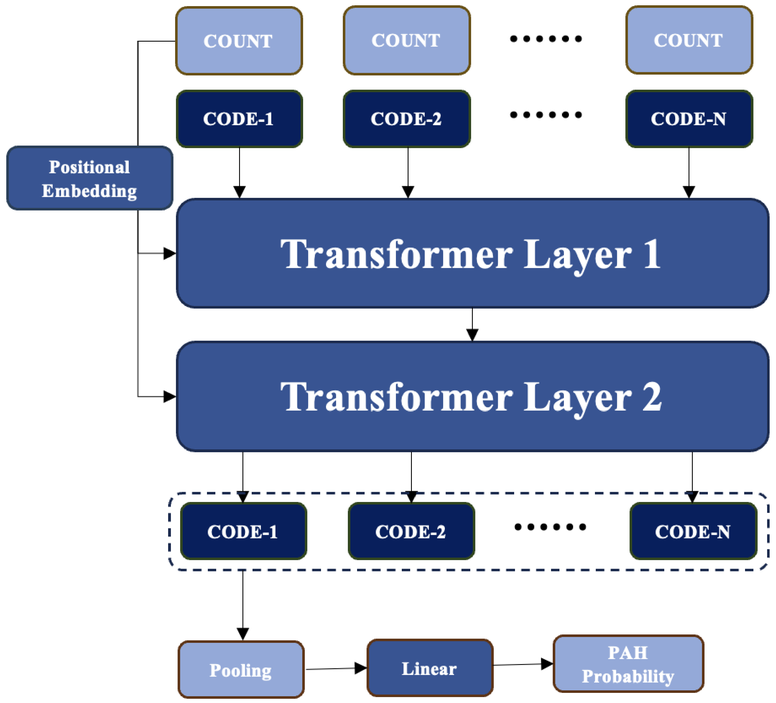
A Semi-Supervised Framework to Enhance Rare Disease Phenotyping from EHR

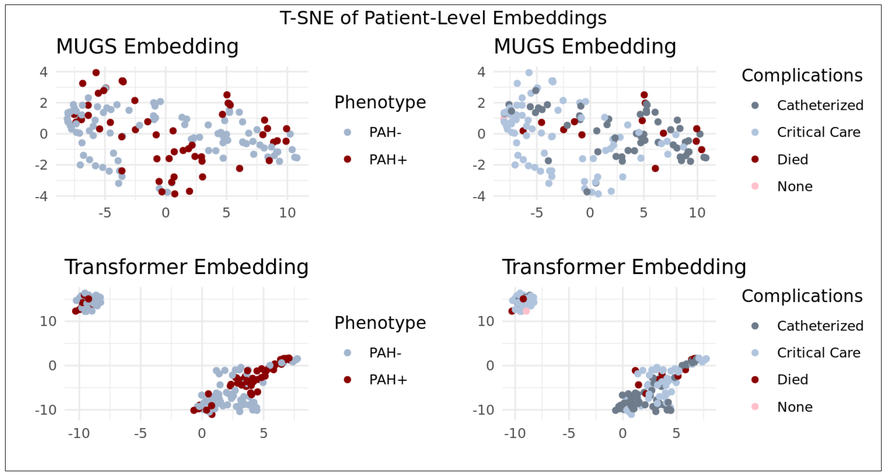
Kimberly F. Greco, Zongxin Yang, Mengyan Li, Han Tong, Alon Geva, Kenneth D. Mandl, Tianxi Cai

Background

- **Rare Diseases as a Global Challenge**: Rare diseases collectively affect an estimated 300 million people worldwide. Despite their collective impact, the individual low prevalence of these conditions creates significant challenges for timely diagnosis and treatment. Accurate phenotyping is necessary to drive advancements in research and clinical care.
- **Pulmonary Arterial Hypertension (PAH) as a Case Study**: PAH exemplifies the challenges faced in rare disease research. As a progressive disease with an estimated prevalence of 0.5 to 2 cases per million children, PAH is characterized by increased pulmonary vascular resistance, ultimately leading to heart failure and premature death. Early diagnosis is difficult due to nonspecific symptoms and reliance on specialized clinical expertise.
- **Challenges in Phenotyping Rare Diseases**: Identifying phenotypes for rare diseases like PAH is complicated by the resource-intensive nature of manually labeling cases and controls via expert review of patient records. This labor-intensive process limits sample sizes, compromises the scalability of phenotyping models, and hinders accurate disease characterization. In many cases, researchers only have access to disease registries, leaving positive-only datasets as the primary resource for training algorithms.



- **Semi-Supervised Framework for Phenotyping**: To address the challenges of phenotyping rare diseases, we propose a semi-supervised framework that utilizes **positive-only gold-standard labels** to train a robust phenotyping algorithm. These gold labels are augmented with **silver-standard labels** inferred from electronic health records (EHRs), creating a scalable and iterative approach. This framework includes:
  1. **Positive-Only Data**: Using gold-standard labels exclusively from disease registries to initialize the phenotyping process and train the algorithm.
  2. **Silver-Standard Labels**: Refining algorithm performance with labels inferred from EHR patterns, such as diagnostic code frequency and clinical context. These silver labels expand the dataset while maintaining sufficient accuracy.
  3. **Patient-Level and Temporal Embeddings**: Training embeddings that represent a patient's entire medical history or specific temporal windows. These embeddings capture dynamic, time-specific features, enabling the prediction and visualization of disease progression.
- By combining positive-only labels with iteratively refined silver labels, this framework enhances the scalability and accuracy of phenotyping models. The application of this approach to PAH demonstrates its potential to overcome the unique challenges of rare diseases, including limited labeled data and the need for insights into disease progression.



Methods

EHR Data Processing

- **Cohort**: Data collected from 6536 patients at Boston Children's Hospital (BCH) with at least one PheCode indicating potential PAH (PheCode 415.2).
  1. **Gold-Standard Patients**: 262 patients with true PAH labels, verified through a PAH registry for training or detailed physician chart reviews for validation.
  2. **Silver-Standard Patients**: 6274 patients without gold-standard labels are labeled based on having ≥2 PAH-related codes in their EHR.
- **Temporal Representation**: EHR data summarized monthly, with months containing PAH codes flagged for analysis.
- **Time-Truncation Augmentation**: To address the limited size of the gold-standard training cohort, time-truncation augmentation is applied to simulate variations in patient timelines, allowing the model to generalize across incomplete or irregular data and improve robustness.

Algorithm

The proposed phenotyping algorithm leverages a transformer-based architecture to generate patient-level and temporal embeddings for rare disease phenotyping. Key components include:

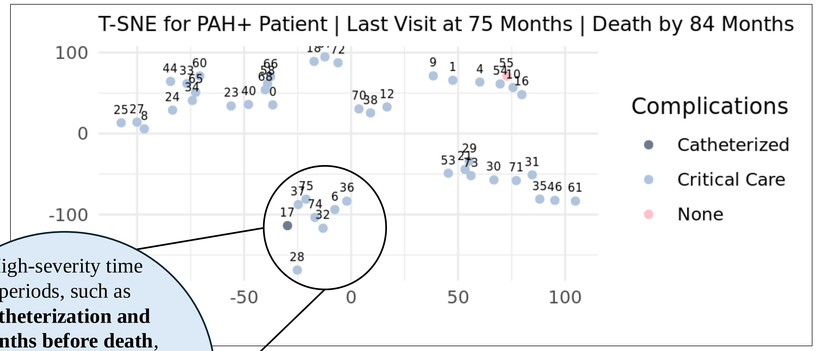
1. **Input Representation**
   - Each patient's EHR is encoded as a sequence of **medical codes** (diagnoses, medications, procedures) and their associated **counts**.
   - **MUGS Embeddings**: The model is initialized with pre-trained Multisource Graph Synthesis (MUGS) embeddings for each medical code, iteratively refining these representations during training to capture temporal and contextual relationships.
2. **Transformer Layers**
   - The encoded sequence is processed through **two transformer layers**, learning contextual relationships between codes and capturing temporal dependencies.
   - The training process minimizes a weighted combination of:
     - **Binary Cross-Entropy Loss**: To optimize PAH prediction.
     - **Contrastive Loss**: To improve representation quality by aligning current model embeddings with EMA embeddings (Exponential Moving Average of model parameters).
3. **Estimation**
   - **Pooling**: Code-level embeddings are aggregated using mean pooling, consolidating the sequence into a single **patient-level representation**.
   - **Prediction**: The pooled embedding passes through a linear layer, and the output is transformed using a sigmoid activation function, generating a probability score for PAH likelihood.

Results

1. **Model Performance**: Transformer-based embeddings significantly improved disease classification performance compared to baseline PheCode counts, achieving an AUC of 76.1 vs. 70.5 across a range of label cutoffs.
2. **Clustering**: t-SNE visualizations revealed distinct clusters of PAH+ patients, highlighting the model's ability to capture patient-level phenotypes. These clusters suggest improved phenotyping, where embeddings encode meaningful distinctions between disease states.
3. **Temporal Insights**: Time-specific embeddings demonstrated patterns of disease progression within individual patients. These embeddings provide insights into how clinical events and disease severity evolve over time.

Next Steps

- Validate the model with alternative sets of silver-standard labels to evaluate and track improvements in classification performance.
- Incorporate time-to-death data and embeddings to better capture periods of severe disease, further refining temporal representations.



High-severity time periods, such as **catheterization and months before death**, cluster closely within individual patients, indicating that the **embeddings capture temporal patterns** of severe disease.

Gradients in the plots below reflect **shifts in disease trajectory over time**, with periods of fewer or no complications positioned farther from clusters of critical events, highlighting **gradual changes in disease severity throughout the follow-up period**.