# Team B4 Project Proposal

Team Members: Devin Cortese, Ethan Tran
CSE-6250: Big Data Health, Fall 2023

Preferred Papers:

**Paper 1: 39, "Learning Patient Representation From Text", NAACL HLT 2018 - Lexical and Computational Semantics, SEM 2018, Proceedings of the 7th Conference; Dmitriy Dligach, Timothy Miller**

**Task**: Effectively represent patient data from clinical texts for phenotyping and various ML tasks
**Innovation**: Simpler alternative phenotyping method involving neural networks via billing codes; obtaining state-of-the-art performance on a standard comorbidity detection task
**Advantage**: Simpler model outperforms standard phenotyping models, on average
**Disadvantages**: Simpler model performed worse/tied on 3/16 diseases – quoting it likely due to "scarcity of positive training examples"
**Data Accessibility**: MIMIC III Database (CCU Patients, ICD9, CPT codes), Informatics for Integrating Biology to the Bedside (i2b2) Obesity Challenge (publicly available dataset)
**Code Accessibility**: https://github.com/dmitriydligach/starsem2018-patient-representations; cTAKES (open-source system for processing clinical texts for identifying UMLS CUIs)

**Paper 2: 42, "StageNet: Stage-Aware Neural Networks for Health Risk Prediction"; Proceedings of The Web Conference 2020; Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M. Glass, Jimeng Sun**

**Task**: Health Risk Prediction using EHR data
**Innovation**: "StageNet" model to capture disease progression stages without supervision, disease progression patterns & indicative features in each stage to help predict patient's health risks
**Advantage**: Outperforms baseline standard models against two real-world patient datasets in risk prediction & patient subtyping tasks (up to 12% AUPRC for risk prediction tasks); Able to adaptively emphasize indicative features to help predict patient's health risk at various disease progression stages; considers that disease progression speed depends on the underlying disease stage
**Disadvantages**: N/A
**Data Accessibility**: MIMIC III Database, ESRD Dataset
**Code Accessibility**: https://github.com/v1xerunt/StageNet

**Paper 3: 41, "Learning Tasks for Multitask Learning: Heterogenous Patient Populations in the ICU"; Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery; Harini Suresh, Jen J. Gong, John V. Guttag**

**Task**: Learn relevant patient subgroups & predict outcomes for these subgroups

**Innovation**: Data-driven multi-task framework for discovering patient subgroups via underlying physiological patient data w/ a sequence-to sequence autoencoder & mortality prediction risk for these
subgroups (from their 24/48-hour duration stay in ICU)
**Advantage**: Able to identify physiologically distinct cohorts of patients in a data-driven way (as opposed to expert-knowledge driven methods) which is beneficial for clinical risk model; can be used as a supplement to expert groups (first care unit labels) to help improve model performance, but performs well without expert knowledge labels (in cases of those labels not existing in a clinical population)
**Disadvantages**: 48-hour duration mortality prediction model didn't perform as well (due to scarcity of physiological data); cohort discovery to morbidity prediction doesn't do well with extreme sparsity in the data
**Data Accessibility**: MIMIC III Database
**Code Accessibility**: https://github.com/mit-caml/multitask-patients

**Target Paper**

1) **Which paper in the candidate you will replicate.**
   ***Paper 1:*** *Learning Patient Representations from Text*
2) **Why you choose the paper.**
   *With the LLM hype and constant move towards automation, it is important for big data experts to have a solid understanding of text embeddings and what is required to make text readable to machines. This is important in the healthcare industry, as phenotyping is a critical to the innovation and advancement of research and treatment, and we believe reimplementing this paper will improve our understanding of this big data practice along with deep learning concepts*.
3) **What are the specific hypotheses from the paper that you plan to verify in your reproduction study?**
   *Our goal is to verify that the deep learning methodology presented in the paper has stronger performance when phenotyping patients than baseline models, such as SVM's*.
4) **Briefly state how you are assured that you can obtain appropriate data and computational resources including software and hardware demanded in the paper.**
   *The data used in the paper is accessible to us, including the MIMIC-III database, so data gathering should not be a concern. We plan to use open-source libraries, such as Numpy, PyTorch and/or Tensorflow to construct our models. The biggest concern we have is accumulating computational power for training, as the authors used a Titan X GPU for theirs. We anticipate being able to use Google Colab for training at the very least, as powerful GPU's like V100 are made available there. Spinning up an instance on AWS, Azure or even renting a GPU from Vast.AI could also be options. Given the time constraints, we anticipate having to use Colab, which is also beneficial from a team collaboration standpoint.*