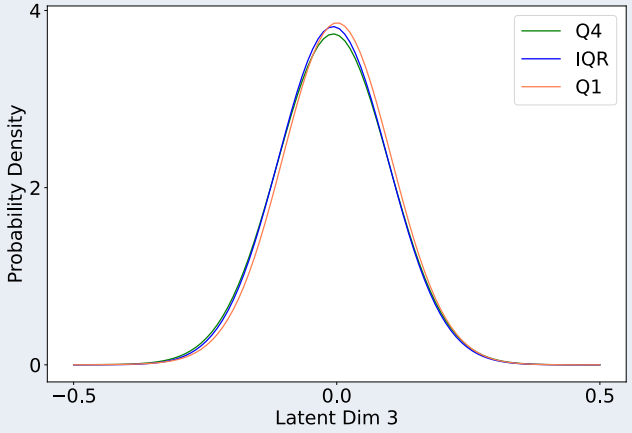
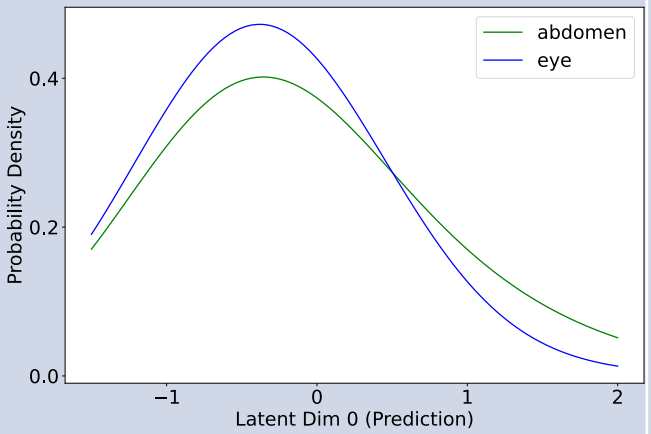
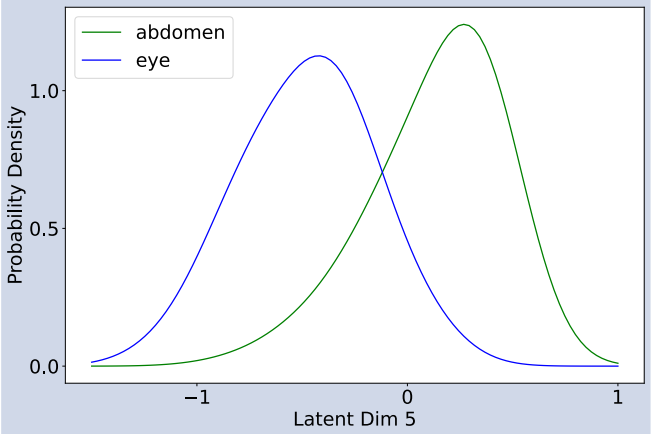
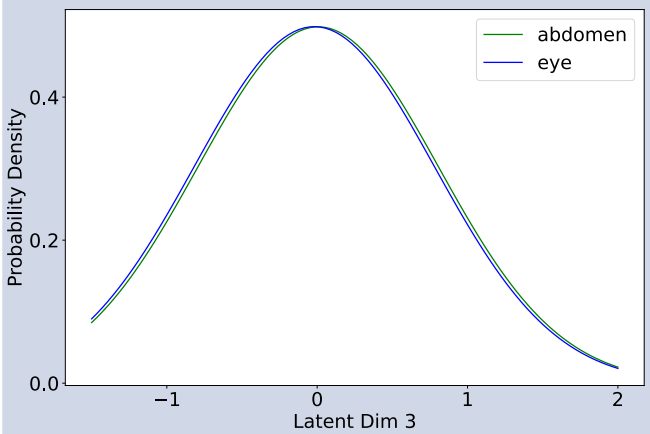
Case study: Disentanglement

- Latent space: clustering (phenotyping) and disentanglement:



Case study: Disentanglement

- Latent space: clustering (phenotyping) and disentanglement

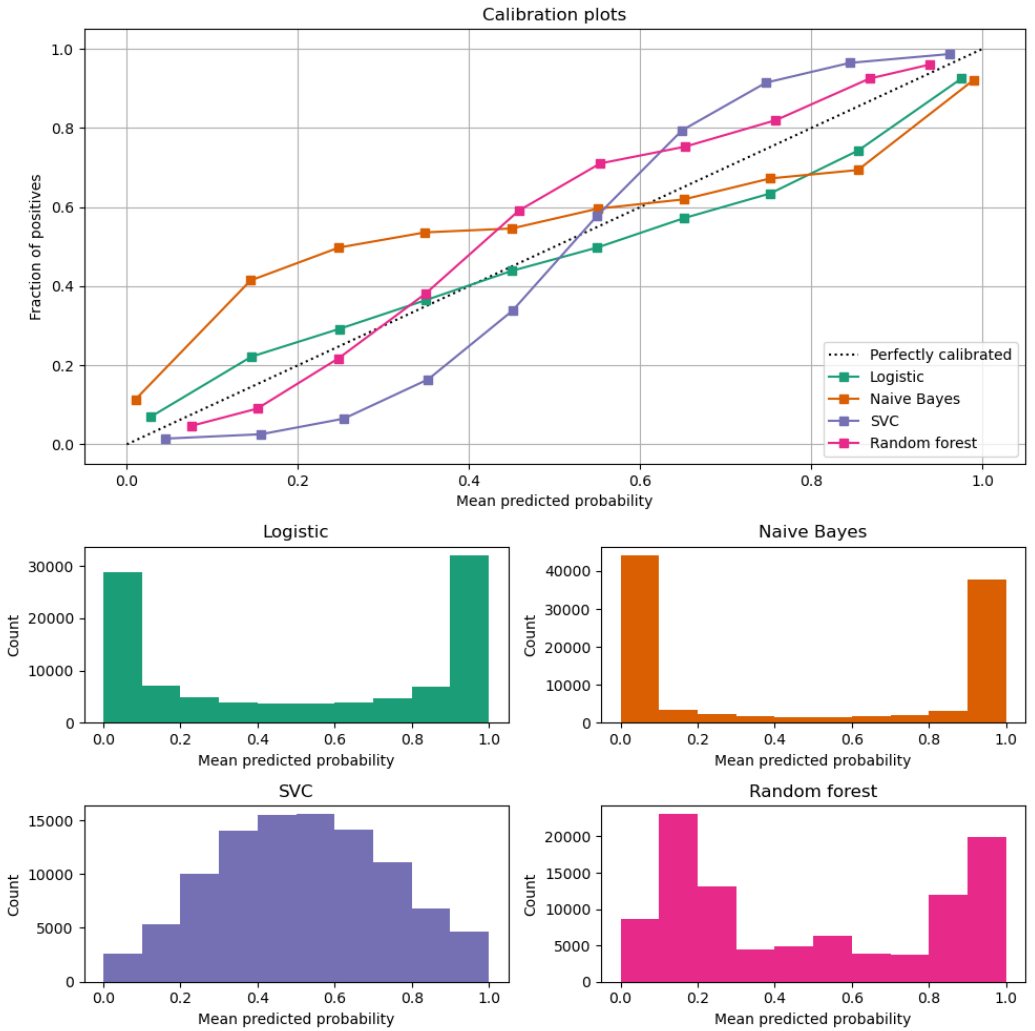
	Dimension 0 (Prediction)	Dimension 5	Other Dimensions
Surgery Duration			
Surgery type			

Open problems

- Model interpretation for OR-ICU handoff: ECHO project
  - Given a prediction, how should clinicians trust/accept the model and interpret the predicted probability?
    - A predicted probability is not a probability
    - Is the risk high?
    - Why is the risk high?
    - How should I mitigate the risks?

Open problems

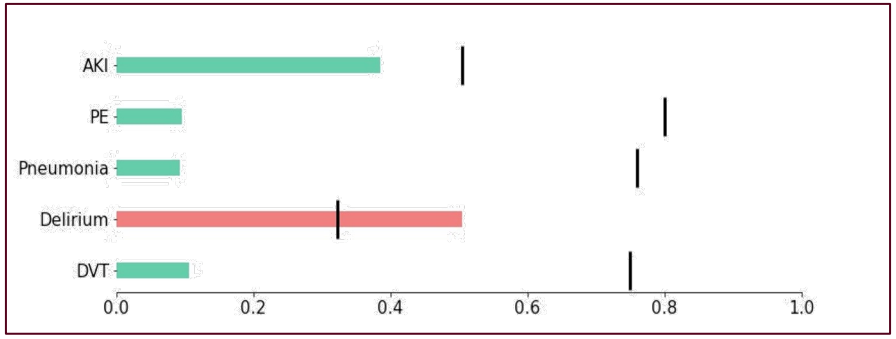
- We discussed in the previous presentation about model calibration
  - Probabilities are not comparable between ML models
  - Calibration is hard



Open problems

- Model interpretation for OR-ICU handoff: ECHO project
  - Given a prediction, how should clinicians trust/accept the model and interpret the predicted probability?
    - A predicted probability is not a probability
    - Is the risk high?
    - Why is the risk high?
    - How should I mitigate the risks?

Open problems



Postop Complications	Predicted likelihood risk for complication	Accuracy of prediction
AKI	13	78
Delirium	49	66
DVT	4	92
Pulmonary embolism	6	97
Pneumonia	10	92

C5: “You could consider showing a **confidence interval**, maybe using an **error bar** on the bar chart.” (C5, Anest)

- **Confidence interval and percentages of prediction accuracy**

- **Identifying clinically meaningful risk thresholds**

If you **set [the alert threshold] too low**, you'll get way more alerts than might be clinically present. And you'll likely get fatigued and potentially [ignore alerts]. **If you set them too high,** you make it unlikely that these would be prevented.” (C12, Anest)

Open problems

- Model interpretation for OR-ICU handoff: ECHO project
  - Given a prediction, how should clinicians trust/accept the model and interpret the predicted probability?
    - A predicted probability is not a probability
    - Is the risk high?
    - Why is the risk high?
    - How should I mitigate the risks?

Open problems

- **Color coding to differentiate risk levels**

"I kinda like red, yellow and green in terms of things that need to catch my attention. **Yellow, like meh. Green, don't worry about.**" (C2, Anest/Crit Care)

- **Risk factors grouping – decreasing and increasing risks over pre-op and intra-op risks**

"I think grouping it **according to those that increase, and those that decrease.**" (C12, Anest)



Open problems

- Model interpretation for OR-ICU handoff: ECHO project
  - Given a prediction, how should clinicians trust/accept the model and interpret the predicted probability?
    - A predicted probability is not a probability
    - Is the risk high?
    - Why is the risk high?
    - How should I mitigate the risks?

Summary

- **cVAE**: a representation learning framework for challenging clinical applications and others
  - Multi-modality
  - high dimensionality
  - performance critical
  - demanding interpretation
- Prediction-guided encoding
  - Improve predictive performance without a separate predictor
- Disentangled latent space
  - Interpretable phenotypes of patients