Project Proposal

Paper 1: 69, *Dynamic Survival Transformers for Causal Inference with Electronic Health Records*, NeurIPS TS4H, Prayag Chatha, Yixin Wang, Zhenke Wu, Jeffrey Regier

1) Task: Utilize a transformer to model the effects of treatment on a patient's probability of survival as it changes with time.

2) Innovation: Introduction of a transformer to recalculate the probability of failure for patients for each time interval throughout the patient's history.

3) Adv/Dis: Patient survival probabilities are modeled at each timestep, independent of data from future timesteps, allowing the prediction to change over time. Timesteps correspond to exactly one hour of time.

4) Dataset and Accessibility: MIMIC III Dataset utilizing patient stays longer than 16 hours (inclusive) and limited to the first 128 hours of length. Exactly 30323 patient records in length. https://physionet.org/content/mimiciii/1.4/

5) Code Accessibility: https://github.com/prob-ml/DynST?utm_source=chatgpt.com

6) Feasibility: The dataset used in this paper is MIMIC-III, which will be available and accessible to us. For data preprocessing, the authors utilized MIMIC-Extract, a pipeline that is also publicly available. To effectively replicate their methodology, we will need to understand how they configured and applied this pipeline. Regarding computational resources, the authors specified using an NVIDIA GTX 2080 Ti GPU. Since we have multiple computers equipped with NVIDIA GPUs, we can match their computational power, ensuring that hardware constraints do not pose a major challenge. The most challenging aspect of replicating this paper will be the integration of static and dynamic patient data using transformer layers. While the authors have made their codebase publicly available, successful replication will still require experimentation with different hyperparameters to optimize model performance and validate reproducibility.

Paper 2: 46, *Improving clinical outcome predictions using convolution over medical entities with multimodal learning*, Artificial Intelligence in Medicine, Batuhan Bardak, Mehmet Tan

1) Task: Enhance the prediction of patient mortality and ICU length of stay by introducing clinical notes to the already existing structured dataset for patients.

2) Innovation: Utilizes the med7 named entity recognition model to extract diagnosis, drug dosage and frequency, and other medical information from unstructured clinical notes.

3) Adv/Dis: Clinical notes tend to be difficult to process for meaning. Dataset is imbalanced, with a significant portion of patients surviving and having short ICU stays.

4) Dataset and Accessibility: MIMIC III Dataset utilizing patient stays between 12 hours and 10 days (inclusive) for patients older than 15 years of age with clinical notes within the first 24 hours of stay, excluding discharge summaries and clinical notes without a timestamp. https://physionet.org/content/mimiciii/1.4/

5) Code Accessibility: https://github.com/tanlab/ConvolutionMedicalNer

6) Feasibility: The dataset used in this paper is MIMIC-III, which will be available and accessible to us. The authors employed a data pipeline for extracting and preprocessing the medical records, utilizing Med7, a publicly available NLP tool for medical entity recognition. However, we will need to familiarize ourselves with its functionality before integrating it into our workflow.

The paper does not specify the exact computational resources used. However, given the complexity of deep learning models, particularly convolutional architectures, it is reasonable to assume that a high-performance GPU will be required for efficient training.
The most challenging aspect of replicating this study would be designing and optimizing the convolutional layers to effectively process medical records. Tuning the model parameters and preprocessing pipelines for structured and unstructured medical data will also require careful experimentation.

Paper 3: 30, *Towards automated clinical coding*, International Journal of Medical Informatics, Finneas Catling, Georgios P. Spithourakis, Sebastian Riedel
1) Task: Develop a method to automate the process of turning clinical notes into structured clinical code.
2) Innovation: Broadens the scope of current automation studies to include more than just common diagnoses. Trains discharge notes (with known diagnoses) on their symptoms presented to create a model to predict diagnoses upon patient admission, before it has been confirmed.
3) Adv/Dis: Clinical notes generally lack structure, are full or errors, and medical coding is fuzzy (there are many synonyms and codes used to denote the same concept). Clinical notes are, by nature, subjective. Incredibly rare diagnoses sometimes lack specific medical codes to identify them. Predictive models are prone to systematic errors caused on the institution level.
4) Dataset and Accessibility: The free-text discharge summaries and associated ICD-9-CM codes from the MIMIC III Dataset v1.4, excluding those with errors or no associated code.
https://physionet.org/content/mimiciii/1.4/
5) Code Accessibility: The codebase for this research is not provided.
6) Feasibility: The dataset for this paper would be easy to obtain, as it is the MIMIC-III dataset. Additionally, multiple named entity recognition models could be implemented and tested for reading the clinical notes. The difficulty with this paper would be improving upon the research done there-in, as we would be using the same dataset, if not a similar dataset from the same institution, prone to the same disadvantages outlined above.

Target Paper: Paper 1: 69, *Dynamic Survival Transformers for Causal Inference with Electronic Health Records*, NeurIPS TS4H, Prayag Chatha, Yixin Wang, Zhenke Wu, Jeffrey Regier
1) Why this paper: Both team members are interested in the paper's use of transformers to train the electronic health records as a function of time.
2) Hypotheses to Verify: Utilizing transformers allowed the model to improve upon baseline models for predicting patient survivability by updating the probability at each unit in time.
3) Methods: We are certain that the dataset used in the study is available, as it is utilizing the MIMIC-III dataset and the filtering data is provided. Our computers have at least half of the power of the proposed GPU utilized in the paper. Alternatively, we may leverage Google Colab compute nodes should either computer have troubles in training or testing the dataset.

Appendix

1. MIMIC-Extract:

   MIMIC-Extract is a publicly available data extraction and preprocessing pipeline designed to streamline the use of the MIMIC-III clinical database for machine learning applications. Developed to address the challenges of working with raw electronic health records (EHRs), MIMIC-Extract provides a structured framework for processing patient data into a format suitable for predictive modeling.

   Key Features:

   - Standardized Data Representation: Transforms raw ICU patient records into a uniform, time-aligned dataset for easier analysis.
   - Predefined Clinical Variables: Extracts vital signs, lab measurements, demographics, and treatment indicators from the MIMIC-III database.
   - Temporal Data Structuring: Organizes time-series data into fixed-hourly intervals to facilitate longitudinal analysis.
   - Open-Source Implementation: Available on GitHub (https://github.com/MLforHealth/MIMIC_Extract), allowing for modifications and extensions based on research needs.

   Relevance to This Study:

   In our work, MIMIC-Extract is utilized to preprocess patient records, ensuring a structured and consistent input format for machine learning models. The extracted features serve as the foundation for training predictive models, including those based on transformers and survival analysis techniques. Understanding and adapting MIMIC-Extract is essential for maintaining data integrity and replicability in clinical outcome predictions.

2. Propensity Score: The odds of a patient receiving a treatment dependent on their observed characteristics. In this study, the propensity score is used to calculate a synthetic binary treatment variable, A (whether the patient receives treatment). Denoted by whether a patient is severely ill, Z_*, which is 1 when a patient has been diagnosed with more than one of the following conditions: hypertension, coronary atherosclerosis, and atrial fibrillation. Propensity score is defined as:

$$P(A = 1 \mid X) \equiv \pi(X_i) = \begin{cases} 0.8, if\ Z_* = 1 \\ 0.2, if\ Z_* = 0 \end{cases}$$

3. Hazard Function: The risk of death at an instant in time given that a patient has survived up to that point in time. Limited to the range $(10e^{-8}, 0.1)$ and defined as the product of the baseline hazard (a), the treatment effect (b), static variables (c), the temporal interaction term (d), and dynamic variables (e), such that:

$$h(t) = \underset{(a)}{H_0\, e^{-\lambda t}} \cdot \underset{(b)}{e^{\theta A}} \cdot \underset{(c)}{e^{(\sum_{j=1}^{4} \beta_j Z_j)}} \cdot \underset{(d)}{e^{\log(1.02) t Z_*}} \cdot \underset{(e)}{e^{\left(\sum_{j=1}^{4} \gamma_j g(V_j^t)\right)}}$$

Where:

a. Let $H_0 = 0.001$ and $\lambda = 0.25$, such that a patient's risk decreases the longer they survive.

b. Let A be the binary treatment variable and $\theta = -0.5$, such that treatment intervention reduces a patient's risk.

c. Let $Z_1 \ldots Z_4$ be the binary variables for whether the patient is male, and if they have been diagnosed with hypertension, coronary atherosclerosis, and atrial fibrillation respectively. Let β_j be coefficients generated with a Uniform(0.7,1.2) distribution.

d. Let Z_* be the indicator variable for severely ill patients as denoted in the propensity score function. When $Z_* = 1$ the hazard increases by 2% each time step to replicate a severely ill patient's deteriorating health.

e. Let γ_j be coefficients generated with a Uniform(0.1,0.3) distribution. Let $V_1 \ldots V_4$ be the vital readings for hematocrit, hemoglobin, platelets, and mean blood pressure, clipped to avoid inflated hazards such that

$$g(V_j^t) = \begin{cases} 0, & if\ V_j^t \geq 0 \\ max\left\{\left(V_j^t\right)^2, 3\right\}, & else \end{cases}$$

4. Estimated Survival Probability:

$$\widehat{S}_i(t) = \prod_{\tau}^{t} \hat{q}(t; Z_i, \overline{V_i^t})$$

Where τ is the cutoff timestep, $\overline{V_i^t}$ is the time series vector for dynamic variables to time t, and $\hat{q}\left(t; Z_i, \overline{V_i^t}\right)$ is the compliment of the hazard function, such that $\hat{q}\left(t; Z_i, \overline{V_i^t}\right) = 1 - h(t; Z_i, \overline{V_i^t})$

5. Predicted Survival Time:

$$\hat{T}_i = \sum_{t=1}^{t_{max}} \widehat{S}_i(t)$$

6. Loss Function: The difference between the model's predicted output and the actual target value.

a. Maximizing survival probabilities up to failure time for uncensored patients.

$$\mathcal{L}_1^i = \left[\sum_{t=1}^{O_i - 1} \log \widehat{S}_i(t) + \sum_{t=O_i}^{t_{max}} \log\left(1 - \widehat{S}_i(t)\right)\right] * \delta_i - \left[\sum_{t=1}^{O_i} \log \widehat{S}_i(t)\right] * (1 - \delta_i)$$

b. Penalizing error of predicted survival time such that for censored patients, only predicted times before censoring are penalized.

$$\mathcal{L}_2^i = \left|O_i - \hat{T}_i\right| * \delta_i + \max\{0, O_i - \hat{T}_i\} * (1 - \delta_i)$$

c. Summed over all patients

$$\mathcal{L} = \sum_{i=1}^{n} (1 - \alpha)\mathcal{L}_1^i + \alpha \mathcal{L}_2^i$$

Where α is the tuning hyperparameter.

7. True Average Treatment Effect: Defined on Restricted Mean Survival Time (RMST) from two copies of the dataset, one in which every patient receives treatments ($A_i = 1$) and one in which no patients receive treatment ($A_i = 0$).

$$\psi \approx \frac{1}{n} \sum_{i=1}^{n} (Y_{i,\tau}(1, X_i) - Y_{i,\tau}(0, X_i))$$

8. Mean Absolute Error:

$$C_{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left[\left| O_i - \hat{T}_i \right| * \delta_i + \max\{0, O_i - \hat{T}_i\} * (1 - \delta_i) \right]$$

Where O_i is the observed survival time or censor time, $\hat{T}_i$ is the predicted survival time, and δ_i is a binary indicator for if the patient record is censored (for patient records longer than 128 hours)

Note: Our team chose not to use any LLMs during the project proposal.

References

- *Dynamic Survival Transformers for Causal Inference with Electronic Health Records*, NeurIPS TS4H, Prayag Chatha, Yixin Wang, Zhenke Wu, Jeffrey Regier
    - **Kaplan, E. L., & Meier, P. (1958).** *Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association*, 53(282), 457–481.
    - **Cox, D. R. (1972).** *Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187–202.
    - **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, & Polosukhin, I. (2017).** *Attention Is All You Need. Advances in Neural Information Processing Systems*, 30.
    - **Johnson, A. E. W., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016).** *MIMIC-III, a Freely Accessible Critical Care Database. Scientific Data*, 3, 160035.
    - **Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018).** *DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network. BMC Medical Research Methodology*, 18(1), 24.
    - **Lee, C., Zame, W. R., Yoon, J., & van der Schaar, M. (2019).** *Temporal Quilting for Survival Analysis. Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 32(1).
    - **Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019).** *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, 4171–4186.
    - **Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982).** *Evaluating the Yield of Medical Tests. JAMA*, 247(18), 2543–2546.
    - **Kingma, D. P., & Ba, J. (2015).** *Adam: A Method for Stochastic Optimization. International Conference on Learning Representations*.
    - **Ishwaran, H., & Kogalur, U. B. (2007).** *Random Survival Forests for R. R News*, 7(2), 25–31.
    -
- *Improving clinical outcome predictions using convolution over medical entities with multimodal learning*, Artificial Intelligence in Medicine, Batuhan Bardak, Mehmet Tan
    - **Johnson, A. E. W., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016).** *MIMIC-III, a Freely Accessible Critical Care Database. Scientific Data*, 3, 160035.
    - **Kormilitzin, A., Vaci, N., Liu, Q., & Nevado-Holgado, A. (2020).** *Med7: A Transfer Learning Framework for Clinical Natural Language Processing. arXiv preprint arXiv:2003.01271*.
    - **Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013).** *Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781*.
    - **Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017).** *Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics*, 5, 135–146.

    - **Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016).** *Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. Machine Learning for Healthcare Conference*, 301–318.
    - **Suresh, H., Hunt, N., Johnson, A., Celi, L. A., Szolovits, P., & Ghassemi, M. (2017).** *Clinical Intervention Prediction and Understanding with Deep Neural Networks. Machine Learning for Healthcare Conference*, 322–337.
    - **Hochreiter, S., & Schmidhuber, J. (1997).** *Long Short-Term Memory. Neural Computation*, 9(8), 1735–1780.
    - **Kingma, D. P., & Ba, J. (2015).** *Adam: A Method for Stochastic Optimization. International Conference on Learning Representations*.
    - **He, K., Zhang, X., Ren, S., & Sun, J. (2016).** *Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

- *Towards automated clinical coding*, International Journal of Medical Informatics, Finneas Catling, Georgios P. Spithourakis, Sebastian Riedel
    - **Perotte, A., Pivovarov, R., Natarajan, K., Weiskopf, N., Wood, F., & Elhadad, N. (2014).** *Diagnosis code assignment: models and evaluation metrics. Journal of the American Medical Informatics Association*, 21(2), 231–237.
    - **Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., & Elhadad, N. (2018).** *Multi-label classification of patient notes: case study on ICD code assignment. arXiv preprint arXiv:1709.09587.*
    - ***Shi, X., Wang, H., Peng, Y., Lu, L., & Yan, R. (2017).*** *Towards automated ICD coding using deep learning. arXiv preprint arXiv:1711.04075.*
    - ***Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., & Eisenstein, J. (2018).*** *Explainable prediction of medical codes from clinical text. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 1101–1111.*
    - ***Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016).*** *Doctor AI: Predicting clinical events via recurrent neural networks. Machine Learning for Healthcare Conference, 301–318.*
    - ***Johnson, A. E. W., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016).*** *MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.*
    - ***Hochreiter, S., & Schmidhuber, J. (1997).*** *Long short-term memory. Neural Computation, 9(8), 1735–1780.*
    - ***Kingma, D. P., & Ba, J. (2015).*** *Adam: A method for stochastic optimization. International Conference on Learning Representations.*
    - ***LeCun, Y., Bengio, Y., & Hinton, G. (2015).*** *Deep learning. Nature, 521(7553), 436–444.*
    - ***Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R. (2016).*** *Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:1511.03677.*