Clinical prediction on the MIMIC-III dataset

Mohammad Hashir Khan

Master's thesis

A hierarchical CNN-RNN for predicting ICU mortality using clinical notes

Part 1

- Overview of MIMIC-III dataset
- What questions to consider when creating an EHR prediction model

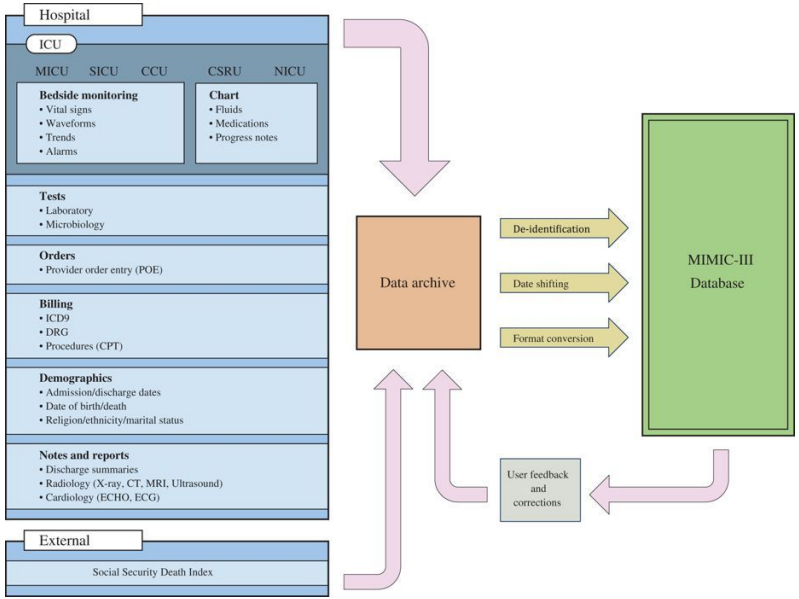
Part 2

- Existing literature on clinical prediction tasks on MIMIC III
- Towards unstructured prediction: predicting mortality on MIMIC III with notes

MIMIC III

Boston hospital from 2001-2012

- 60k admissions
  - 39k patients
  - Average age: 65.8 years
  - Average LOS: 7 days
  - 11.5% mortality rate
- 23 tables
- Multiple ICUs
  - Neonatal, coronary, trauma, cardiac, medical, surgical



Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." *Scientific data* 3 (2016): 160035

A single patient can have multiple hospital stays and a single hospital stay can have multiple ICU stays

Tables in MIMIC

Clinical data

From inside the ICU

CHARTEVENTS	Heart rate, oxygen, blood pressure, GCS scores
DATETIMEEVENTS	Time of dialysis or insertion of lines
INPUTEVENTS	IV drugs, tube feedings
NOTEEVENTS	Progress notes, discharge summary, external radiology reports
OUTPUTEVENTS	Bodily fluids
PROCEDUREEVENTS	Lumbar puncture

Other clinical data

CPTEVENTS, DRGCODES	Admin codes for diagnosis & procedures
DIAGNOSES_ICD PROCEDURES_ICD	ICD codes for diagnosis & procedures
LABEVENTS	Potassium, vitamins
MICROBIOLOGY	Culture growth, antibiotic reaction
PRESCRIPTIONS	Medication orders

Stay info

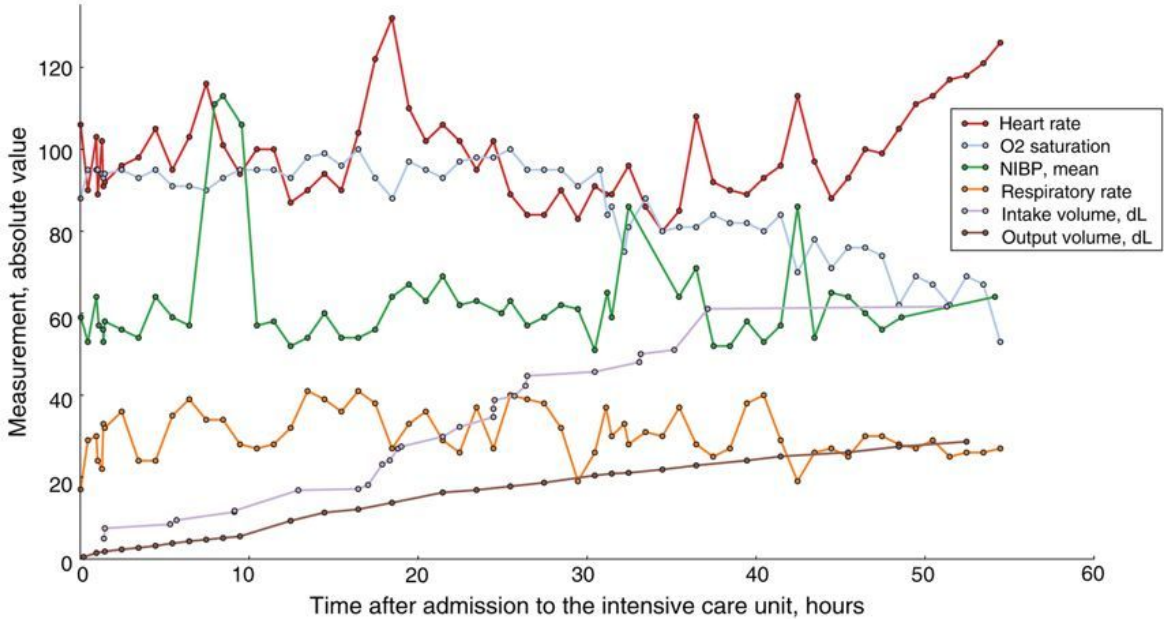
ADMISSIONS, CALLOUT, CAREGIVERS, ICUSTAYS , PATIENTS, SERVICES, TRANSFERS

Externally derived/calculated tables

Severity scores, comorbidities, sepsis, height/weight

Example of patient trajectory

Code status	Full code					Comfort measures
GCS: Verbal GCS: Motor GCS: Eye	Oriented Obeys commands Spontaneously	Oriented Obeys commands Spontaneously	Oriented Obeys commands Spontaneously	Confused Obeys commands To speech	Confused Obeys commands To speech	Incomprehensible sounds Flex-withdraws None
Platelet, K/uL Creatinine, mg/dL White blood cell, K/uL Neutrophil, %	48 53 0.7 9.1 12.4 37	46 0.7 16.8			45 0.8 23.2	
Morphine Sulfate Vancomycin (1 dose) Piperacillin (1 dose) NaCl 0.9% Amiodarone Dextrose 5%			10.0 mL/hour 1mg/min 50mL/hour	10.0mL/hour 0.5mg/min 25mL/hour	0.5mg/min 25mL/hour	10.0mL/hour



Creating models for the EHR

1. Define problem and select cohort
2. Extract and preprocess data
3. Create model and evaluate

Benchmark papers

1. Purushotham, Sanjay, et al. “Benchmarking Deep Learning Models on Large Healthcare Datasets.” *Journal of Biomedical Informatics*, vol. 83, July 2018, pp. 112–34.
2. Harutyunyan, Hrayr, et al. “Multitask Learning and Benchmarking with Clinical Time Series Data.” *arXiv [stat.ML]*, 22 Mar. 2017, http://arxiv.org/abs/1703.07771. arXiv.
3. Johnson, Alistair E. W., and Roger G. Mark. “Real-Time Mortality Prediction in the Intensive Care Unit.” Age , vol. 63, lcp.mit.edu, 2017, pp. 17–12.
4. Guzman, Uri Shalit, and David Sontag. "An Open Benchmark for Causal Inference Using the MIMIC-III Dataset." (preliminary)

Problem definition

What exactly am I going to predict with what and for who?

MIMIC Paper Generator

I am going to predict ___(column A)___ using ___(column B)___ for patients with ___(column C)___

Column A	**Column B**	**Column C**
ICD Codes	Demographics	No disease criteria
Length of stay	Clinical vitals	Pneumonia
Readmission	ICD codes	Cancer
Mortality	Billing codes	Kidney failure
Early warning scores	Prescriptions	Diabetes
Decompensation	Notes	Heart disease
Prescriptions		

Things to consider when formulating task: outcome

Outcome variables tradeoff

- Popular -> easy to find literature
  - Mortality, readmission, length of stay etc.
  - But loosely defined, e.g. in-hospital vs post-discharge mortality, ICU vs hospital LOS
- Uncommon could be more interesting, clinically relevant and standardized
  - invasive ventilation, non-invasive ventilation, vasopressors, colloid boluses, and crystalloid boluses (Suresh et al, 2017)

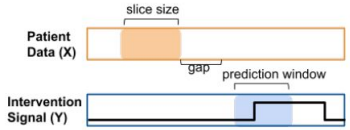


Figure 2: Given data from a fixed-length (6 hour) sliding window, models predict the status of intervention in a prediction window (4 hours) after a gap time (6 hours). Windows slide along the entire patient record, creating multiple examples from each record.

Suresh, Harini, et al. “Clinical Intervention Prediction and Understanding with Deep Neural Networks.” *Proceedings of the 2nd Machine Learning for Healthcare Conference*, edited by Finale Doshi-Velez et al., vol. 68, PMLR, 2017, pp. 322–37.

Things to consider when formulating task: input

Depends on nature of study

- *Post-hoc:* Retrospective, after patient has been discharged
- *Dynamic:* Prediction done during patient stay, eg 24 hours after admission

Input data should make sense for chosen outcome + goal + nature of study

- What if data chosen is usually collected towards end of stay?
- But could be useful for retrospective exploratory study?

Would you even have the data available during the hypothetical situation?

Things to consider when formulating task: phenotypes

Should you look at a certain subset wrt diagnosis?

If certain subset, what criteria?

- ICD/DRG codes?
- Initial assessment by admitting physician?

Most recent papers are disease agnostic

- Model could be good for easy/salient phenotypes?
- Suresh et al (2018): MICU AUC lower than avg AUC across ICU's

Suresh, Harini, Jen J. Gong, and John V. Guttag. "Learning tasks for multitask learning: Heterogenous patient populations in the icu." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018.

Cohort selection

Easy: age (don’t mix pediatric and adults), mortality after N hours

Tricky: multiple hospital stays, DNR, certain phenotypes

For e.g., multiple stays

- Different stays of same person in train & test -> information leakage?

OR

- Different stays possibly represent different phenotypes -> some independence?

Data preprocessing

There is no easy and standard formula to preprocess clinical data

Data collection window

Data up to N hours before EVENT

- Not possible in real life
- Sherman et al (2017) show performance is misleadingly good

Hence, reference point should be **<u>outcome independent</u>**

Example: data from the first N hours after ICU in-time

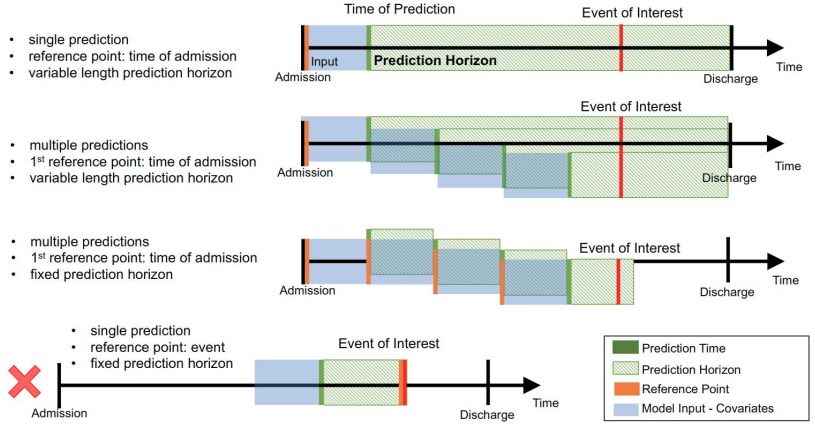


Figure 1: Sample time series' depicting the various ways that data sets can be extracted. The first three correspond to varying numbers of predictions per outcome measurement. The final one uses the event of interest as the reference point which we note should *not* be used for model evaluation since it does not mirror any clinical use case.

Sherman, Eli, et al. "Leveraging Clinical Time-Series Data for Prediction: A Cautionary Tale." *AMIA... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, vol. 2017, ncbi.nlm.nih.gov, 2017, pp. 1571–80.

But what if multiple reference points?

What about data from between admission and ICU in-time or even before admission?

Time series acquisition rates

Temporal data have different rates/irregularly sampled

How to sync up periods? Eg: BP recorded every hour and PaO2 every 4 hours

- Summarize BP over 4 hours?
- Resample PaO2 to every hour? Impute forward/backward/healthy?

Zheng et al (2017) -

EMR regularization: transform biased, irregular multivariate time series data into unbiased, regular ones

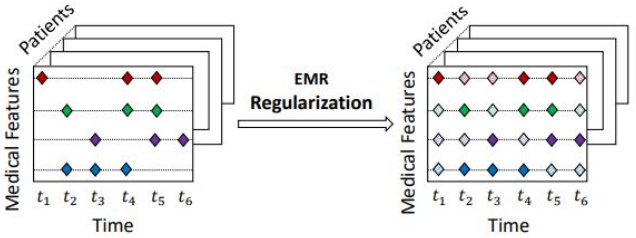


Figure 3: EMR regularization.

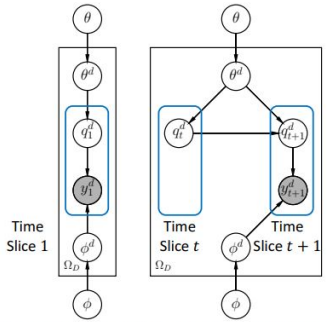


Figure 4: Graphical representation of proposed model.

Zheng, Kaiping, et al. "Resolving the bias in electronic medical records." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.

Missing data

Missing completely at random

No systematic differences between missing and observed values.

- BP missing because of breakdown of an automatic sphygmomanometer

Missing at random

Any systematic difference between missing and observed values can be explained by differences in observed data.

- Missing BP < measured BP
  - But only because younger people more likely to have missing BP measurements

Missing not at random

Even after observed data taken into account, systematic differences remain between missing and observed values.

- People with high BP more likely to miss clinic appts because they have headaches

Sterne Jonathan A C, White Ian R, Carlin John B, Spratt Michael, Royston Patrick, Kenward Michael G et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls BMJ 2009; 338 :b2393

Learn missingness or generate imputations

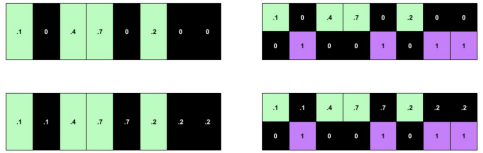


Figure 2: (top left) no imputation or indicators, (bottom left) imputation absent indicators, (top right) indicators but no imputation, (bottom right) indicators and imputation. Time flows from left to right.

Lipton, Zachary C., et al. “Directly Modeling Missing Data in Sequences with RNNs: Improved Classification of Clinical Time Series.” *Proceedings of the 1st Machine Learning for Healthcare Conference*

$\boldsymbol{X}$: Input time series (2 variables);
$\boldsymbol{s}$: Timestamps for $\boldsymbol{X}$;

$$\boldsymbol{X} = \begin{bmatrix} 47 & 49 & NA & 40 & NA & 43 & 55 \\ NA & 15 & 14 & NA & NA & NA & 15 \end{bmatrix}$$

$$\boldsymbol{s} = [\,0 \quad 0.1 \quad 0.6 \quad 1.6 \quad 2.2 \quad 2.5 \quad 3.1]$$

$\boldsymbol{M}$: Masking for $\boldsymbol{X}$;
$\boldsymbol{\Delta}$: Time interval for $\boldsymbol{X}$.

$$\boldsymbol{M} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{\Delta} = \begin{bmatrix} 0.0 & 0.1 & 0.5 & 1.5 & 0.6 & 0.9 & 0.6 \\ 0.0 & 0.1 & 0.5 & 1.0 & 1.6 & 1.9 & 2.5 \end{bmatrix}$$

Figure 2: An example of measurement vectors $\boldsymbol{x}_t$, time stamps s_t, masking $\boldsymbol{m}_t$, and time interval $\boldsymbol{\delta}_t$.

Che, Zhengping, et al. “Recurrent Neural Networks for Multivariate Time Series with Missing Values.” *arXiv [cs.LG]*, 6 June 2016, http://arxiv.org/abs/1606.01865. arXiv.

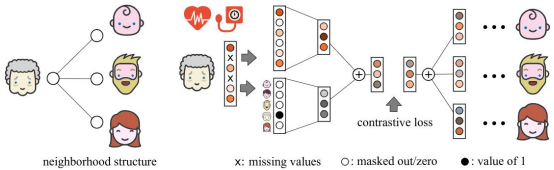


Figure 2: Illustration of the embedding propagation framework for missing data.

Malone, Brandon, et al. “Learning Representations of Missing Data for Predicting Patient Outcomes.” arXiv [cs.LG], 12 Nov. 2018, http://arxiv.org/abs/1811.04752. arXiv.

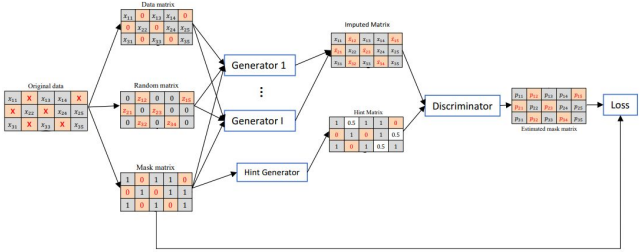


Figure 3: Architecture of matrix imputation by Stackelberg GAN.

Zhang, Hongyang, and David P. Woodruff. "Medical Missing Data Imputation by Stackelberg GAN."

Key-value approach

Instead of `[time samples, features]`, maybe try `[data type, value]`, ordered by time of collection

E.g. - $[[\text{“Heart rate”}, 65]_{t1}, [\text{“Glucose”}, 70]_{t2}, \ldots, [\text{“Arrhythmia”}, \text{Tachycardia}]_{tN}]$

Key: one-hot vector or a dense representation

Event type embedding and attribute encoding

To help the HE-LSTM to trace temporal information of various kinds of events, we use “event type embedding” and “attribute encoding” to embed the type and attributes of the high dimensional events into compact continuous vectors, which can be trained end-to-end with the following HE-LSTM.

An event $e_i = (type, value, time)$ of the sequence will be embedded into three parts to feed the HE-LSTM for the endpoint prediction. The three input including embedding vector of event type $\boldsymbol{s}$, the event attribute encoding vector $\boldsymbol{x}$ and the scale variable time t.

The event type vector $\boldsymbol{s}$ carries the information of the event category of $\boldsymbol{e}_i$, and is constructed only by the one hot representation $\boldsymbol{type}$ of event type. Similar to word embedding (Mikolov et al. 2013), it will provide a low-dimension vector of the event type with semantic meaning in clinical field. The embedding lookup matrix $C_{type} \in \mathbb{R}^{N \times M}$, where N is the embedding dimension and M is the number of event types, is established for further training. The event type vector $\boldsymbol{s}$ is given by:

$$\boldsymbol{s} = C_{type} \times \boldsymbol{type} \quad (1)$$

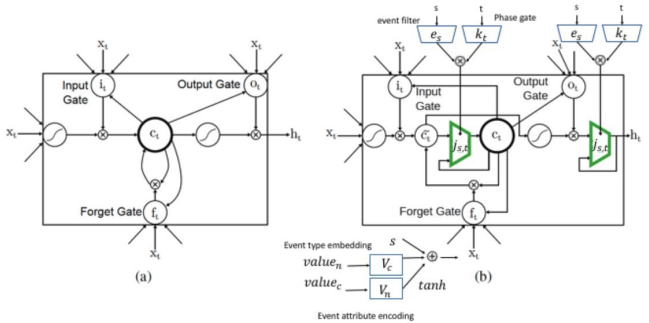


Figure 2: Model architecture. (a) Standard LSTM model. (b) HE-LSTM model, with event gate $\boldsymbol{j}_t$ consist of the event filter $\boldsymbol{e}_s$ and phase gate $\boldsymbol{k}_t$ separately controlled by the event type $\boldsymbol{s}$ and timestamp t. In the HE-LSTM formulation, each neural in the cell value c_t and the hidden output h_t can be updated during an “open” phase by only some certain types of events; otherwise, the previous values are maintained.

Liu, Luchen, et al. “Learning the Joint Representation of Heterogeneous Temporal Events for Clinical Endpoint Prediction.” *arXiv [cs.AI]*, 13 Mar. 2018, http://arxiv.org/abs/1803.04837. arXiv.

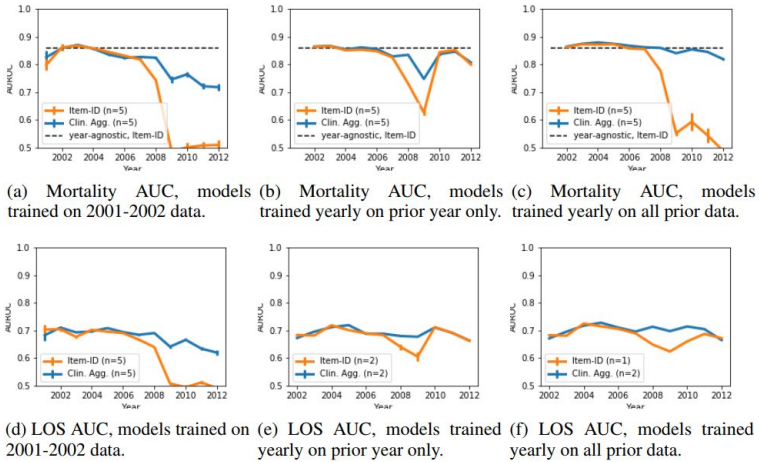
Modeling & evaluation

Simpler models might just work if the data is preprocessed well

Evaluate carefully

- AUROC or AUPRC?
  a. Saito, Takaya, and Marc Rehmsmeier. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PloS One*, vol. 10, no. 3, Mar. 2015, p. e0118432.
  b. Davis, Jesse, and Mark Goadrich. “The Relationship Between Precision-Recall and ROC Curves.” *Proceedings of the 23rd International Conference on Machine Learning*, ACM, 2006, pp. 233–40.

- Inherent issue with MIMIC III: data from early years performs better than later because of system change and date randomisation
  - Nestor, Bret, et al. “Rethinking Clinical Prediction: Why Machine Learning Must Consider Year of Care and Feature Aggregation.” *arXiv [cs.LG]*, 30 Nov. 2018, http://arxiv.org/abs/1811.12583. arXiv.



(a) Mortality AUC, models trained on 2001-2002 data. (b) Mortality AUC, models trained yearly on prior year only. (c) Mortality AUC, models trained yearly on all prior data.

(d) LOS AUC, models trained on 2001-2002 data. (e) LOS AUC, models trained yearly on prior year only. (f) LOS AUC, models trained yearly on all prior data.

Figure 1: Performance of RF classifiers using `Item-Id` and `Clinically Aggregated` representations on mortality (top) and LOS prediction (bottom). Error bars indicate ± standard error.

Reproducibility

- Make sure cohort is easy to reproduce,
  - Just publishing code to do so might not be enough
- Johnson et al (2017) found a disparity in trying to reproduce 38 papers

> We reproduced datasets for 38 experiments corresponding to 28 published studies using MIMIC. In half of the experiments, the sample size we acquired was 25% greater or smaller than the sample size reported. The highest discrepancy was 11,767 patients. While accurate reproduction of each study cannot be guaranteed, we believe that these results highlight the need for more consistent reporting of model design and methodology to allow performance improvements to be compared. We discuss the challenges in reproducing the cohorts used in the studies, highlighting the importance of clearly reported methods (e.g. data cleansing, variable selection, cohort selection) and the need for open code and publicly available benchmarks.

Johnson, Alistair E. W., et al. “Reproducibility in Critical Care: A Mortality Prediction Case Study.” *Proceedings of the 2nd Machine Learning for Healthcare Conference*, edited by Finale Doshi-Velez et al., vol. 68, PMLR, 2017, pp. 361–76.

Some other interesting reads:

- Coiera, Enrico, et al. "Does health informatics have a replication crisis?." *Journal of the American Medical Informatics Association* 25.8 (2018): 963-968.
- McDermott, Matthew, et al. "Reproducibility in Machine Learning for Health." *arXiv preprint arXiv:1907.01463* (2019).

Part 2 Preview

Existing literature on MIMIC III

1. Outcome prediction and benchmarks
2. Notes and other unstructured data
    a. Why notes over structured physiological variables?
3. Causal inference
4. Generative models
5. Transfer learning
6. Sociology/intersectionality

Thank you