FD ID: FD1982
Requestor Name: Vivek Rudrapatna
Requestor Institution: University of California, San Francisco
Project Title: Generalizable AI for clinical language understanding using nationwide datasets

Background/Rationale: Clinical notes are the single richest source of information about patients, treatments, and outcomes. Embedded within them are many insights with immediate applications to precision diagnosis and personalized treatment. Yet they are also the least utilized source of clinical data, due to their unstructured nature and the high costs of manual information extraction. Large language models (LLMs) like chatGPT have shown tremendous promise in their ability to reason over text, and thus could be a potential solution to this major bottleneck. However, due to their nature of training over general text from the internet, these models have known limitations when applied to high-stakes domains like healthcare. Many have proposed the development of a dedicated LLM for healthcare, trained using clinical notes as captured in electronic health records (EHR) systems. However, careful analyses of the LLM training process have concluded that these models require massive quantities of data, on the order of what is typically available across *dozens* of health systems. Establishing such large pooled data repositories is nearly impossible for many reasons, including the notes' sensitive nature.

Our objective is to overcome these challenges and train the first nationwide clinical LLM by advancing a newly developed technology for distributed training called virtual pooling. This method allows for the training of complex machine learning models across a network of otherwise siloed datasets without requiring direct data sharing. The outcome of this project will be a new and state-of-the-art clinical AI, deployable within EHR systems to improve clinical decisions and enable new research discoveries.

Specific Aims: The specific aims of this project focus on (i) scaling the virtual pooling technology, (ii) training the first nationwide clinical LLM, and (iii) benchmarking clinical applications atop the LLM.

Aim 1: Develop the virtual pooling (VP) technology in preparation for AI training over siloed, nationwide datasets of clinical notes. A key aspect of VP is elimination of the need to extract and transfer data. This ability to train AI models without creation of a centralized repository reduces contractual delays. This aim will scale VP to work robustly with LLMs consisting of tens of billions of parameters.

Aim 2: Train state-of-the-art clinical LLM over a consortium of health systems across the US. We will deploy virtual pooling to train the LLaMa2-chat LLM with 65 billion parameters directly over clinical notes. Prior studies have suggested that we will need a training dataset with 1.6 trillion tokens (20x the size of data typically available at a single site).

Aim 3: Benchmark the trained LLM over a wide range of clinical tasks. In this aim, we will set up a suite of existing and new benchmarks to evaluate the performance of our AI. For example, we will create and benchmark performance over a de-novo dataset for extracting key study variables from clinical notes (e.g., cancer grade, treatment-related adverse events, social determinants of health, patient reported outcomes).

Intervention (if applicable): There is no intervention

Collaborators: UCSF will be collaborating with University of California, Santa Barbara (UCSB), and Dataunite, Inc. The co-investigators and PIs include: Madhumita Shushil, PhD, Postdoctoral researcher, UCSF; Divyakant Agrawal, PhD, Computer Science professor, UCSB; Xifeng Yan, PhD, Computer Science professor, UCSB; Trinabh Gupta, PhD, Dataunite and Computer Science assistant professor, UCSB. Dataunite will be responsible for developing and managing the virtual pooling software.