Assisting clinical decision making:
perioperative risk prediction

Bing Xue
CSE 521S Lecture
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_remap_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?

- 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
- Time Series Method
- Concatenated Features
- ML Predictor
### What are clinical texts?

- Clinical texts have complex abbreviations or medical terminologies
- Clinical texts can be structured or unstructured

<table>
<thead>
<tr>
<th>AN_PROC_NAME</th>
<th>AN_DATE</th>
<th>PROCED_ICD</th>
<th>PROCED_ICD_SHORT_TITLE</th>
<th>PROCED_ICD_LONG_TITLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESOPHAGOGASTRODUODENOSCOPY wu prep (N/A); CO...</td>
<td>2019-03-05 00:00:00</td>
<td>[9604, 9671, 9634]</td>
<td>['Insert endotracheal tube', 'Gastric tube irrig...</td>
<td>Insertion of endotracheal tube Other irrigatio...</td>
</tr>
<tr>
<td>BILATERAL MYRINGOTOMY TUBE INSERTION (Bilatera...</td>
<td>2019-03-14 00:00:00</td>
<td>[3613, 3615, 3961, 8872, 9904, 9905, 9907]</td>
<td>['Aortocor bypass-3 cor art', 'Int mam-cor ar...</td>
<td>Aortocoronary bypass of three coronary arterie...</td>
</tr>
<tr>
<td>FUSION SPINAL ANTERIOR LUMBAR/THORACIC WITH IN...</td>
<td>2019-03-28 00:00:00</td>
<td>[151]</td>
<td>'Ex cereb meningea les'</td>
<td>Excision of lesion or tissue of cerebral meninges</td>
</tr>
<tr>
<td>EXCHANGE STENT - URETERAL (Right Ureter)</td>
<td>2019-03-05 00:00:00</td>
<td>[3601, 3607, 3723, 8852, 8856, 9920]</td>
<td>['Ins drug-elut corony st', 'Rl/left heart ca...</td>
<td>Insertion of drugeluting coronary artery stent...</td>
</tr>
<tr>
<td>Robotic Assisted Laparoscopic Partial Fundoplici...</td>
<td>2019-03-25 00:00:00</td>
<td>[3601, 3607, 3722, 8856, 9920, 3722, 8856, 885...</td>
<td>['Ins drug-elut corony st', 'Left heart cardil...</td>
<td>Insertion of drugeluting coronary artery stent...</td>
</tr>
</tbody>
</table>
NLP Method

How to read the texts?

- **Word Embedding**: Convert words to embeddings
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 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.
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

Glove

Time Series Method

Concatenated Features

ML Predictor

Cyber-Physical Systems Laboratory
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
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. $x_1$ to $x_6$ are observed at $s_{1.6} = 0, 2, 7, 9, 14, 15$ respectively. Considering the 2nd feature in $x_6$, the last observation of the 2nd feature took place at $s_2 = 2$, and we have that $\delta_2^6 = s_6 - s_2 = 13$. 

Figure 2: Imputation with unidirectional dynamics.
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

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

Glove

Feature Engineering

Concatenated Features

ML Predictor
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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Delirium</th>
<th>OR</th>
<th>MIMIC-III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Rate</td>
<td>52.65%</td>
<td>2.50%</td>
<td>11.79%</td>
</tr>
</tbody>
</table>
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.
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)
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%)

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
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
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)

<table>
<thead>
<tr>
<th>Transformation Method (d=10)</th>
<th>Direct Prediction</th>
<th>LR</th>
<th>XGBoost</th>
<th>SVM</th>
<th>DNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ROC AUC</td>
<td>Average Precision</td>
<td>ROC AUC</td>
<td>Average Precision</td>
<td>ROC AUC</td>
</tr>
<tr>
<td>PCA</td>
<td>-</td>
<td>-</td>
<td>.717 (.009)</td>
<td>.739 (.015)</td>
<td>.706 (.013)</td>
</tr>
<tr>
<td>ICA</td>
<td>-</td>
<td>-</td>
<td>.747 (.009)</td>
<td>.769 (.012)</td>
<td>.672 (.007)</td>
</tr>
<tr>
<td>GMM</td>
<td>.720 (.007)</td>
<td>.732 (.010)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AE</td>
<td>-</td>
<td>-</td>
<td>.643 (.007)</td>
<td>.657 (.015)</td>
<td>.641 (.009)</td>
</tr>
<tr>
<td>VAE</td>
<td>-</td>
<td>-</td>
<td>.647 (.007)</td>
<td>.667 (.009)</td>
<td>.650 (.010)</td>
</tr>
<tr>
<td>pi-VAE</td>
<td>-</td>
<td>-</td>
<td>.690 (.014)</td>
<td>.714 (.018)</td>
<td>.667 (.011)</td>
</tr>
<tr>
<td>cVAE-P</td>
<td>-</td>
<td>-</td>
<td>.656 (.010)</td>
<td>.672 (.007)</td>
<td>.655 (.011)</td>
</tr>
<tr>
<td>cVAE-D</td>
<td>.760 (.010)</td>
<td>.776 (.015)</td>
<td>.760 (.010)</td>
<td>.778 (.015)</td>
<td>.758 (.010)</td>
</tr>
<tr>
<td>cVAE</td>
<td>.776 (.009)</td>
<td>.794 (.015)</td>
<td>.773 (.010)</td>
<td>.790 (.017)</td>
<td>.774 (.009)</td>
</tr>
<tr>
<td>Raw Data (d=562)</td>
<td>-</td>
<td>-</td>
<td>.758 (.009)</td>
<td>.780 (.015)</td>
<td>.737 (.010)</td>
</tr>
</tbody>
</table>
Case study: Disentanglement

- Latent space: clustering (phenotyping) and disentanglement:

<table>
<thead>
<tr>
<th>Surgery Duration</th>
<th>Dimension 0 (Prediction)</th>
<th>Dimension 5</th>
<th>Other Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgery type</td>
<td>Predicted surgery duration is represented in this dimension</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Surgery type

- Predicted surgery duration is represented in this dimension.
Case study: Disentanglement

Latent space: clustering (phenotyping) and disentanglement:

<table>
<thead>
<tr>
<th>Surgery Duration</th>
<th>Dimension 0 (Prediction)</th>
<th>Dimension 5</th>
<th>Other Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgery type</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Patients with similar surgery types are clustered together.

Surgery type information is represented in this dimension.
# Case study: Disentanglement

- **Latent space:** clustering (phenotyping) and disentanglement

<table>
<thead>
<tr>
<th>Surgery Duration</th>
<th>Dimension 0 (Prediction)</th>
<th>Dimension 5</th>
<th>Other Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="#" alt="Graph 1" /></td>
<td><img src="#" alt="Graph 2" /></td>
<td><img src="#" alt="Graph 3" /></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Surgery type</th>
<th>Dimension 0 (Prediction)</th>
<th>Dimension 5</th>
<th>Other Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="#" alt="Graph 4" /></td>
<td><img src="#" alt="Graph 5" /></td>
<td><img src="#" alt="Graph 6" /></td>
<td></td>
</tr>
</tbody>
</table>
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

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?
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