

STATISTICAL SIGNIFICANCE

For patients with rare diseases, the road to an accurate diagnosis is often long, winding, and filled with uncertainty. This journey – known as the “diagnostic odyssey” – captures the many specialist referrals, inconclusive tests, and misdiagnoses that rare disease patients endure before arriving at a correct diagnosis. But what if we could pave a faster, more reliable path? We present a data-driven deep learning framework, inspired by the same transformer architecture underlying popular large language models, to accelerate rare disease diagnosis and guide patients toward the care they need – sooner.

RARE DISEASES: A GLOBAL PROBLEM HIDDEN IN PLAIN SIGHT

Rare diseases are individually uncommon, but collectively they affect an estimated 300-400 million people worldwide. Strikingly, 7 in 10 rare diseases begin in childhood, and 1 in 3 children affected won't live past the age of five. The stakes are high – but diagnosis remains slow, elusive, and expensive.

With over 7,000 unique rare diseases, most clinicians will never encounter the same condition twice. As a result, patients with rare diseases often embark on a “diagnostic odyssey” lasting 4 to 7 years on average – marked by countless visits, specialist referrals, inconclusive tests, and frequent misdiagnoses. These delays result in lost time, irreversible complications, and missed opportunities for early treatment or clinical trial enrollment.

EHRs TELL A STORY: IDENTIFYING DISEASE PATTERNS FROM DATA

Modern healthcare systems generate rich, real-world data through **electronic health records (EHRs)**. These digital records offer critical diagnostic clues – capturing diagnoses, medications, procedures, lab results, and free-text clinical notes that document a patient's condition in real time.

EHRs are messy – full of incomplete records, inconsistent codes, and clinical shorthand. They're also powerful. Hidden in these data are subtle but meaningful patterns that can signal early signs of rare disease – but recognizing them at scale is beyond human ability. That's where statistics shines.

We built a **computational phenotyping framework** that uses both structured codes and unstructured notes within EHRs to automatically flag patients at risk for rare diseases. It doesn't just look for a diagnostic code – it learns from how conditions evolve, how symptoms cluster, and how physicians document uncertainty or concern.

TRANSFORMERS: FROM LANGUAGE TO LIFE-SAVING PREDICTION

To capture the full complexity of patient histories, we used a powerful deep learning architecture called the **transformer** – originally developed for natural language processing and now powering popular tools like ChatGPT.

Why transformers? They're uniquely suited to model sequential, longitudinal data, like the kind found in EHRs.

We adapted the transformer to learn patient-level representations from a timeline of medical concepts, weighting not just what happened, but how often and in what context.

Here's the twist: unlike traditional supervised machine learning, we don't rely on large sets of perfectly labeled data (which are rare for rare diseases!) to train our model.

Instead,

we use a technique called **weak supervision**, starting with a small set of expert-confirmed cases and iteratively improving noisy labels drawn from EHRs for unlabeled patients. Over time, the model yields more confident,




accurate, and clinically useful predictions.

FROM MODEL TO MEDICINE: A PATH FORWARD FOR RARE DISEASE PATIENTS

We validated our model using data from **Boston Children's Hospital** on two rare pulmonary diseases: pulmonary hypertension and severe asthma. The model not only outperformed traditional approaches in detecting these diseases, but also identified clinically

meaningful subgroups – patients who progressed faster, responded differently to treatment, or presented with distinct risk profiles.

By transforming how we use real-world EHR data, our framework supports:

-  **Earlier recognition of high-risk patients**
-  **Smarter and broader clinical trial enrollment**
-  **Personalized management strategies**
-  **Better outcomes for patients living with rare diseases**

This is more than an algorithm – it's a step toward closing the diagnosis gap for millions of families worldwide. And that's the statistical significance: making data work for those who need answers most.