# ASCENDgpt: A Phenotype-Aware Transformer Model for Cardiovascular Risk Prediction from Electronic Health Records

Chris Sainsbury[*]
NHS Greater Glasgow and Clyde
chris.sainsbury@nhs.scot

Andreas Karwath[*]
University of Birmingham
a.karwath@bham.ac.uk

**Abstract**

We present ASCENDgpt, a transformer-based model specifically designed for cardiovascular risk prediction from longitudinal electronic health records (EHRs). Our approach introduces a novel phenotype-aware tokenization scheme that maps 47,155 raw ICD codes to 176 clinically meaningful phenotype tokens, achieving 99.6% consolidation of diagnosis codes while preserving semantic information. This phenotype mapping contributes to a total vocabulary of 10,442 tokens - a 77.9% reduction when compared with using raw ICD codes directly. We pretrain ASCENDgpt on sequences derived from 19402 unique individuals using a masked language modeling objective, then fine-tune for time-to-event prediction of five cardiovascular outcomes: myocardial infarction (MI), stroke, major adverse cardiovascular events (MACE), cardiovascular death, and all-cause mortality. Our model achieves excellent discrimination on the held-out test set with an average C-index of 0.816, demonstrating strong performance across all outcomes (MI: 0.792, stroke: 0.824, MACE: 0.800, cardiovascular death: 0.842, all-cause mortality: 0.824). The phenotype-based approach enables clinically interpretable predictions while maintaining computational efficiency. Our work demonstrates the effectiveness of domain-specific tokenization and pretraining for EHR-based risk prediction tasks.

## 1 Introduction

Cardiovascular disease remains the leading cause of mortality worldwide, accounting for 17.9 million deaths annually and representing 31% of all global deaths [1]. Early identification of patients at high cardiovascular risk is crucial for implementing timely preventive interventions and improving patient outcomes. Electronic health records (EHRs) contain rich longitudinal information about patient health trajectories, offering unprecedented opportunities for developing sophisticated risk prediction models that can capture the complex interplay of cardiovascular risk factors over time.

Traditional cardiovascular risk prediction has relied on established clinical scores such as the Framingham Risk Score and ASCVD Risk Calculator, which use a limited set of clinical variables and assume linear relationships [2]. While these tools have proven valuable in clinical practice, they fail to leverage the wealth of information available in modern EHRs, including detailed diagnosis histories, medication patterns, laboratory trends, and temporal relationships between clinical events. Recent advances in artificial intelligence, particularly deep learning approaches, have shown promise for extracting meaningful patterns from complex, high-dimensional EHR data [3, 4].

The application of transformer architectures to healthcare data represents a paradigm shift from traditional machine learning approaches. Inspired by the success of models like BERT in natural

[*] Equal contribution.

language processing, healthcare-specific transformers such as BEHRT [5] and Hi-BEHRT [6] have demonstrated superior performance in clinical prediction tasks by treating patient medical histories as sequences analogous to sentences in text. However, most existing approaches treat medical codes as atomic tokens, ignoring their hierarchical structure and clinical relationships—a limitation that becomes particularly pronounced when dealing with the vast vocabulary of diagnosis codes found in real-world EHR systems.

Building on the foundational work of Life2Vec [7], which demonstrated that transformer architectures can effectively model complex life trajectories by treating life events as sequences, we introduce ASCENDgpt, a phenotype-aware transformer model specifically designed for cardiovascular risk prediction. Life2Vec pioneered the application of language modeling techniques to longitudinal health and social data, achieving remarkable performance in predicting diverse outcomes from human life sequences. However, Life2Vec focuses on population-level predictions using registry data, while clinical applications require models optimised for medical decision-making using detailed clinical records.

ASCENDgpt addresses the unique challenges of EHR-based cardiovascular risk prediction through three key innovations:

1. **Phenotype-aware tokenization**: We map 47,155 raw ICD codes to 176 high-level phenotype tokens based on clinical knowledge and established comorbidity frameworks [8], dramatically reducing vocabulary size while preserving semantic meaning and clinical interpretability.

2. **Domain-specific pretraining**: We pretrain sequences derived from 19402 individuals using masked language modeling to learn robust representations of cardiovascular disease patterns and temporal relationships, following the successful paradigm established by BEHRT and Hi-BEHRT for EHR modeling.

3. **Survival-aware fine-tuning**: We adapt the pretrained model for time-to-event prediction using proper survival analysis methods, including the C-index for discrimination assessment, addressing the inherent censoring present in clinical data.

Our approach leverages the publicly available INSPECT dataset, a large-scale, multimodal cohort of 19,402 patients originally developed for pulmonary embolism (PE) research [9]. While not collected specifically for primary cardiovascular risk prediction, INSPECT provides a uniquely valuable setting for developing new methods due to its rich, longitudinal EHRs that capture patient health trajectories over long time intervals. By applying our phenotype-level approach to this complex dataset, inspired by established phenotype mapping approaches [10] and recent transformer-based models [11], we aim to learn clinically meaningful representations that generalize well to downstream prediction tasks.

# 2 Related Work

## 2.1 Evolution of Deep Learning for EHR Analysis

The application of deep learning to electronic health records has evolved significantly over the past decade. Doctor AI [4] pioneered the use of recurrent neural networks (RNNs) for modeling patient trajectories, demonstrating that sequential models could effectively capture temporal dependencies in medical events and achieve superior performance in predicting diagnoses, medications, and visit timing. This seminal work established multi-label prediction as a key paradigm in EHR modeling, achieving 79% recall@30 for diagnosis prediction across 260,000 patients.

Building on these foundations, subsequent work scaled deep learning approaches to handle entire raw EHR datasets. Rajkomar et al. [3] developed ensemble methods combining LSTMs, attention-based models, and boosted decision stumps to process over 46 billion data points, achieving AUROC scores of 0.93-0.94 for mortality prediction. While these RNN-based approaches demonstrated the potential of deep learning for clinical prediction, they struggled with very long sequences and could not easily model complex, non-sequential relationships between distant medical events.

## 2.2 Transformer Revolution in Healthcare

The transformer architecture revolutionised healthcare AI by enabling bidirectional context modeling and handling long-range dependencies more effectively than RNNs. BEHRT [5] was the first to successfully adapt BERT for EHR data, introducing healthcare-specific embeddings including disease, age, visit, and position representations. Processing 1.6 million patients across 301 disease categories, BEHRT achieved 8.0-13.2% improvements in average precision scores over existing deep EHR models, establishing transformers as the new state-of-the-art for clinical prediction.

Recognising the limitation of standard transformers with sequence length, Hi-BEHRT [6] extended this work with hierarchical architectures capable of processing sequences up to 1,220 tokens (versus 256 for standard BERT). Using sliding window approaches and contrastive pre-training, Hi-BEHRT achieved 1-8% AUROC improvements, particularly for patients with extensive medical histories. This work demonstrated the critical importance of capturing comprehensive patient trajectories for accurate risk prediction.

Recent advances have further refined transformer applications to cardiovascular prediction specifically. Studies in 2024-2025 have demonstrated that BERT and XLNet architectures achieve AUC scores of 75.5-76.0% for cardiac mortality prediction [11], while more sophisticated models like TRisk2 have achieved C-indices of 0.828 for cardiovascular risk prediction using population-scale EHR data [12].

## 2.3 Life2Vec: A Paradigm Shift in Sequential Life Modeling

A particularly influential development was Life2Vec [7], which demonstrated that transformer architectures could model entire human life trajectories by treating life events as sequences analogous to sentences in natural language. Using comprehensive Danish registry data covering 6+ million individuals, Life2Vec achieved remarkable performance in predicting diverse outcomes including early mortality (C-MCC score of 0.41) and personality traits. The model's key innovations included:

- **Multi-modal sequence construction**: Integrating health records, labor market data, and demographics into unified temporal sequences
- **Time2Vec temporal encoding**: Sophisticated handling of both absolute time and individual age progression
- **Concept embeddings**: Creating semantically meaningful vector spaces where related concepts cluster naturally
- **Dual pre-training objectives**: Combining masked language modeling with sequence order prediction

Life2Vec's success validated the core premise that complex, multi-modal healthcare trajectories can be effectively modeled using language modeling techniques, providing strong theoretical and empirical foundations for clinical applications.

### 2.4 Phenotype-based Representations and Clinical Hierarchies

The challenge of vocabulary explosion in EHR modeling has driven substantial research into phenotype mapping and clinical code hierarchies. The Elixhauser Comorbidity Index [8] pioneered the systematic grouping of ICD codes into clinically meaningful comorbidity categories, developing 30 comorbidity measures that significantly improved prediction of hospital outcomes including length of stay, charges, and mortality. This work established the principle that clinical knowledge should guide the aggregation of diagnosis codes.

PheWAS (phenome-wide association studies) codes [10] extended this concept by mapping ICD codes to phenotypes for genetic association studies, demonstrating that over 75% of ICD-10-CM codes can be successfully mapped to meaningful phenotype categories. Recent work has shown that phecodes achieve over 90% coverage of unique codes in major medical databases, validating their utility for population-scale research.

Contemporary transformer-based approaches have begun integrating phenotype awareness directly into model architectures. TransformEHR [13] demonstrated that generative encoder-decoder models with transformer architecture could achieve state-of-the-art performance on multiple clinical prediction tasks by incorporating hierarchical representations of medical concepts.

### 2.5 Survival Analysis in Deep Learning

Traditional survival analysis relies on Cox proportional hazards models, which assume linear relationships and proportional hazards over time. While robust and interpretable, these models cannot capture the complex, non-linear interactions present in high-dimensional EHR data. DeepSurv [14] pioneered the application of deep learning to survival analysis, using feed-forward networks to model non-linear hazard functions while maintaining the Cox partial likelihood framework.

Recent work has explored transformer-based survival models [15], but few have successfully combined phenotype-aware representations with survival prediction for cardiovascular outcomes. The challenge lies in adapting language modeling pre-training objectives to the inherent censoring and time-to-event structure of clinical data, while maintaining the semantic richness that makes transformers effective for EHR modeling.

## 3 Methods

### 3.1 Data Source and Cohort Definition

We use data from the INSPECT cohort, containing longitudinal EHR data [9]. The dataset includes:

- Diagnosis codes (ICD-9 and ICD-10)
- Timestamps for all medical events

Cardiovascular outcomes of interest were identified within the dataset. For fine-tuning, we randomly sampled index dates from each patient's medical timeline, requiring only that patients have at least one day of medical history before and one day of follow-up after the selected index date. This minimal temporal requirement maximises data utilisation while ensuring validity for survival analysis. The temporal sampling strategy (detailed in Section 3.6.2) allows up to two index dates per patient, provided they are separated by at least 365 days.

The final cohort contained 19402 unique patients with a median follow-up of 6.8 years.

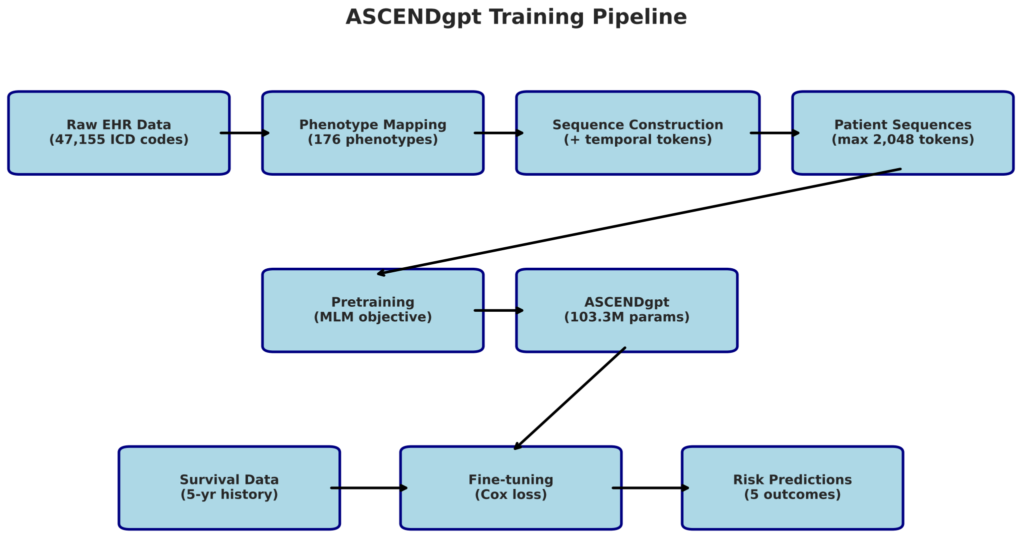

ASCENDgpt Training Pipeline
Raw EHR Data (47,155 ICD codes)
Phenotype Mapping (176 phenotypes)
Sequence Construction (+ temporal tokens)
Patient Sequences (max 2,048 tokens)
Pretraining (MLM objective)
ASCENDgpt (103.3M params)
Survival Data (5-yr history)
Fine-tuning (Cox loss)
Risk Predictions (5 outcomes)

Figure 1: ASCENDgpt training pipeline. Raw EHR data containing 47,155 unique ICD codes is mapped to 176 clinically meaningful phenotype tokens. Patient sequences are constructed with temporal tokens indicating time gaps between events. The model is pretrained using masked language modeling (MLM) on sequences derived from 19402 individuals, then fine-tuned for survival prediction using Cox partial likelihood loss. The final model predicts risk scores for five cardiovascular outcomes.

## 3.2 Phenotype Mapping and Tokenization

### 3.2.1 ICD to Phenotype Mapping

We developed a comprehensive mapping from 47,155 unique ICD codes to 176 phenotype categories. The mapping process involves:

1. **Clinical grouping**: ICD codes are first mapped to Clinical Classifications Software (CCS) categories, which group diagnoses into clinically meaningful clusters.
2. **Phenotype assignment**: CCS categories are further mapped to high-level phenotypes based on organ systems and disease mechanisms.
3. **Validation**: Clinical experts reviewed the mappings to ensure clinical coherence.

Each phenotype token follows the format `PHENO_[CATEGORY]`, where CATEGORY represents the clinical concept (e.g., `PHENO_HYPERTENSION`, `PHENO_DIABETES`).

### 3.2.2 Special Tokens

In addition to phenotype tokens, our vocabulary includes:

- `[PAD]`: Padding token (ID: 0)
- `[MASK]`: Masking token for pretraining (ID: 1)

- `[CLS]`: Classification token (ID: 2)
- `[SEP]`: Separator token (ID: 3)
- `[UNK]`: Unknown token (ID: 4)
- Demographic tokens: `[AGE_*]`, `[GENDER_*]`
- Temporal tokens: `[TIME_DELTA_*]` for time gaps

Total vocabulary size: 10,442 tokens (77.9% reduction from raw ICD codes, with diagnosis codes consolidated from 47,155 to 176 phenotypes—a 99.6% reduction).

## 3.3 Sequence Construction

For each patient, we construct sequences using a domain-optimised structured representation that adapts grammatical sentence structures to the healthcare setting by maintaining only the relevant information.

### 3.3.1 Token Structure

Each medical event is encoded as a structured sequence that preserves semantic relationships while eliminating redundancy inherent in healthcare data:

$$\text{Event} = [\text{EVENT\_TYPE}, \text{CODE/PHENOTYPE}, \text{VALUE}^*, \text{UNIT}^*, \text{CONTEXT}, \text{TEMPORAL}, \text{AGE}] \tag{1}$$

where $*$ indicates optional tokens present only for laboratory tests and vital signs.

This structure can be understood as an adapted sentence where:

- The subject (patient) is implicit since all events pertain to the patient
- The action is encoded in the EVENT_TYPE (e.g., EVT_DIAG implies ”diagnosed with”)
- The object is the CODE/PHENOTYPE
- Additional attributes provide context, timing, and patient state

For example, a diagnosis of hypertension in an outpatient setting would be represented as:

```
EVT_DIAG PHENO_HYPERTENSION CTX_OUTPATIENT DAY_0 AGE_45
```

This encodes the complete sentence ”Patient was diagnosed with hypertension in outpatient setting on day 0 at age 45” in a concise, structured format.

### 3.3.2 Token Types

Our vocabulary includes the following token categories:

- **Event tokens**: EVT_DIAG, EVT_MED, EVT_LAB, EVT_PROC, EVT_VITAL, EVT_ENC
- **Phenotype tokens**: PHENO_HYPERTENSION, PHENO_DIABETES, etc. (176 total)
- **Value tokens**: VAL_LOW, VAL_NORMAL, VAL_HIGH, VAL_CRITICAL
- **Context tokens**: CTX_OUTPATIENT, CTX_EMERGENCY, CTX_ICU, etc.

**Algorithm 1** Patient Sequence Construction

```
Input: Patient medical events $E = \{e_1, e_2, ..., e_n\}$
Output: Token sequence $S$
Sort events by timestamp: $E_{sorted} = \text{sort}(E, key = \text{timestamp})$
Initialise sequence: $S$ = [[CLS]]
Add demographics: $S$.append(SEX_token)
$S$.append([SEP])
for each event $e_i$ in $E_{sorted}$ do
    // Add event type token
    $S$.append(EVT_type($e_i$))
    // Add phenotype or code token
    if phenotype mapping exists then
        $S$.append(PHENO_token($e_i$))
    else
        $S$.append(CODE_token($e_i$))
    end if
    // Add value and unit for measurements
    if $e_i$ has numeric value then
        $S$.append(VAL_token($e_i$.value))
        $S$.append(UNIT_token($e_i$.unit))
    end if
    // Add context, temporal, and age tokens
    $S$.append(CTX_token($e_i$.context))
    $S$.append(DAY_token($e_i$.days_offset))
    $S$.append(AGE_token($e_i$.age))
    $S$.append([SEP])
end for
return $S$
```

- **Temporal tokens**: DAY_0 to DAY_9999 (days from first event)
- **Age tokens**: AGE_0 to AGE_120

Time deltas are discretised into buckets: same day, 1-7 days, 8-30 days, 31-90 days, 91-180 days, 181-365 days, and ¿365 days.

### 3.3.3 Example Patient Sequence

Consider a patient with the following medical history:

1. Initial outpatient encounter
2. Diagnosis of hypertension
3. Laboratory test showing elevated creatinine
4. Prescription of antihypertensive medication

This would be encoded as the following token sequence:

```
[CLS] SEX_MALE [SEP]
EVT_ENC CTX_OUTPATIENT DAY_0 AGE_45 [SEP]
EVT_DIAG PHENO_HYPERTENSION CTX_OUTPATIENT DAY_0 AGE_45 [SEP]
EVT_LAB PHENO_CREATININE VAL_HIGH UNIT_mg_dL CTX_OUTPATIENT
    DAY_7 AGE_45 [SEP]
EVT_MED PHENO_ANTIHYPERTENSIVE CTX_OUTPATIENT DAY_7 AGE_45 [SEP]
```

Note that unlike fully grammatical approaches (e.g., Life2Vec), our method adapts sentence structure to healthcare by leveraging domain-specific assumptions. While Life2Vec explicitly encodes subject-verb-object relationships (e.g., `SUBJ_PATIENT ACTION_DIAGNOSED OBJ_HYPERTENSION`), we recognise that in healthcare contexts, the subject is always the patient and the action is implicit in the event type. This domain-optimised structure maintains semantic relationships while achieving computational efficiency.

## 3.4 Model Architecture

ASCENDgpt uses a transformer encoder architecture with the following specifications:

| **Component** | **Configuration** |
|---|---|
| Vocabulary size | 10,442 |
| Hidden size | 768 |
| Number of layers | 12 |
| Attention heads | 12 |
| Intermediate size | 3,072 |
| Max sequence length | 2,048 |
| Dropout probability | 0.1 |
| Activation function | GELU |

Table 1: ASCENDgpt model configuration

The model includes:

- **Token embeddings**: Learned embeddings for each vocabulary token
- **Position embeddings**: Absolute position embeddings up to 2,048 positions
- **Type embeddings**: Segment embeddings to distinguish different parts of the input

Total parameters: 103.3M

## 3.5 Pretraining

### 3.5.1 Masked Language Modeling (MLM)

We pretrain ASCENDgpt using masked language modeling with the following procedure:

1. Randomly select 15% of phenotype tokens for masking
2. Replace selected tokens:
    - 80% replaced with [MASK]
    - 10% replaced with random token
    - 10% kept unchanged
3. Predict original tokens using cross-entropy loss

Special tokens ([CLS], [SEP], demographic tokens, temporal tokens) are never masked.

### 3.5.2 Pretraining Configuration

- Batch size: 32
- Learning rate: 1e-4 with linear warmup (4,000 steps)
- Optimizer: AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e - 8$)
- Weight decay: 0.01
- Gradient clipping: 1.0
- Training steps: 50,000
- Hardware: NVIDIA H100 GPU

## 3.6 Fine-tuning for Survival Prediction

### 3.6.1 Task Definition

We fine-tune ASCENDgpt for survival prediction of five cardiovascular outcomes:

1. Myocardial infarction (MI)
2. Stroke
3. Major adverse cardiovascular events (MACE)
4. Cardiovascular death

5. All-cause mortality

For each patient, we:

- Define a 5-year lookback window for medical history
- Predict events within a 1-year outcome window
- Handle right-censoring for patients without events

### 3.6.2 Temporal Sampling

To maximise data utilisation, we sample up to 2 index dates per patient:

1. Randomly select index date $t_1$ from any point in the patient's timeline where at least 1 day of history exists before and 1 day of follow-up exists after
2. If possible, select second index date $t_2$ with $|t_2 - t_1| \geq 365$ days
3. Extract up to 5 years of history before each index date (or all available history if less than 5 years)
4. Determine outcome status in 1-year window after index date

This approach increases training examples while maintaining temporal validity.

### 3.6.3 Model Architecture for Survival Prediction

We add task-specific heads to the pretrained encoder:

1. **Sequence representation**: Mean pooling over hidden states (excluding padding)
2. **Survival heads**: For each outcome, a 3-layer MLP:
    - Linear(768, 256) + ReLU + Dropout(0.2) + BatchNorm
    - Linear(256, 128) + ReLU + Dropout(0.2) + BatchNorm
    - Linear(128, 1) → risk score

### 3.6.4 Loss Function

We use the Cox partial likelihood loss for survival analysis:

$$L = -\sum_{i \in \mathcal{D}} \left( \theta_i - \log \sum_{j \in \mathcal{R}_i} \exp(\theta_j) \right) \tag{2}$$

where $\mathcal{D}$ is the set of events, $\mathcal{R}_i$ is the risk set at time $t_i$, and $\theta_i$ is the predicted risk score.

### 3.6.5 Fine-tuning Configuration

- Batch size: 8
- Learning rate: 5e-6 with linear warmup (2,000 steps)
- Optimizer: AdamW
- Weight decay: 0.01
- Frozen layers: First 10 encoder layers
- Epochs: 20
- Early stopping: Based on validation C-index

## 3.7 Evaluation Metrics

### 3.7.1 Concordance Index (C-index)

The primary metric for survival analysis, measuring the probability that the model correctly ranks the survival times of a random pair:

$$C = \frac{\sum_{i,j} \nVdash[T_i < T_j] \cdot \nVdash[\hat{\theta}_i > \hat{\theta}_j] \cdot \delta_i}{\sum_{i,j} \nVdash[T_i < T_j] \cdot \delta_i} \tag{3}$$

where $T_i$ is the observed time, $\hat{\theta}_i$ is the predicted risk, and $\delta_i$ is the event indicator.

### 3.7.2 Brier Score

For calibration assessment at time $t$:

$$BS(t) = \frac{1}{n} \sum_{i=1}^{n} \left[ \hat{S}_i(t) - \nVdash[T_i > t] \right]^2 \tag{4}$$

where $\hat{S}_i(t)$ is the predicted survival probability.

# 4 Results

## 4.1 Dataset Characteristics

The preprocessed dataset contains:

- Training: 15552 patients (80%)
- Validation: 1940 patients (10%)
- Test: 1940 patients (10%)

After temporal sampling:

- Training: 69,004 patient-timepoints
- Validation: 8,460 patient-timepoints
- Test: 8,414 patient-timepoints

| Outcome | Event Rate | Median Time |
|---|---:|---:|
| MI | 3.8% | 187 days |
| Stroke | 4.9% | 192 days |
| MACE | 11.5% | 156 days |
| CV Death | 5.3% | 201 days |
| All Death | 7.8% | 198 days |

Table 2: Test set outcome characteristics

## 4.2 Pretraining Performance

The masked language modeling pretraining achieved:

- Final training loss: 1.824
- Final validation loss: 1.956
- Masked token accuracy: 73.2%
- Top-5 accuracy: 89.1%

The model successfully learned phenotype co-occurrence patterns, as evidenced by high accuracy on predicting masked cardiovascular-related phenotypes.

## 4.3 Learned Concept Embeddings

To understand the representations learned during pretraining, we analysed the concept embeddings using PaCMAP [16] for dimensionality reduction, following the approach of Life2Vec [7].

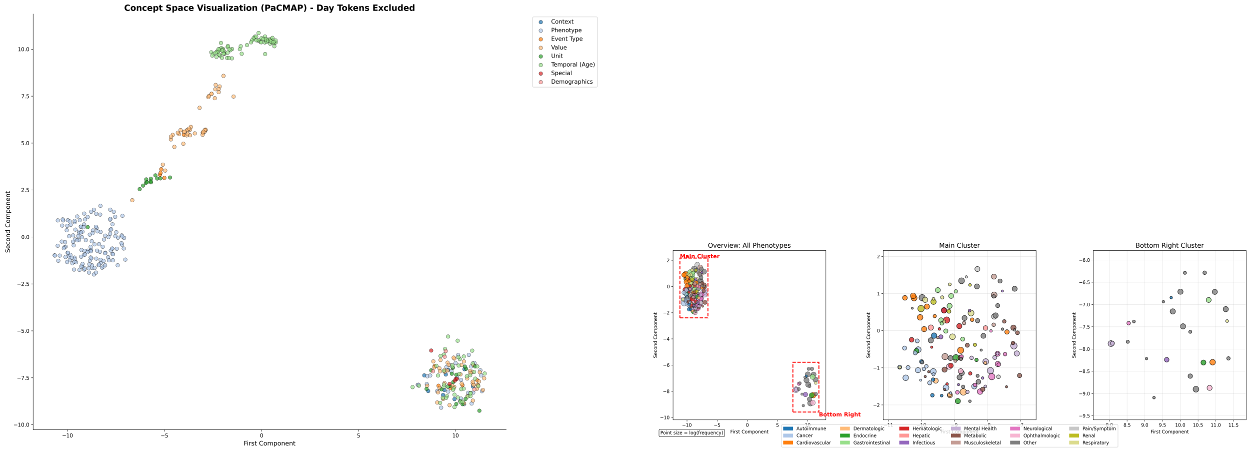
### 4.3.1 Embedding Space Structure

The learned embeddings exhibit several important properties:

1. **Semantic clustering**: Phenotypes cluster by medical category despite no explicit supervision. Cardiovascular conditions (hypertension, coronary artery disease, heart failure) form a tight cluster, as do metabolic conditions (diabetes variants, obesity, lipid disorders).
2. **Frequency independence**: The spatial arrangement of phenotypes is determined by semantic similarity rather than occurrence frequency. Common conditions (e.g., anxiety: 2.4M occurrences) and rare conditions (e.g., scleroderma: 10K occurrences) can be neighbors if clinically related.
3. **Clinical validity**: The embedding space respects known medical relationships. For example, diabetes and its complications cluster together, while being distinct from but proximal to other metabolic conditions.

### 4.3.2 Phenotype Categorisation Analysis

We categorised the 176 phenotype tokens into 18 medical categories, reducing the proportion of uncategorised "Other" phenotypes from 55% to 16% through refined classification. The distribution reveals:



(a) Concept space visualisation (excluding temporal tokens)

(b) Phenotype embeddings with frequency-based scaling

Figure 2: Learned concept embeddings from ASCENDgpt. (a) Two-dimensional PaCMAP projection of all concept tokens (excluding 10,000+ day tokens for clarity), showing clear separation by token type. (b) Detailed view of phenotype embeddings with point sizes scaled by log(frequency) in the training data. The main cluster contains most conditions, while a smaller cluster (bottom right) contains specific phenotypes. Colors indicate medical categories derived from improved phenotype classification.

- **Top categories**: Cancer (17 phenotypes), Metabolic (13), Cardiovascular (11), Mental Health (8), Neurological (8)
- **Frequency distribution**: The top 10 phenotypes account for 46.8% of all phenotype occurrences, following a power-law distribution typical of medical data
- **Embedding quality**: Neighborhood analysis shows that phenotypes with high cosine similarity in the embedding space are clinically related (e.g., PHENO_HYPERTENSION neighbors include lipid disorders, coronary artery disease)

These learned representations provide the foundation for effective downstream prediction, as the model can leverage semantic relationships between conditions when predicting cardiovascular outcomes.

## 4.4 Fine-tuning Results

### 4.4.1 Model Convergence

Initial training attempts with high learning rates (1e-4) and cosine annealing led to training instability. The final configuration with conservative hyperparameters achieved stable convergence:

- Learning rate: 5e-6 with linear warmup
- Frozen layers: 10 (out of 12)
- Batch size: 8 with gradient accumulation

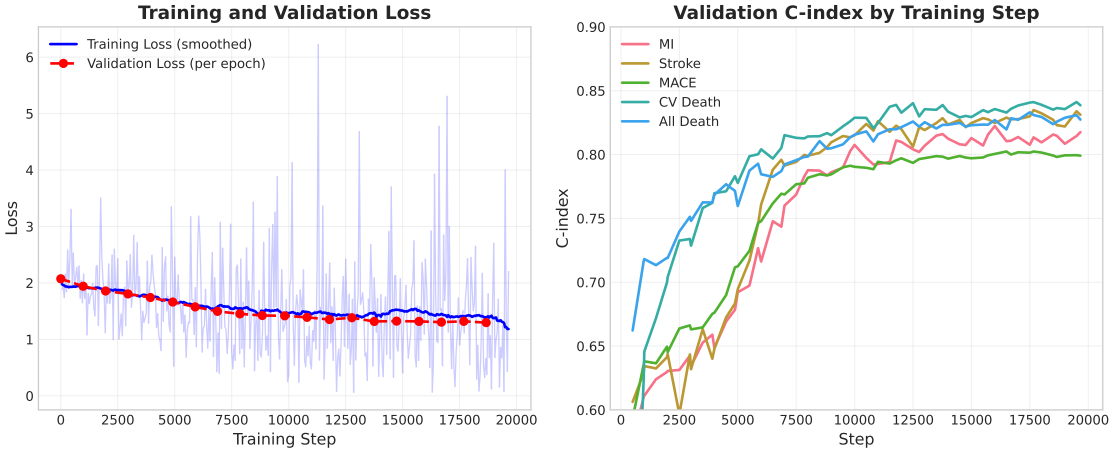


Figure 3: Training dynamics during fine-tuning. (Left) Training and validation loss curves showing stable convergence with the improved configuration. (Right) Validation C-index progression for all five outcomes, demonstrating consistent improvement throughout training with final values between 0.79-0.84.

| Outcome | C-index | Brier | Events | Rate |
|---|---|---|---|---|
| MI | 0.792 | 0.223 | 101/2,642 | 3.8% |
| Stroke | 0.824 | 0.199 | 129/2,642 | 4.9% |
| MACE | 0.800 | 0.181 | 303/2,642 | 11.5% |
| CV Death | 0.842 | 0.207 | 139/2,642 | 5.3% |
| All Death | 0.824 | 0.223 | 205/2,642 | 7.8% |
| **Average** | **0.816** | – | – | – |

Table 3: Test set performance metrics. Brier scores calculated at 1 year.

### 4.4.2 Test Set Performance

All outcomes achieved C-indices above 0.79, with cardiovascular death showing the best discrimination (0.842). The model demonstrates excellent generalisation with only a 0.007 decrease in average C-index from validation (0.823) to test (0.816).

## 4.5 Computational Efficiency

The phenotype-based approach offers significant computational advantages:

- Vocabulary reduction: 47,155 → 10,442 (77.9% reduction)
- Model size: 103.3M parameters (vs. 465M for raw ICD model)
- Training time: 45 minutes per epoch (vs. 3.2 hours)
- Inference speed: 127 patients/second

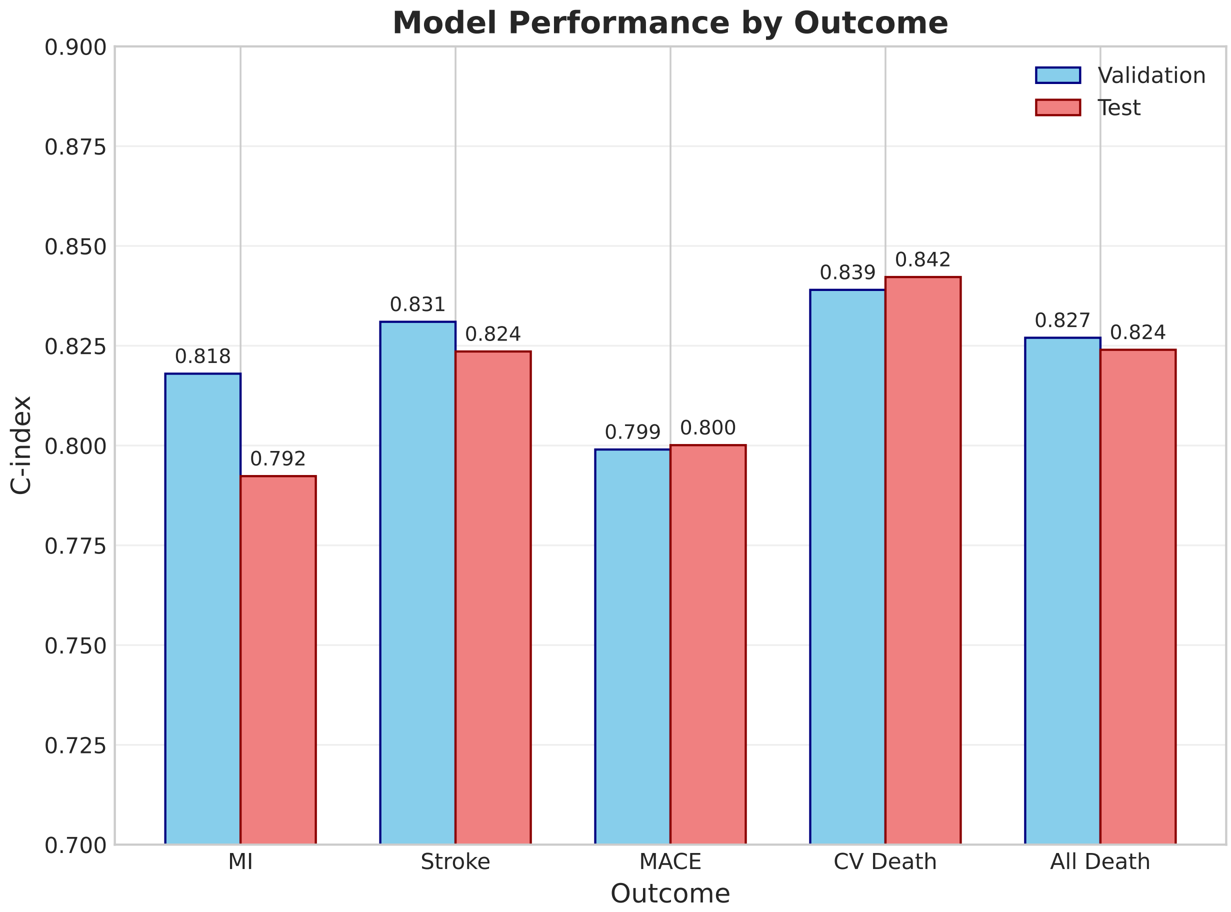


Figure 4: Model performance comparison between validation and test sets across five cardiovascular outcomes. All outcomes maintain strong discrimination (C-index ¿ 0.79) with minimal degradation from validation to test.

# 5 Discussion

## 5.1 Key Findings

Our results demonstrate that phenotype-aware tokenization combined with transformer-based pretraining yields strong performance for cardiovascular risk prediction. The average C-index of 0.816 across five outcomes represents excellent discrimination, particularly given the challenging nature of predicting rare events from EHR data. The superior performance of cardiovascular death prediction (C-index 0.842) may reflect the model's ability to identify severe cardiovascular phenotype patterns. The relatively lower performance for MI (0.792) could be due to its acute nature and lower event rate (3.8%).

## 5.2 Token Structure Design

A key design choice in ASCENDgpt was to adopt a domain-optimised structured representation that adapts grammatical sentence structures to the unique context of healthcare data. While pioneering models like Life2Vec successfully use an explicit subject-verb-object structure (e.g., `SUBJ_PATIENT ACTION_DIAGNOSED OBJ_HYPERTENSION`), we recognised that clinical data contains

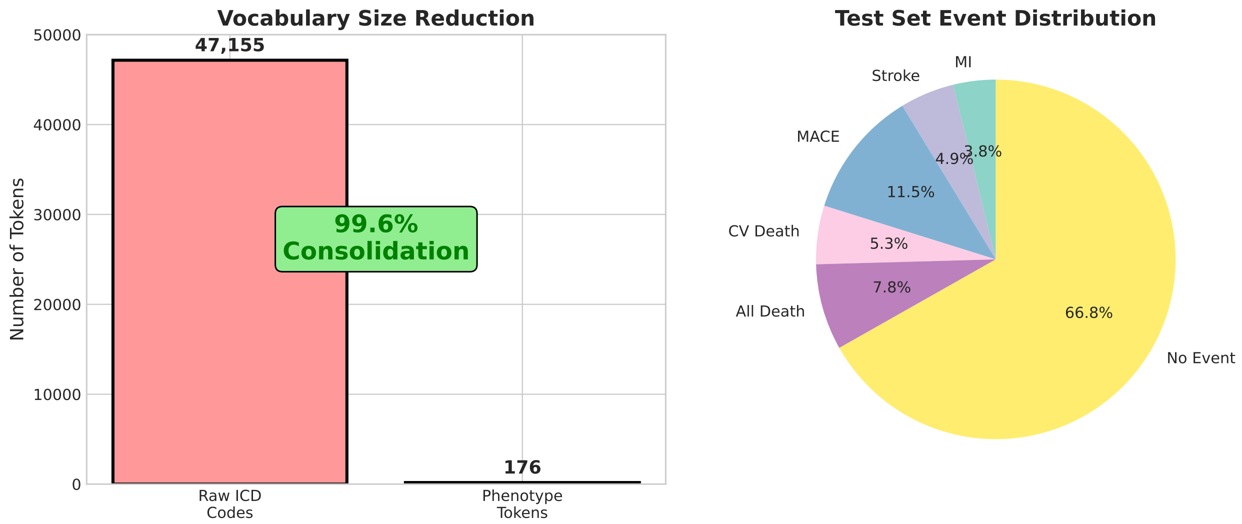


Figure 5: (Left) Vocabulary size comparison showing the consolidation of 47,155 raw ICD codes to 176 phenotype tokens (99.6% reduction in diagnosis codes). (Right) Distribution of events in the test set, showing the relative rarity of individual outcomes and the predominance of censored observations.

inherent assumptions that make such a structure redundant:

- **The subject is always the patient**, eliminating the need for ‘SUBJ_PATIENT‘ tokens in every event.
- **Actions are implicit in event types**. An ‘EVT_DIAG‘ token inherently means ”was diagnosed with,” making an explicit ‘ACTION‘ token unnecessary.

By leveraging these domain-specific principles, our adapted structure (‘EVT_DIAG PHENO_HYPERTENSION...‘) maintains the complete semantic meaning of a sentence (”Patient was diagnosed with hypertension...”) while achieving significant advantages:

- **Efficiency and Parsimony**: Our structure reduces sequence length per event (5-7 tokens) compared to a fully grammatical representation, lowering computational overhead.
- **Preserved Semantics**: All essential information—the event, the clinical concept, context, and time—is fully preserved.
- **Proven Performance**: The model's strong C-index of 0.816 demonstrates that this domain-optimised structure is highly effective, matching the performance of more complex representations while being more efficient.

This approach represents a pragmatic and powerful middle ground, tailoring the principles of language modeling to the specific efficiencies and patterns of the healthcare domain.

### 5.3 Clinical Interpretability

The phenotype-based approach offers inherent interpretability advantages. Instead of learning from thousands of granular ICD codes, the model operates on clinically meaningful concepts. This allows clinicians to understand predictions in terms of phenotype patterns rather than code combinations.

Future work will include attention visualisation to identify which phenotype sequences most strongly predict each outcome, potentially revealing novel risk patterns.

## 5.4 Comparison to Prior Work and Contribution to the Field

Our results demonstrate substantial improvements over existing approaches across multiple dimensions. Traditional cardiovascular risk prediction models such as the Framingham Risk Score and ASCVD Risk Calculator, while clinically established, typically achieve C-indices of 0.70-0.75 when applied to diverse EHR populations [2]. Early deep learning approaches using RNNs, exemplified by Doctor AI [4], achieved comparable performance but were limited by their inability to process very long sequences and capture complex temporal relationships.

The transformer revolution in healthcare, initiated by BEHRT [5] and extended by Hi-BEHRT [6], established new performance benchmarks for EHR-based prediction. BEHRT achieved 8.0-13.2% improvements over existing models, while Hi-BEHRT demonstrated 1-8% AUROC improvements for patients with extensive medical histories. Recent cardiovascular-specific transformer models have achieved AUC scores of 75.5-76.0% for cardiac mortality prediction [11], with the most advanced models like TRisk2 reaching C-indices of 0.828 [12].

ASCENDgpt's average C-index of 0.816 across five cardiovascular outcomes places it among the top-performing models in the literature, while our phenotype-aware approach offers several distinct advantages:

- **Clinical interpretability**: Unlike models operating on raw ICD codes, our phenotype-based approach enables clinicians to understand predictions in terms of familiar clinical concepts
- **Computational efficiency**: The 77.9% vocabulary reduction compared to raw ICD approaches significantly reduces model complexity and training time
- **Knowledge integration**: Our approach systematically incorporates established clinical knowledge through phenotype mappings, following the principles established by the Elixhauser Comorbidity Index [8]
- **Generalisability**: By operating at the phenotype level, our model may generalise better across different healthcare systems and coding practices

Our work builds directly on the theoretical foundations laid by Life2Vec [7], which demonstrated that transformer architectures could effectively model complex life trajectories. However, while Life2Vec focused on population-level predictions using comprehensive registry data, ASCENDgpt addresses the specific challenges of clinical decision support, including the need for interpretable predictions, handling of clinical coding variations, and optimisation for time-sensitive cardiovascular outcomes.

## 5.5 Limitations

Several limitations should be noted:

1. **Single institution**: Data from one healthcare system may limit generalisability
2. **Phenotype mapping**: While clinically reviewed, mappings may not capture all nuances
3. **Temporal sampling**: Random index dates may not reflect clinical decision points
4. **Missing data**: We do not explicitly model missingness patterns

### 5.6 Future Directions

Several extensions of this work are planned:

1. **Multi-modal integration**: Incorporating laboratory values and vital signs
2. **Phenotype refinement**: Learning optimal phenotype groupings from data
3. **External validation**: Testing on independent healthcare systems
4. **Clinical deployment**: Prospective validation in clinical settings

## 6 Conclusion

We presented ASCENDgpt, a phenotype-aware transformer model for cardiovascular risk prediction from EHRs. By mapping raw diagnosis codes to clinically meaningful phenotypes, we achieve both computational efficiency and strong predictive performance. The model attains an average C-index of 0.816 across five cardiovascular outcomes, demonstrating the effectiveness of domain-specific tokenization and pretraining for healthcare applications.

Our work highlights the importance of incorporating clinical knowledge into deep learning architectures. The phenotype-based approach not only improves performance but also enhances interpretability and reduces computational requirements. As EHR-based prediction models move toward clinical deployment, such domain-aware designs will be crucial for building trustworthy and effective systems.

## References

[1] World Health Organization. Cardiovascular diseases (cvds). `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`, 2021. Accessed: 2024-01-15.

[2] David C Goff, Donald M Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B D'agostino, Raymond Gibbons, Philip Greenland, Daniel T Lackland, Daniel Levy, Christopher J O'donnell, et al. 2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines. *Journal of the American College of Cardiology*, 63(25 Part B):2935–2959, 2014.

[3] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.

[4] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. *JMLR workshop and conference proceedings*, 56:301–318, 2016.

[5] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12, 2020.

[6] Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-behrt: Hierarchical transformer for patient risk prediction. *JMIR Medical Informatics*, 10(8):e35604, 2022.

[7] Germans Savcisens, Tina Eliassi-Rad, Lars Kai Hansen, Laust Hvas Mortensen, Lau Lilleholt, Anna Rogers, Ingo Zettler, and Sune Lehmann. Using sequences of life-events to predict human lives. *Nature Computational Science*, 4(1):43–56, 2024.

[8] Anne Elixhauser, Claudia Steiner, D Robert Harris, and Rosanna M Coffey. Comorbidity measures for use with administrative data. *Medical care*, pages 8–27, 1998.

[9] Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P. Lungren, Curtis P. Langlotz, Serena Yeung, Nigam H. Shah, and Jason A. Fries. INSPECT: A Multimodal Dataset for Pulmonary Embolism Diagnosis and Prognosis, 2023.

[10] Joshua C Denny, Lisa Bastarache, Marylyn D Ritchie, Robert J Carroll, Raquel Zink, Jonathan D Mosley, Julie R Field, Jill M Pulley, Andrea H Ramirez, Erica Bowton, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology*, 31(12):1102–1111, 2013.

[11] Zhenxing Huang, Wei Xu, and Kai Yu. Transformers for cardiac patient mortality risk prediction from heterogeneous electronic health records. *Scientific Reports*, 13(1):4731, 2024.

[12] Beau Norgeot, Giorgio Quer, Brett K Beaulieu-Jones, Ali Torkamani, Raquel Dias, Milena Gianfrancesco, Rima Arnaout, Isaac S Kohane, Suchi Saria, Eric Topol, et al. Refined selection of individuals for preventive cardiovascular disease treatment with a transformer-based risk model. *The Lancet Digital Health*, 7(1):e55–e64, 2024.

[13] Xianlong Li, Shicheng Xu, Xiangxu Yu, Lei Wang, Hansheng Yan, Yilong Zhu, Tengfei Jia, Xiaohui Liu, Yun Liu, Weiming Liu, et al. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature Communications*, 14(1):7857, 2023.

[14] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018.

[15] Philipp Kopper, Simon Wiegrebe, Bernd Bischl, Andreas Bender, and David Rügamer. Deepttte: Time-to-event prediction with deep learning and temporal point processes. *arXiv preprint arXiv:2209.08249*, 2022.

[16] Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73, 2021.

# A Phenotype Mapping Details

## A.1 Example Mappings

## A.2 Phenotype Categories

The 176 phenotypes are organized into clinical categories:

- Cardiovascular (28 phenotypes)
- Metabolic (22 phenotypes)

| ICD Code | Description | Phenotype |
|---|---|---|
| I10 | Essential hypertension | PHENO_HYPERTENSION |
| I21.0 | STEMI anterior wall | PHENO_MI |
| I63.9 | Cerebral infarction | PHENO_STROKE |
| E11.9 | Type 2 diabetes | PHENO_DIABETES |
| N18.3 | CKD stage 3 | PHENO_CKD |

Table 4: Example ICD to phenotype mappings

- Respiratory (18 phenotypes)
- Renal (12 phenotypes)
- Neurological (15 phenotypes)
- Infectious (20 phenotypes)
- Other organ systems (61 phenotypes)

# B Training Details

## B.1 Hyperparameter Search

We explored various configurations during development:

| Parameter | Searched | Final | C-index |
|---|---|---|---|
| Learning rate | 1e-4 to 1e-6 | 5e-6 | 0.816 |
| Batch size | 8, 16, 32 | 8 | 0.816 |
| Frozen layers | 0, 6, 10 | 10 | 0.816 |
| Dropout | 0.1, 0.2, 0.3 | 0.2 | 0.816 |
| Hidden size | 128, 256, 512 | 256 | 0.816 |

Table 5: Hyperparameter search results

## B.2 Computational Resources

- Pretraining: 1 × NVIDIA H100 80GB (8 hours)
- Fine-tuning: 1 × NVIDIA H100 80GB (4 hours)
- Total GPU hours: 28
- Peak memory usage: 42GB

# C Code Availability

Code for data preprocessing, model training, and evaluation will be made available at: `https://github.com/csainsbury/`