**## 1a. Patient Funnel**

We constructed two cohorts—SPS and control—from a claims database with defined time coverage. For both cohorts, patients were required to be aged 18+ at the starting point of their clinical trajectory in our database (**Forian Version... need to check with Ambit**) and have continuous enrollment check for both cohort with 2 years looking back and 12 month time window, filtering out patient who has at least one clinical records every 12 months 2 years prior to the index SPS or mimic diagnosis. To further ensure the data quality, we only keep those patients in SPS cohort with at least one clinical record within one year following the index diagnosis and no mimic codes are observed within this time period. For the control cohort, we included individuals with at least two control condition codes appearing ≥30 days apart (based on an updated control condition code list), and no diagnosis of G25.82 in the entire trajectory. To ensure misdiagnosed SPS patients were not included in the control group, we required at least 2 years of follow-up after the index diagnosis, allowing time for potential reclassification.

**## 1b. Downsampling**

Due to the computational power, we randomly sampled 30,000 control patients while keep all the SPS patients to train the model, resulting in a 1:9.6 positive-to-negative ratio.

**## 2. Use of ICD-10 codes**

We convert each event with a level-5 ICD-10, and eventually, we get 12867 codes in the beginning. Considering that a huge amount of individual-specific code and irrelevant codes are included, we made a further selection with the following criteria: 1. the top 1k codes present in the largest number of patients in the SPS group. 2. top 1k codes present in the largest number of patients in the control group. 3. top 1k codes presenting in the largest number of patients in both groups combined. 4. top 1k codes with absolute value of log-transformed code frequency ratio between groups, penalized by extremely rare observations of some codes in the SPS group because of the class imbalance. The final code size is 1734.

**## 3. Construction of trajectories**

patient trajectory is the sequence of time and clinical records (in our work, selected level-5 ICD-10 codes only) pairs. For the SPS group, we truncate the patient trajectory at the index diagnosis, while for the control group, we truncate the patient trajectory 2 years prior to the end of the entire trajectory they have in the database. This avoids the uncertainty of cases in which SPS diagnosis appears on control patient trajectories within the 2-year buffer time period. Then, we set a limit for the length of trajectories: we filter out samples with fewer than 30 data points. For samples that have trajectories longer than 225, we only kept the latest 225 data points, as the codes observed early on may be irrelevant. Variable-length patient trajectories require standardization to fixed-length sequences for batch processing while preserving authentic medical information. Sequences shorter than the maximum length (225 events) undergo pre-padding with designated tokens: empty strings for diagnostic codes (mapped to index 0) and -1 values for corresponding age positions. Longer sequences are pre-truncated to retain the most recent medical events, prioritizing recent disease trajectory information most relevant for SPS risk assessment.

**## 4. Model Architecture**

The Transformer model was employed to learn from the patient trajectory we constructed1–3. We embedded each medical event denoted with a level-5 ICD-10 diagnosis code from a patient's medical trajectory in a continuous 128-dimensional latent space. Such embedding is data-driven and trained jointly with other components of the model through supervised learning. Each diagnosis code is mapped to a learnable dense vector representation via a neural network embedding layer, where the embedding weights are optimized during training based on the SPS prediction task.

After embedding, temporal information is incorporated through age-based positional encoding. Patient age at each medical event (measured in days since birth) is encoded using cosine-based functions with 32 different frequency components, with periods ranging from 1 day to 7,300 days. This temporal encoding captures both short-term and long-term temporal patterns in the medical trajectory.

The token embeddings and temporal encodings are then fused using Feature-wise Linear Modulation (FiLM), where two learned linear transformations project the age encodings to modulate and shift the token embeddings. This combined representation is subsequently processed through 4 stacked Transformer blocks with multi-head attention mechanisms to capture complex dependencies across the medical history. The encoded features are aggregated using global average pooling and passed through a cumulative probability layer to predict SPS risk at multiple time horizons (6, 12, 24, 36, and 60 months).

After the preprocessing stage, the model parameter was trained with focal loss 4 that is tailored for unbalanced class

Formula: FL(pt)=−αt(1−pt)γlog(pt).

pt represents the predicted probability for the true class, αt provides class-specific weighting and γ controls the modulation strength. This formulation substantially reduces loss contribution from high-confidence predictions, enabling the model to concentrate learning capacity on challenging boundary cases while improving sensitivity for SPS detection across the severe class imbalance.

**## 5. Model Evaluation**

Robust performance assessment employs multiple trajectory sampling to account for the stochastic nature of sequence processing. For each patient in the test set, ten different trajectory samples are generated through random subsequence selection from their complete medical history, mimicking the trajectory augmentation used during training. Each trajectory sample represents a different temporal window of the patient's medical experience, enabling demonstration of consistent risk assessment across various historical perspectives. Patient-level predictions are computed by averaging risk scores across the ten trajectory samples, reducing prediction variance while preserving the benefits of temporal augmentation.

Model performance was evaluated using standard binary classification metrics. For rare disease detection, we prioritize the successful prediction of the positive class, and therefore, the area under the precision-recall curve (AUPRC) is our primary interest. Precision and recall are defined as:

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

where TP, FP, TN, FN are true positive, false positive, true negative, false negative, respectively. AUPRCs are calculated separately over each time horizon (6, 12, 24, 36, and 60 months).

**## 6. Model Interpretability**

To interpret the model, we employed Integrated Gradients (IG) to attribute prediction scores to individual medical codes in patient trajectories. IG computes the integral of gradients along a straight path from a baseline input to the actual input, providing an attribution score that satisfies key axioms, including sensitivity and implementation invariance. **For each patient, we identified the medical codes with the highest attribution scores to understand which diagnoses most strongly influenced the model's SPS risk predictions (the method here not finalized).** This analysis revealed clinically meaningful patterns, with neurological and autoimmune conditions showing consistently high importance scores across positive predictions.

**References:**

1. Vaswani, A. *et al.* Attention is All you Need. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).

2. Placido, D. *et al.* A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nat. Med.* **29**, 1113–1122 (2023).

3. Alsentzer, E. *et al.* Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. 2022.12.07.22283238 Preprint at https://doi.org/10.1101/2022.12.07.22283238 (2024).

4. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 318–327 (2020).