Fairness Innovation Challenge

Technical brief

Use-Cases

Challenge participants must demonstrate how they will use the funding to develop and implement their solution in practice for their chosen use-case.

Your project must focus on one of the following use-cases:

1. Provided healthcare use-case
2. Open use-case

1. Healthcare

This use-case asks participants to submit fairness solutions to address bias and discrimination in the CogStack Foresight model developed by Kings Health Partners and Health Data Research UK, with the support of NHS AI Lab. This is a generative AI model for predicting patient outcomes based on Electronic Health Records.

What is CogStack?

CogStack is a platform that has been deployed in several NHS Hospitals and includes tools for unstructured (text) health data centralisation, natural language processing for curation as well as generative AI for longitudinal data analytics, forecasting and generation. The latter components have been applied to large real-world secondary care health datasets (>1m individuals in a mix of urban and suburban ethnically and socioeconomically diverse populations).

What is Foresight?

Foresight is a generative AI component of the CogStack platform, developed from secondary care data in two CogStack instances (one Foresight model trained at Kings College Hospital and one trained at The South London and Maudsley mental health Hospitals). Further models are trainable at CogStack instances in other NHS Trusts. This generative AI is a Generative Pretrained Transformer (GPT) model which can forecast next diagnostic codes (and any other standardised medical codes including medications and symptoms) based on their source dataset. Foresight can also generate synthetic longitudinal health records that match the probability distributions of the source data, allowing pilots on synthetic data without direct access to private data. The generated health data follows international health data standards (SNOMED-CT) with interoperability to ICD10.

Potential risks to fairness

As these AI models have been trained on real-world data, they contain biases of their historical datasets, including demographic biases, styles of historical practice and biased missingness from data capture. Additionally, some biases will have been introduced by the privacy-preserving process (e.g. all conditions n<100 out of 1million have been excluded from the training dataset), meaning that the multitude of Rare Diseases (prevalence <1 per 100,000) are excluded unless the source NHS institutions are national centres for those conditions.

Some biases are desirable (e.g. older patients are more likely to get a particular disease, so any AI model should reflect that), while others are only desirable in certain contexts (e.g. patients of a certain socioeconomic demographic having higher rates of diabetes or HIV). Furthermore, others are likely to be undesirable in many contexts (e.g. predicting health conditions associated with criminal activities in certain demographics). Inequalities of historical healthcare delivery may also be captured as biases, in situations when specific ethnic groups had less healthcare access in the past for example, or there had been changes in treatment guidelines (e.g. certain treatments being less prioritised historically).

Description of training algorithms

Foresight is a novel transformer-based pipeline for modelling biomedical concepts from clinical narratives. It uses named entity recognition and linking tools to convert document text into structured, coded concepts, followed by providing probabilistic forecasts for future medical events such as disorders, substances, procedures and findings. It is built on top of the Generative Pretrained Transformer (GPT) v2 architecture enabling causal language modelling, the main difference between a standard language model and Foresight is that our tokens represent biomedical concepts instead of words (or subwords).

Description of the dataset (e.g., variables, size)

We used two electronic health record (EHR) datasets to train/test Foresight:

1) King's College Hospital (KCH) NHS Foundation Trust - all available free text from EHRs from 1999 to January 2021). We collected a total of 18,436,789 documents from 1,459,802 patients (both inpatients and outpatients).

2) South London and Maudsley (SLaM) NHS Foundation Trust Hospitals - all available free text for patients with a serious mental illness diagnosed prior to August 2019. SLaM is one of Europe's largest providers of secondary mental healthcare, serving a geographical catchment of approximately 1.32million residents, and providing almost complete coverage of secondary mental healthcare provision to all age groups. We collected 14,995,092 documents from 27,929 patients with a serious mental illness diagnosis.

The Medical Concept Annotation Toolkit (MedCAT) software was used to extract biomedical concepts from free text and link them to the SNOMED-CT UK Clinical Edition and Drug Extension concept database. SNOMED-CT is an international open health standard and an NHS-Digital-required standard, these codes are fully accessible from SNOMED international as well as from the NHS National Terminology Server which is accessible openly (https://termbrowser.nhs.uk/?). The dataset schema can be found on the Health Data Research Innovation Gateway (https://tinyurl.com/5ybv926x) Once the concepts were extracted, we removed all concepts that occurred <100 times (to remove rare concepts that could identify patients) and grouped them by patient and organised into a timeline (Table 1 and 2). There are 195,416 different biomedical concepts from SNOMED-CT that are organised as timelines, belonging to top level concept categories including Disorder, Substance, Finding and Procedures (See Table 3 for a summary of patient timeline characteristics). The datasets were split randomly into a train set (95%) and a test set (5%) for the Foresight algorithm.

	KCH		**SLaM**	
	Train	Test	Train	Test
Patients	710194	37301	21910	1155
Patients by Ethnicity				
Asian	34616 (5%)	1764 (5%)	1405 (6%)	63 (6%)
Black	131216 (18%)	6980 (19%)	4822 (22%)	281 (24%)
Mixed	8484 (1%)	441 (1%)	572 (3%)	28 (2%)

Other	34434 (5%)	1798 (5%)	4167 (19%)	213 (19%)
Unknown	154132 (22%)	8071 (21%)	1150 (5%)	48 (4%)
White	347312 (49%)	18247 (49%)	9794 (45%)	522 (45%)
Patients by Sex				
Female	381155 (54%)	19873 (53%)	10054 (46%)	544 (47%)
Male	328866 (46%)	17422 (47%)	11777 (54%)	607 (53%)
Unknown	173 (0%)	6 (0%)	79 (0%)	4 (0%)
Patients by Age				
0-18	119297 (14%)	6402 (14%)	1437 (4%)	81 (4%)
18-30	122137 (14%)	6435 (15%)	7372 (21%)	378 (20%)
30-41	138706 (16%)	7232 (17%)	9009 (26%)	500 (27%)
41-50	120187 (15%)	6390 (14%)	7283 (21%)	393 (21%)
51-64	161799 (19%)	8391 (19%)	6044 (18%)	345 (19%)
64+	183423 (22%)	9489 (21%)	3346 (10%)	170 (9%)

Table 1. Selected characteristics from King's College Hospital (KCH) and South London and The Maudsley (SLaM) after preprocessing and timeline creation.

	KCH	**SLaM**
Annotations (Unique)	56,736,380 (10512)	8,958,567 (2182)
Annotations per Semantic Type - Total (Unique)		

Disorder	19,003,851 (5632)	1,743,625 (674)
Substance	12,191,307 (1185)	2,245,368 (255)
Finding	17,282,165 (2868)	4,747,863 (929)
Procedure	3,056,147 (63)	45,189 (26)

Table 2. Four common clinically relevant semantic types after dataset annotation from KCH and SLaM. Counts are calculated after data pre-processing and timeline formation.

	KCH		**SLaM**	
	Train	Test	Train	Test
Mean Timeline Length in concepts (in years from first to last admission)	75 (3.3)	75 (3.3)	387 (6.9)	414 (7.3)
Mean Timeline Length by Ethnicity in concepts (in years from first to last admission)				
Asian	80 (3.6)	78 (3.5)	361 (6.9)	344 (7.4)*
Black	77 (4.7)	79 (4.6)	524 (8.9)	596 (9.2)
Mixed	55 (3.7)	58 (3.6)	516 (7.7)	307 (6.9)*
Other	66 (3.2)	65 (3.2)	372 (6.3)	367 (6.6)
Unknown	55 (2.1)	55 (2.0)	92 (1.6)	58 (1.0)*
White	86 (3.4)	85 (3.3)	357 (6.7)	382 (7.4)
Mean Timeline Length by Sex in concepts (in years from first to last admission)				

Female	74 (3.5)	74 (3.4)	369 (6.8)	394 (7.3)
Male	78 (3.2)	77 (3.2)	404 (7.0)	434 (7.4)
Unknown	88 (1.5)	16 (0.4)*	238 (5.0)*	109 (3.8)*
Mean Timeline Length by Age in concepts (in years from first to last admission)				
0-18	47 (3.2)	48 (3.2)	237 (1.6)	226 (1.6)*
18-30	43 (2.8)	42 (2.7)	359 (3.6)	373 (3.6)
30-41	50 (3.2)	49 (3.2)	405 (6.2)	438 (6.7)
41-50	67 (3.7)	66 (3.5)	414 (8.1)	448 (8.0)
51-64	87 (3.8)	88 (3.8)	432 (9.5)	444 (10.2)
64+	122 (3.4)	121 (3.4)	321 (7.7)	365 (8.4)
Mean Number of Concept by Type per Timeline				
Disorder	25	25	75	81
Substance	16	16	97	102
Finding	23	23	205	221
Procedure	4	4	2	2

Table 3. Selected timeline characteristics from KCH and SLaM. For *mean timeline length by age*, we took the most recent age of a patient and used that to determine the age group. * means the calculation was done on less than 100 timelines (patients).

The Foresight model was evaluated using the test sets from each individual NHS Foundation Trust (KCH, SLaM). We compared the performance of the model to the ground truth (i.e., what really happened in the hospital data) on the task of future biomedical concept prediction. The average precision when forecasting a disorder in a patient's future was 0.9 when considering the top 10 most confident model suggestions. Additionally, our

clinical team manually created 34 hypothetical patient vignettes and Foresight was used to predict the top 5 next concepts for each patient. Clinicians validated the top 5 predictions for relevancy, finding the top prediction was relevant in 97% of cases and 88% of the top 5 predictions were relevant.

Figure 1 shows examples of Foresight generated synthetic timelines for a 43-year-old, black, female (top – KCH model, middle – SLaM model). The right side of the timelines (orange part) was forecasted by Foresight to simulate the medical future (the distances in the figures do not represent real temporal distances, only the order of concepts in the timelines is relevant)

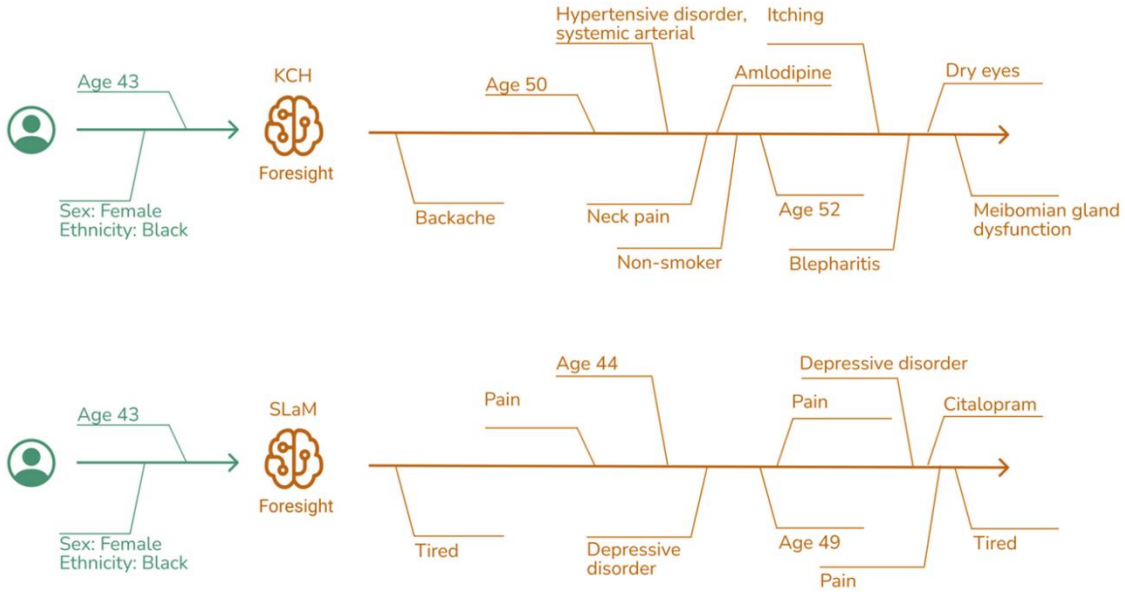

Figure 1: Examples of synthetic timelines generated by Foresight models trained on KCH (upper timeline) and SLaM Hospital data (lower timeline) when provided with the prompt 43-year-old, black, female.

How will selected participants be able to interact with the model/data?

There are multiple modes with differing levels of security and privacy:

- **Model access:** API-based access, Access to model within a TRE/SDE-environment, Licensed access to model within participant's own digital environment
- **Synthetic SNOMED-coded longitudinal data:** Access to synthetic dataset within a TRE/SDE-environment, Synthetic dataset delivered to participant's own digital environment

- **SNOMED-coded longitudinal data only:** Access to dataset only within a TRE/SDE-environment
- **Source data:** Not directly available; bring dockerised code to data, and code run by internal teams

Managed access with associate contracts for roles-based access to TRE/SDE environments would be expected.

Additional links for participants:

- https://doi.org/10.1016/j.artmed.2021.102083
- https://transform.england.nhs.uk/ai-lab/explore-all-resources/understand-ai/cogstack/
- https://arxiv.org/abs/2212.08072
- https://aiforhealthcare.substack.com/p/11530e1e-c46a-4ee6-bce0-766eed035e6d
- https://cogstack.org/cogstack-foresight-beta-launched/
- https://aiforhealthcare.substack.com/p/should-you-use-ai-to-diagnose-yourself?r=1icmzl&utm_campaign=post&utm_medium=web
- https://cogstack.org/new-cogstack-large-language-model-releases/
- https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000218

2. Other (Open submission)

For this option, you can propose your own use-case. This includes AI models, systems and solutions at different stages of development or deployment that are believed to be at risk of bias and discrimination.

If you are proposing your own use-case, you must provide additional information in your application about:

- **Background or Context**: What are you using an AI-enabled system for? What is the model? Why is it being used? What problems does it solve?
- **Potential risks to fairness**: what are the fairness challenges associated with this system for this specific use-case or context? Why is it difficult to make this system fairer?
- **Technical details**: describe the data set, including the size of the data set and any variables, as well as the learning algorithms used to train the models

Your use-case and proposed solutions will need to be published or shareable. This challenge is only open to use-cases that are transparent about their models, tools, and data, as well as the challenges and potential solutions to fairness.