

Foresight

A national-scale foundation model for
generative medical event prediction,
across the COVID pandemic

Presenter: Simon Ellershaw

Collaborators: Chris Tomlinson, Richard Dobson

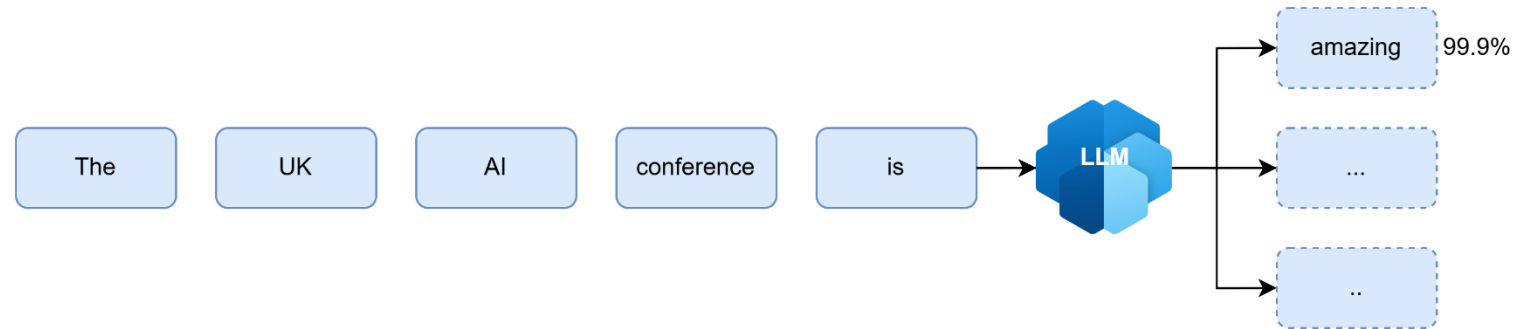


Foresight is to EHR,

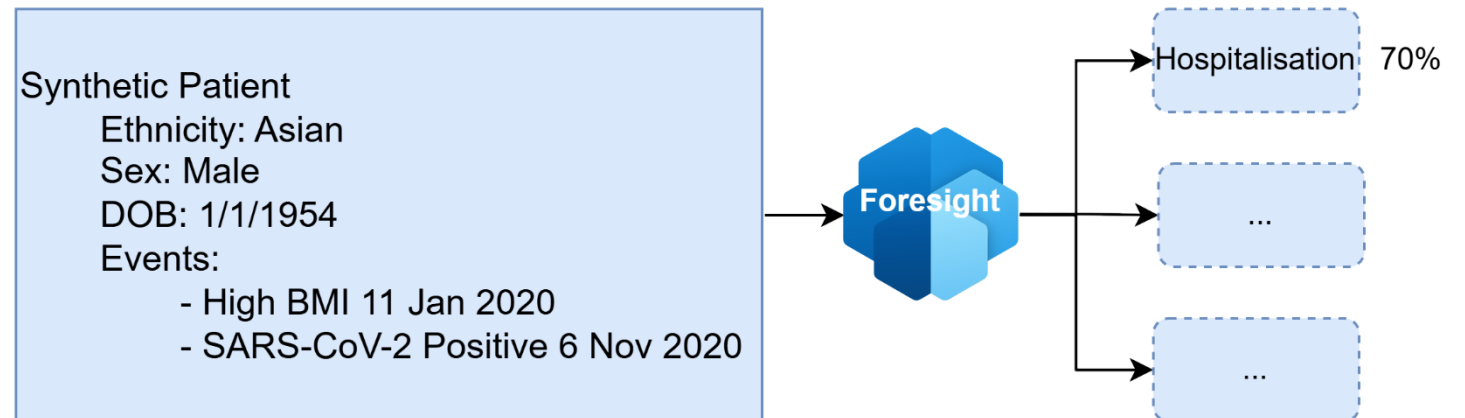
as GPT is to text

Next Token Prediction

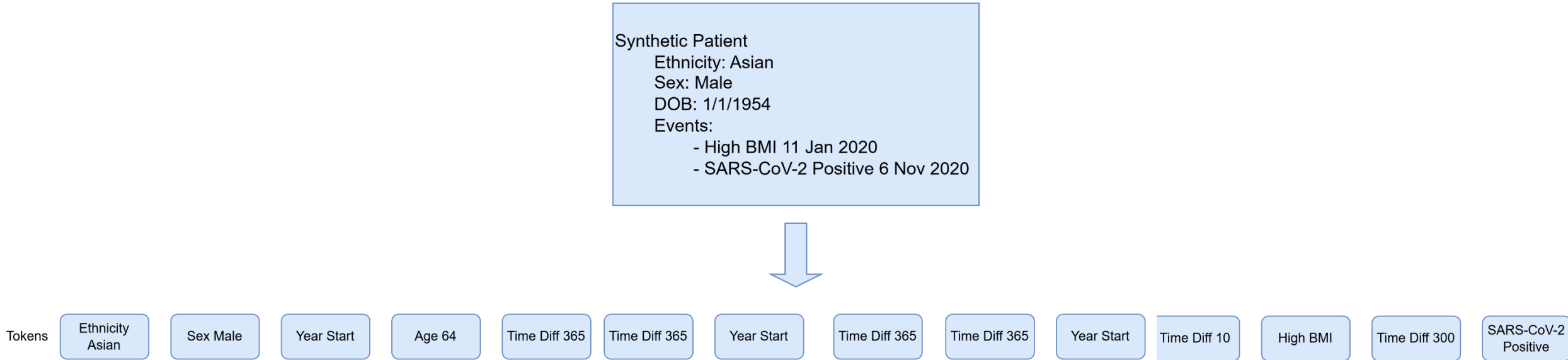
Text



EHR



Tokenization

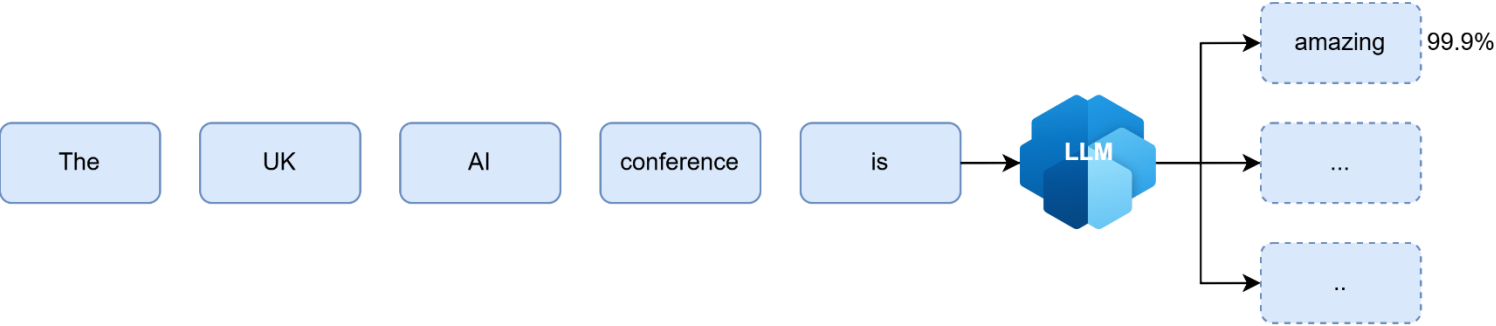


Minimum schema required to encapsulate

- Static variables
- Absolute time (e.g. 2018 or 2020)
 - Extending to years not seen in training data
- Progressing age of the patient
- Bounds the number of time tokens
- An arbitrary number of coded medical events

Next Token Prediction

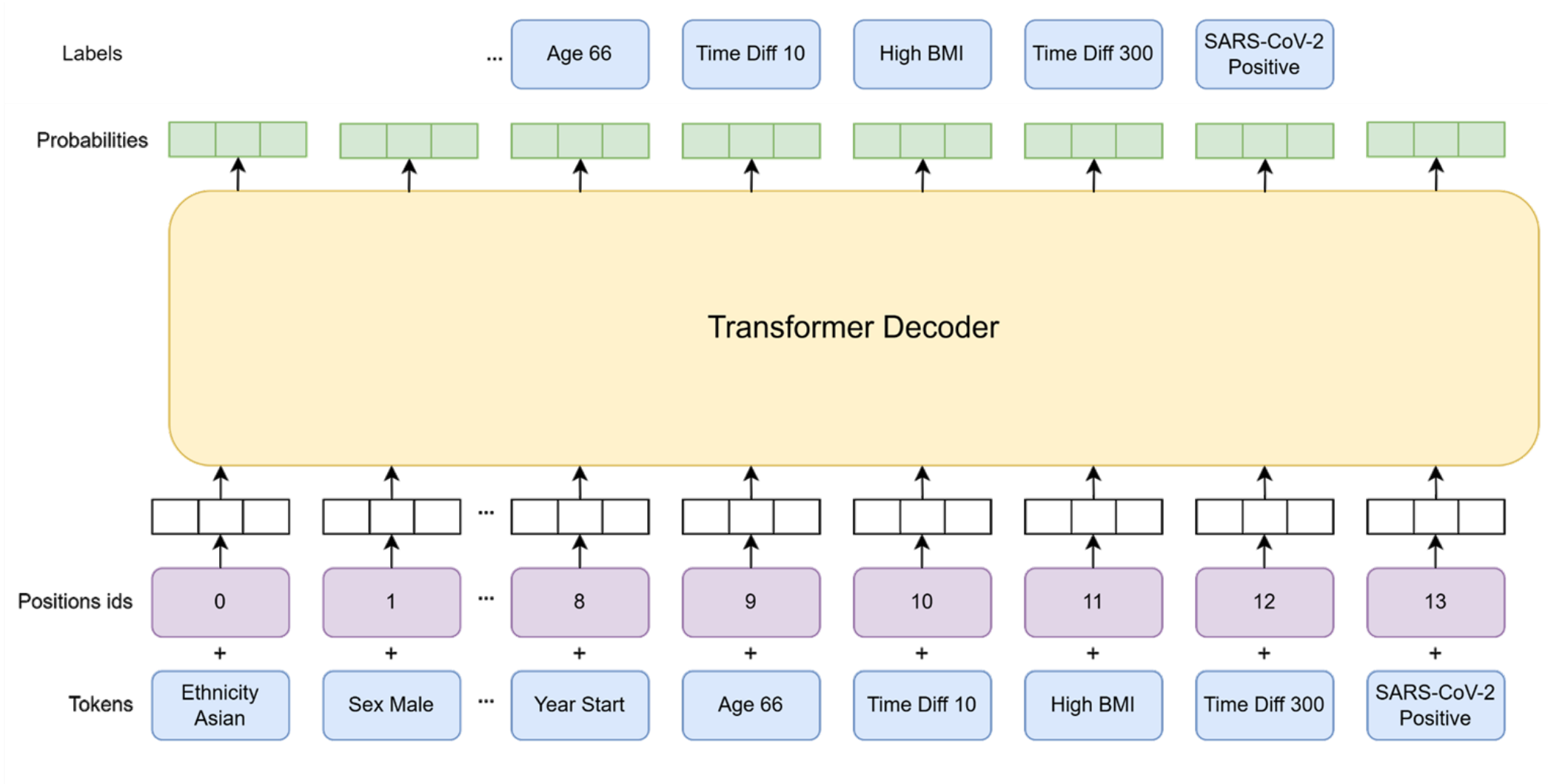
Text



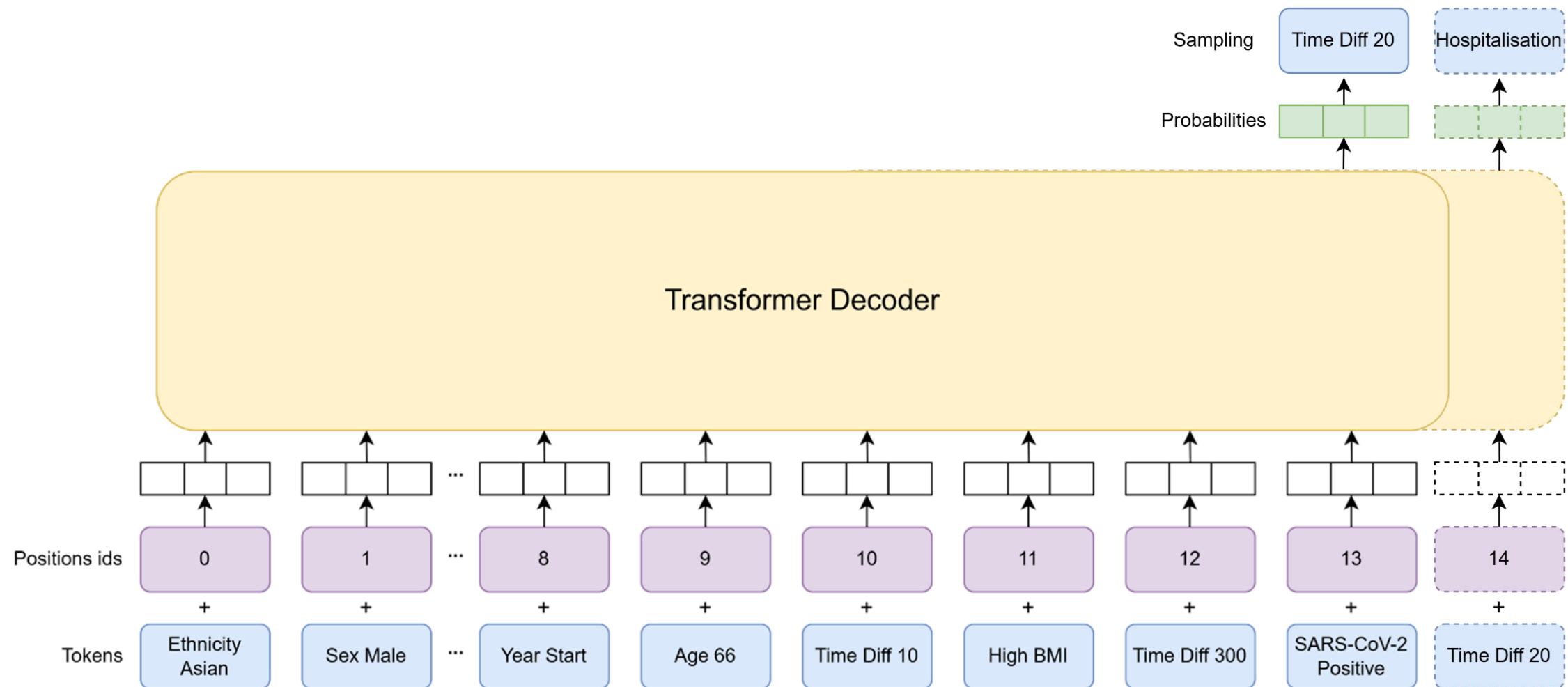
EHR



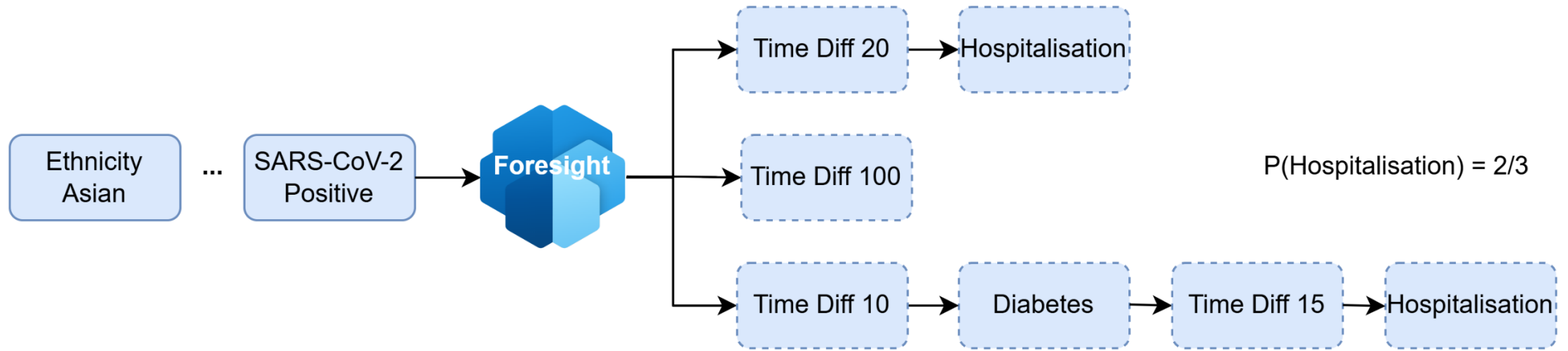
Model Training



Model Inference



Model Inference- Multiple Trajectories



What is Foresight?

Given a patient's coded medical history, the model can:

- Predict the likelihood of **any outcome** in the model's vocabulary
- Forecast these outcomes at **any point in time**

Evaluation

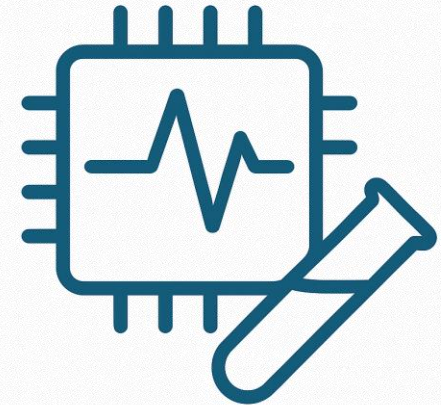
Potential Clinical Use Cases



**Operational
Forecasting**



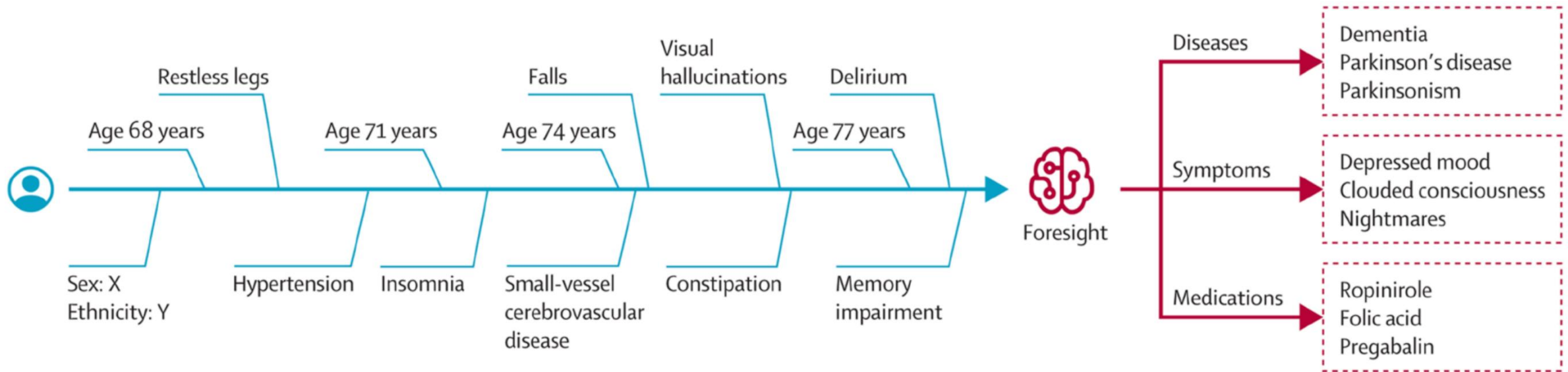
**Preventive
Care**



**In Silco
Clinical Trials**

Foresight V1

Learning from past patient data, to predict the future



THE LANCET
Digital Health

Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study

Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Baston, Jack Ross, Esther Idowu, James T Teo*, Richard J B Dobson*

**Foresight-v1 trained on 1.5M patients at Kings College Hospital*



Foundation model recipe

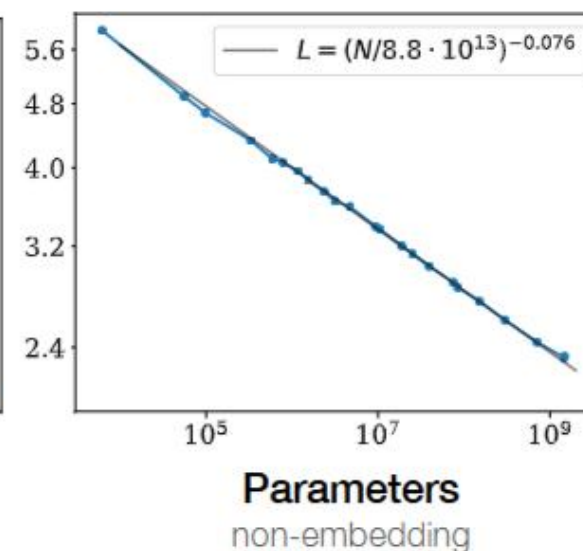
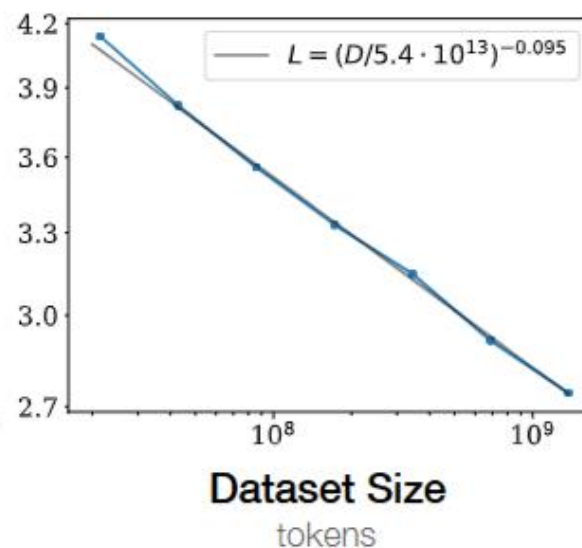
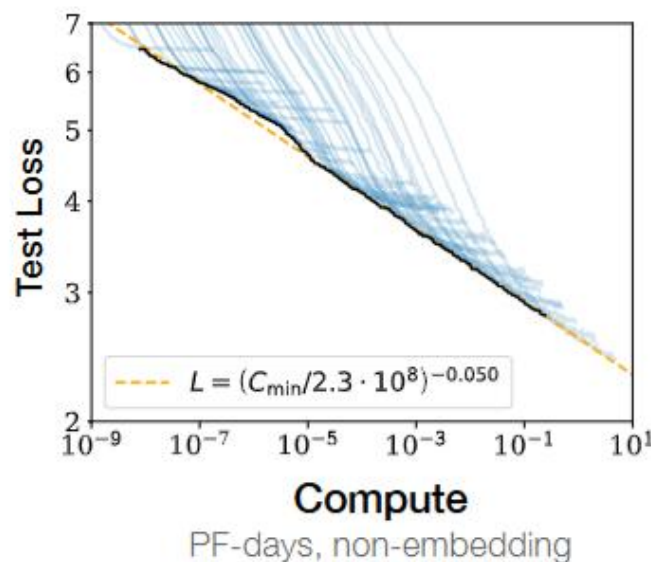
Big
Data

x

Big
Compute

x

Expertise



Foundation model recipe

Big Data* x Big Compute** x Expertise



Part of the
NHS Research Secure Data
Environment Network

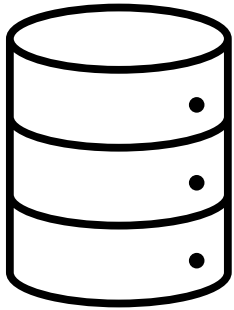


databricks



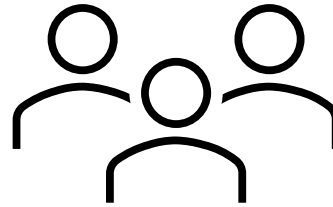
- * Data currently available for COVID-19 related research only
 - ** Industry partners cannot access NHS data, or the model, nor influence research questions
- The model, and its predictions, remain within the SDE, restricted to the same DSA

Data



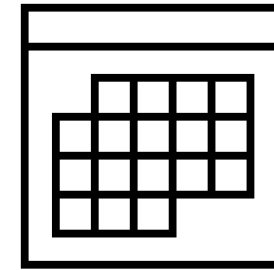
8

Datasets*



57M

Individuals**



~8B

Events



Part of the
NHS Research Secure Data
Environment Network



**British Heart Foundation
Data Science Centre**

Led by Health Data Research UK

* Data currently available for COVID-19 related research only

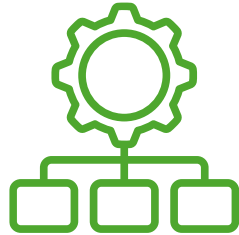
** De-identified data

Foresight: Progress

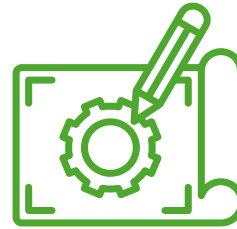
Foresight-SDE is the world's first population-scale EHR foundation model



Governance



Infrastructure



Training



Evaluation



**Translation
(future)**

Series of firsts



Part of the
**NHS Research Secure Data
Environment Network**



**British Heart Foundation
Data Science Centre**

Led by Health Data Research UK

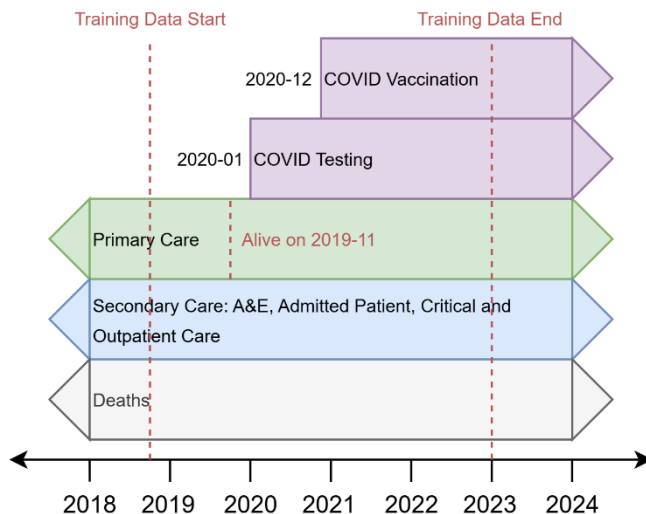
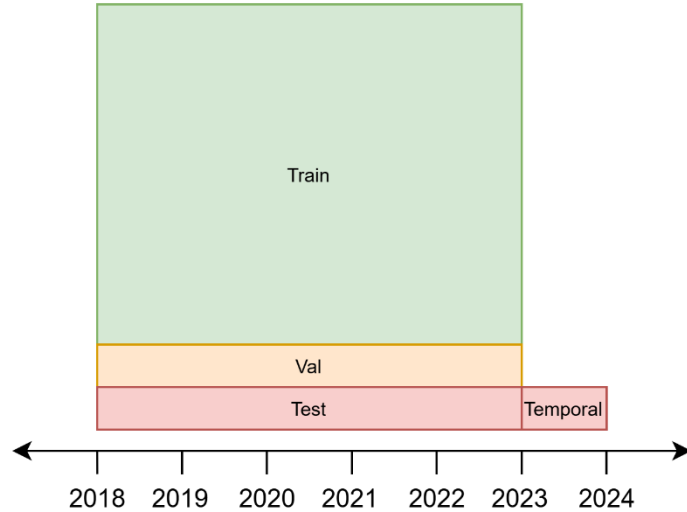


databricks



Training

Dataset and model overview



Model architecture

- 243 million parameter llama architecture
- Custom vocabulary of 40k tokens
 - Including ICD-10, OPCS-4, SNOMED-CT
- Randomly initialized

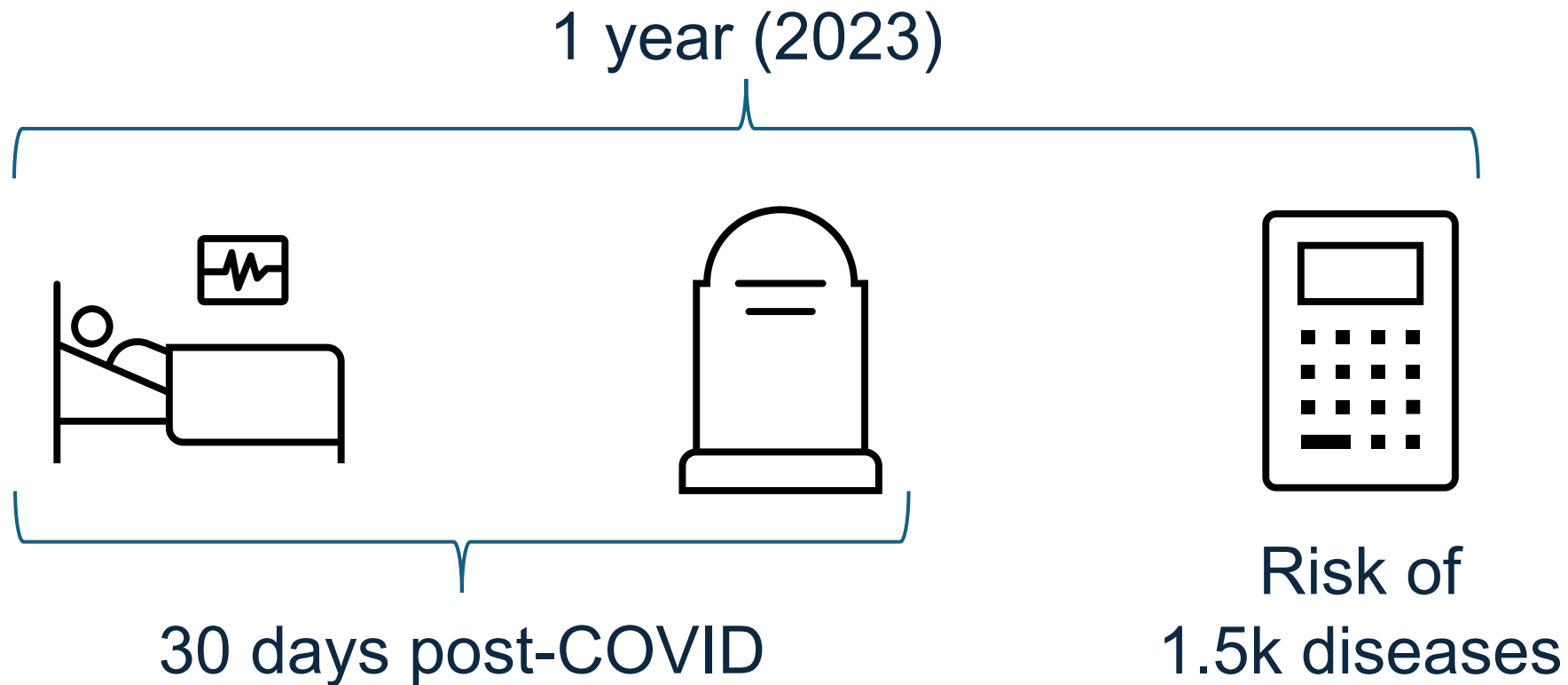


Training Setup

- 8 x A10 Nvidia GPUs (located in UK)
- Completed in 4 days

Evaluation

Direct & Indirect effects of COVID-19



Discrimination

- ROC Curve + AUC
- PR Curve + AUC

Calibration

- Calibration Curve
- Brier Scores

Subgroup analysis

- Age
- Sex
- Ethnicity

Timeline properties

- Number of events
- Temporal/Phase of pandemic




Baseline

- Bag-of-words logistic regression

Data currently available for COVID-19 related research only

On May 29th the British Medical Association and Royal College of General Practitioners Joint GP IT Committee emailed NHS England in relation to Foresight to recommend that they **self-refer** to the **ICO** to ascertain principles of **GDPR** have been **upheld** and that the **project is paused** whilst that process is ongoing.

Project Aims

1. Show that **right now** the **NHS** has the **unique ability** to use national routinely collected data to **train AI models** at scale 
2. Show the effectiveness of this approach in **predicting direct and indirect COVID-19 outcomes** 
3. **Stimulate the debate** on how the NHS should use the data it collects to optimize all patient care 



Questions

simon.ellershaw.20@ucl.ac.uk



Part of the
**NHS Research Secure Data
Environment Network**



**British Heart Foundation
Data Science Centre**

Led by Health Data Research UK

