Foresight-England: Development of a National-Scale Generative AI Model of Patient Electronic Health Records for General Medical Event Prediction across the COVID-19 Pandemic

Anonymous Author(s)
Affiliation
Address
email

Abstract

We developed Foresight-England (Foresight-E), the first national-scale generative foundation model of electronic health records (EHRs), to support COVID-19-related research. We evaluated its ability to model the direct and indirect effects of the COVID-19 pandemic. The 243M-parameter transformer decoder was trained from scratch using a cohort of 54.9 million routinely collected, de-identified, longitudinal EHRs, including primary and secondary care, national death registrations, and COVID-19 testing/vaccination data. Foresight-E models patient timelines autoregressively to enable zero-shot generative prediction across its vocabulary of $\sim$40,000 coded medical events. Our tokenisation scheme preserves the recorded clinical granularity of ICD-10, OPCS-4, and SNOMED CT codes, while jointly encoding absolute and relative temporal context. We designed and implemented an evaluation framework spanning 30-day COVID-19 hospitalisation and mortality using Brier scores and the area under the receiver operating characteristic (AUROC) and precision–recall (AUPRC) curves. We further evaluated the ability to model the pandemic's indirect effects by testing temporal generalisation on the held-out year of 2023, simulating prospective deployment. We benchmarked model performance against logistic regression and XGBoost baselines using a test set of 6.1 million patients. Following concerns raised by the British Medical Association and Royal College of General Practitioners' Joint GP IT Committee, NHS England has paused access to data for the Foresight project while a review is carried out. That pause means quantitative results are not available pending the outcome of ongoing discussions. Instead, we share our strategy for tokenisation, model architecture, training, inference, and evaluation, as a methodological template and a case study in the challenges of building population-scale, EHR foundation models and operationalising generative AI for national health systems.

1 Introduction

Prognostic clinical risk prediction models have traditionally relied on expert-defined, static features from electronic health records (EHRs) [1, 2]. While effective for narrow, well-defined tasks, these approaches cannot easily scale to predict the whole range of clinically relevant medical events or adapt to shifting dynamics of healthcare, such as the COVID-19 pandemic. They also often underutilise the rich temporal information in patients' longitudinal records.

Neural network transformer models offer a way to utilise the temporal and longitudinal information in EHRs for prediction, drawing on techniques initially developed for modelling language [3], first

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

with masked-token objectives such as BERT [4] and later with autoregressive next-token prediction as popularised by GPT-style models [5]. This shift enabled zero- and few-shot performance across diverse tasks without task-specific retraining, inspiring analogous efforts in healthcare [6–10]. Generative EHR models hold the promise of supporting early detection, risk stratification, and simulation of clinical scenarios in a single model, without the need to retrain for each outcome of interest.

A key barrier to realising this vision has been scale: most prior work is restricted to single institutions [9] or population subsets [6, 10], limiting performance [11] and fairness across population subgroups, so-called 'algorithmic bias' [12]. In England, due to the COVID-19 pandemic, permissions were granted to enable COVID-19 research on national-scale de-identified EHR data [13, 14]. This data encompasses primary care, secondary care, COVID-19 testing, vaccination, and mortality data [15], and is securely stored within the NHS England Secure Data Environment (NHSE SDE) [16]. Despite the potential of such a resource for the development and evaluation of AI models for COVID-19 research, computational restrictions within the NHSE SDE have limited previous projects to small cohorts [17, 18].

In this work, we developed Foresight-England (Foresight-E), a 243-million-parameter transformer trained entirely within the NHSE SDE on a national scale, de-identified EHRs. Designed for zero-shot prediction of COVID-19 outcomes, Foresight-E targets the whole English GP-registered population, enabling research into both direct COVID-19 outcomes (e.g., hospitalisation, mortality) and indirect effects on emergency admissions, all-cause mortality, and over 1,400 phenotypes in a temporally held-out test set (2023).

Following concerns raised by the British Medical Association and Royal College of General Practitioners' Joint GP IT Committee, NHS England has paused access to data for the Foresight project while a review is carried out [19]. That pause means quantitative results are not available pending the outcome of ongoing discussions. Therefore, we report here the data pipeline, tokenisation, architecture, training, inference, and evaluation framework underpinning the model, assessed against the TRIPOD+AI [20] and PROBAST-AI [21] reporting guidelines (see Appendix E and F).

We contribute the following:

1. Development of Foresight-E, the first population-scale foundation model of UK EHRs for COVID-19 research.
2. First demonstration of GPU-accelerated foundation model training on national-scale NHS data within the NHSE SDE.
3. A reproducible methodology for longitudinal EHR tokenisation and model development.
4. A six-million-patient evaluation framework to assess zero-shot prediction of COVID-19's direct and indirect effects.

Together, these provide both a technical blueprint for future EHR foundation models and a case study in the challenges of national-scale generative AI for healthcare.

2 Methods

We developed Foresight-E for COVID-19 research using linked, de-identified, routinely collected national datasets [22, 14]. All processing, model training, inference, and evaluation occurred entirely within the 'Five Safes' framework of the NHSE SDE [16].

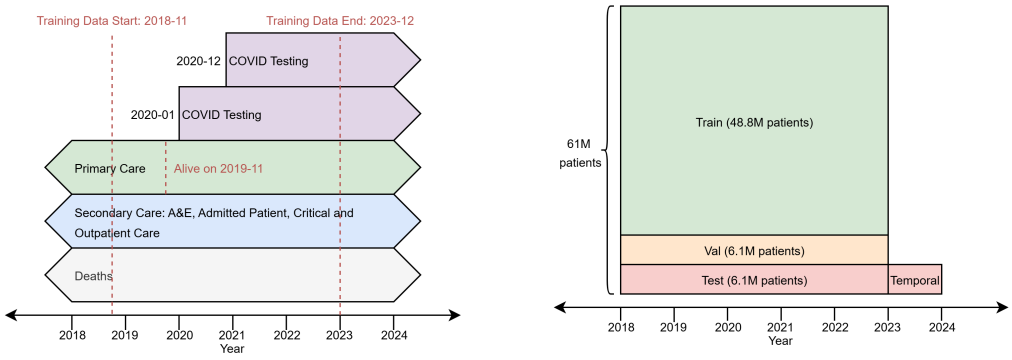
Here we outline the datasets, patient timeline construction, tokenisation, model architecture, training regime, and our inference and evaluation strategy, including uncertainty estimation and comparative baselines. Each step was designed to support robust, generalisable, zero-shot medical event prediction for the direct and indirect effects of COVID-19.

2.1 Data

We used eight linked national datasets covering 1 November 2018 to 31 December 2023 [22, 14], spanning primary care [23], secondary care [24], national death registrations [25], COVID-19 testing [26], and vaccination records [27] (full details in Table 1).

The base cohort comprised individuals alive on or after 1 November 2019 (primary care dataset inclusion date), with known age and sex, resident in England, GP-registered, and without conflicting death dates. We required patients to be at least one year old during training to ensure a minimum history and avoid immortal-time bias between the start of records (1 November 2018) and the first observed deaths (1 November 2019). This gave a nationally representative cohort of 61 million patients [15].

We created training (48.8M patients), validation (6.1M patients), and test sets (6.1M patients) via disjoint 10% patient samples. All 2023 events were reserved to test temporal generalisation, mimicking a prospective deployment on future, unseen data. Figure 1 shows temporal coverage and partitioning of the datasets.



(a) Linked primary/secondary care, deaths, testing, and vaccination datasets (1 Nov 2018–31 Dec 2023).

(b) Disjoint 10% validation/test cohorts; 2023 held out for temporal generalisation evaluation.

Figure 1: Datasets and cohort splits used in the training and evaluation of Foresight-E.

2.2 Patient Timelines

Each patient's history was encoded as a chronological sequence of dated, coded events, including diagnoses, procedures, medications, and other healthcare interactions. Clinical codes followed standard terminologies: ICD-10 [28] for Hospital Episode Statistics (HES) diagnoses and Office for National Statistics (ONS) causes of death, OPCS-4 [29] for HES procedures, and SNOMED CT [30] for General Practice Extraction Service Data for Pandemic Planning and Research (GDPPR) records and COVID-19 vaccinations. We supplemented these with custom tokens for events not covered by standard vocabularies, such as positive SARS-CoV-2 tests, hospital admission indicators, and primary or secondary diagnosis flags.

Because event timestamps were available only at day-level precision, we applied a consistent within-day ordering: COVID-19 vaccination; SARS-CoV-2 test; GDPPR events; HES Outpatients (OP); HES Accident & Emergency (A&E); HES Admitted Patient Care (APC); HES Critical Care (CC); and, lastly, death. Within each dataset, events were further ordered by admission date, then by primary diagnoses, secondary diagnoses, procedures, and finally by alphanumeric code.

Geographic (region) and deprivation (Indices of Multiple Deprivation) data were excluded from the dataset to avoid learning existing biases. Codes deemed sensitive were removed, using a codelist supplied by NHS England.

2.3 Tokenisation

We converted each patient's timeline into a sequence of discrete tokens representing clinical codes and temporal gaps. Each sequence began with static demographic tokens for sex (e.g., `SEX_FEMALE`) and ethnicity (e.g., `ETHNICITY_ASIAN`). We then added the sequence of codes, ordered as described in Section 2.2, from each patient's timeline as tokens. If two consecutive events occurred on different days, we inserted a time-difference token (e.g., `TIME_DIFFERENCE_1`) to represent the temporal gap in days; no time-difference token was added between consecutive same-day events. To provide absolute temporal context, we inserted a `YEAR_START` token at the beginning of each calendar year,

followed by an AGE_N token for the patient's integer age (or AGE_UNBORN if not yet born). These yearly markers ensured that time gaps never exceeded 366 days and allowed the representation of age-related and absolute temporal patterns to generalise temporally.

A lookup-based tokeniser mapped all unique tokens in the training data to integer IDs, including clinical codes, age (integer years and unborn), 1–366 day gaps, and special tokens for padding and sequence end. The resulting fixed tokeniser vocabulary contained ∼40,000 tokens. Unseen codes at inference time were dropped from the sequence without substitution. During training, sequences were left-truncated to a maximum length of 1,024 tokens to retain recent context. Truncation was performed at the event level, with retokenisation, to preserve demographic, year-start, and age tokens, and to adjust time-difference tokens. Sequences were then right-padded to the batch maximum length. For inference, the implementation of an on-GPU valid autoregressive truncation strategy was avoided to reduce complexity. Instead, the input length was constrained to $1{,}024 - L_{\text{forecast}}$, where L_{forecast} denotes the tokens allocated for forecast generation.

2.4 Training

Foresight-E was trained with an autoregressive next-token prediction objective, analogous to the paradigm used for LLMs [5, 8, 9]. All experiments were conducted entirely within the NHSE SDE [16] on 8× NVIDIA A10 GPUs (AWS g5.48xlarge instance in region eu-west-2) [31, 32].

2.4.1 Objective

Let N be the batch size, T the maximum sequence length, and V the vocabulary size. At each position t, given preceding tokens $y_{n,<t}$, the model outputs a distribution over V. A binary mask $m_{n,t}$ excludes padding and the initial demographic tokens from loss calculation. Excluding the initial tokens aims to mitigate the risk of biasing the model by learning to predict based solely on demographic features. The loss, $\mathcal{L}$, given the model weights, θ and the true token, $y_{n,t}$, is:

$$\mathcal{L} = -\frac{1}{\sum_{n,t} m_{n,t}} \sum_{n=1}^{N} \sum_{t=1}^{T} m_{n,t} \log P(y_{n,t} \mid y_{n,<t}, \theta) \tag{1}$$

2.4.2 Architecture

We adapted the Llama 2 transformer-decoder [33, 34] architecture with Rotary Positional Embeddings (ROPE) [35] and FlashAttention-2 [36]. Pretrained weights were not imported into the SDE due to governance restrictions and the custom vocabulary; therefore, training was conducted from scratch. To train a Llama decoder on NVIDIA A10 GPUs, the model architecture was scaled down to 243 million parameters, comprising 12 layers, a hidden size of 1024, a feed-forward size of 4096, 8 attention heads, and a 1024-token context window. Input/output embeddings were tied.

2.4.3 Training Protocol

We used bfloat16 mixed precision, attention dropout 0.1, gradient clipping (max norm 1.0), and weight decay 0.1. The Adam optimizer had $\beta_1 = 0.9$, $\beta_2 = 0.95$, with a linear warm-up over 3% of steps to a peak learning rate of 5×10^{-4}, then cosine decay. The global batch size was 128, achieved via 8-way data parallelism and gradient accumulation (factor 2). Sequences were right-padded to 1024 tokens. The model was trained for one epoch, with validation loss evaluated every 1,000 steps on 32k sampled sequences. Training was completed in 4 days. We did not conduct hyperparameter tuning due to resource constraints; instead, training parameters were selected based on standard defaults from prior literature [8, 37].

2.5 Zero-Shot Inference

After training, Foresight-E can predict future clinical events from patient histories without task-specific fine-tuning ('zero-shot inference'). At inference, the model is prompted with a tokenised patient timeline (see Section 2.3) and outputs a probability distribution for the next token. A token is sampled and appended to the input. This autoregressive process continues until a stopping condition is met: a specific event token (e.g., death), a maximum forecast horizon (e.g., 30 days, 1 year), or

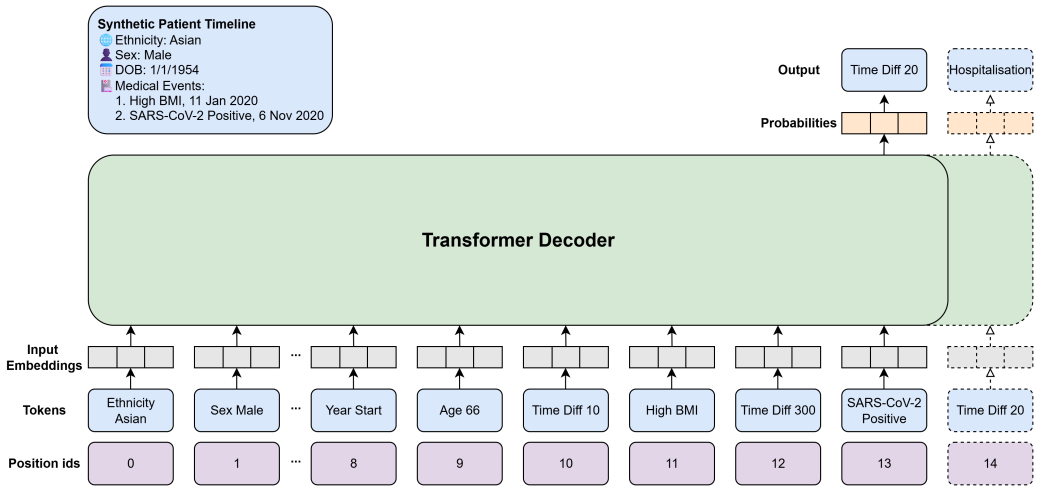


Figure 2: Illustration of idealized zero-shot inference on a synthetic patient timeline. Foresight-E receives a history ending with a positive SARS-CoV-2 test and predicts a hospital admission within the next 30 days. Tokens 2–7 are omitted for clarity.

a maximum of newly generated tokens (e.g., 300). Generated tokens are detokenized to output the predicted event timeline.

Greedy decoding produced repetitive outputs, so stochastic multinomial sampling was adopted to yield diverse and clinically plausible sequences. We did not use parameter-dependent sampling techniques such as top-k or temperature, as they would have required additional task-specific hyperparameter tuning. KV cache was used to increase generation speed.

To estimate prediction uncertainty, critical for safe and reliable clinical decision-making [38], we performed S independent rollouts per patient. In our experiments, $S = 48$, which maximised the available memory capacity of a single A10 GPU [32]. To scale inference across patients, we parallelised inference across the eight available GPUs [31].

The probability of event i occurring within t days was then calculated as [9]:

$$P(\text{Event}_{i,t}) = \frac{s_{i,t}}{S}, \tag{2}$$

where $s_{i,t}$ is the number of samples in which the event appears within the horizon. This formulation supports probability estimates for any of the ∼40,000 medical events seen during training, enabling probabilistic zero-shot predictions across diverse clinical outcomes.

2.6 Evaluation

We evaluated Foresight-E on two direct COVID-related outcomes: hospitalisation and death within 30 days of a positive SARS-CoV-2 test. Eligible patients in the test set had at least one positive test between 1 November 2018 and 1 December 2023. Timelines were truncated after the first positive test, and predictions were generated for the next 30 days. Hospitalisation was defined as an admission in HES APC (`APC` token), and death as any Office for National Statistics (ONS) death registry entry (`DEATH` token). Inference was completed in 2 days.

Secondary outcomes examined Foresight-E's potential to model the indirect effects of COVID-19 [39, 40] on: (i) emergency hospitalisation, (ii) all-cause mortality, and (iii) onset of over 1,400 Phecode-defined diseases, in the held-out year of 2023. Here, timelines were truncated after the 2023 year-start token and timelines predicted over a one-year horizon. Emergency admissions were identified via admission method codes in HES APC, and mortality as above. Phenome-wide disease onset was defined by the first occurrence of each Phecode, a previously validated collection of over 1,400 groups of ICD-10 codes (example shown in Table 2) [41]. Inference was completed in 15 days.

2.6.1 Metrics

For evaluation, each outcome was treated as a binary classification task: given a patient's history, estimate the probability of event i occurring within t days, and compare against the observed outcome. This allowed for the evaluation of single medical event prediction as well as bespoke phenotypes composed of multiple events or more complex definitions.

Discriminative performance was assessed using the area under the receiver operating characteristic curve (AUROC) [42] and the area under the precision–recall curve (AUPRC) [43], the latter being more commonly used under class imbalance [44, 45]. Overall accuracy and calibration were jointly measured using the Brier score [46]. We also generated ROC, precision–recall, and calibration curves for visual inspection.

Confidence intervals (95%) were obtained via non-parametric patient-level bootstrapping with the percentile method, using 1,000 resamples for all analyses apart from the Phecode tasks (where 100 were used) due to computational cost. To reflect potential use as a clinical screening tool, we also evaluated recall at a fixed 10% false-positive rate (FPR10), consistent with prior studies [47] and representing a hypothetical operational threshold.

2.6.2 Subgroup Analyses

We conducted subgroup analyses of mortality predictions at both the 30-day post–SARS-CoV-2–positive test and one-year horizons, stratifying test-set patients independently by age, sex, ethnicity, and vaccination status. For each subgroup, we calculated the AUROC and AUPRC in comparison with the overall cohort. Additionally, we investigated how performance on these metrics varied with the number of historical patient events provided as model input. Finally, we assessed changes in AUROC and AUPRC relative to the timing of the COVID-19 test to determine whether the Foresight model could capture evolving patterns in pandemic dynamics.

2.7 Baseline Methods

We benchmarked Foresight-E against two supervised classifiers trained separately for all-cause mortality and hospitalisation at 30 days after a positive SARS-CoV-2 test, and for the same outcomes over a one-year horizon in 2023.

All baselines used the same training split as Foresight-E but were task-specific, in contrast to Foresight-E's zero-shot capability. The first was a logistic regression model [48] using age, sex, and ethnicity (one-hot encoded), with vaccination status added for the SARS-CoV-2 tasks. The second was an XGBoost model [49] using one-hot vectors of all medical codes from the Foresight-E vocabulary, providing equivalent structured clinical data but without temporal ordering.

To allow training outcomes to be observed, input cutoffs were set to 1 December 2022 for the 30-day tasks and 1 January 2022 for the one-year tasks. This requirement underscores a key advantage of Foresight-E's self-supervised learning framework: it does not rely on outcome labels during training, eliminating the need to reserve training data for labelling temporally. Performance for all models was evaluated using AUROC, AUPRC, and Brier score, as described in Section 2.6.1.

3 Discussion

Following concerns raised by the British Medical Association and Royal College of General Practitioners' Joint GP IT Committee, NHS England has paused access to data for the Foresight project while a review is carried out [19]. That pause means quantitative results are not available pending the outcome of ongoing discussions. Therefore, we discuss the main methodological strengths and limitations of Foresight-E, including data sources, modelling approach, inference strategy, evaluation design, and potential applications.

3.1 Data

Foresight-E's key strength is its first use of national-scale, routinely collected EHRs spanning primary care, secondary care, COVID-19, and death registrations. This ensures Foresight is trained on diverse data representative of the general population [15], helping to mitigate algorithmic bias [50], enabling

prediction of rare diseases [51, 52], and facilitating onward translation through aligning training data with intended use populations. Given that primary care accounts for most healthcare delivery, integrating this data provides a more complete representation of an individual's health, enables earlier risk stratification, and may ultimately guide preventive interventions to avert disease onset, complications and costly secondary care.

Despite its breadth, the datasets reflect typical challenges of routinely collected health data: incomplete or imprecise coding [53], historical biases in care [54, 50], and shifts in recording practices [55], especially during the COVID-19 pandemic [56]. Notably, the GDPPR primary care dataset contains only a subset of codes, those deemed relevant for COVID-related research, omitting many common and rare diseases, as well as signs and symptoms data crucial for understanding disease presentation and evolution. Furthermore, it only contains those alive on or after 1 November 2019. This restricts the available joint history, which may limit the model's ability to capture long-term trajectories. Expanding the temporal window would be expected to improve performance but require methods that efficiently handle longer sequences [57].

Future enhancements could come from incorporating additional modalities such as clinical notes or medical imaging. The development of foundational multi-modal transformer models in the general domain shows how the methods presented here could be extended [58, 59]. However, such data is not currently available at a national scale, and provisioning would require appropriate governance, infrastructure, and technical methods, including de-identification and pre-processing.

3.2 Model

Foresight-E uses a 243-million-parameter Llama-2–style transformer decoder with Rotary Positional Embeddings (ROPE) and FlashAttention-2 for efficiency. While larger than most prior EHR models [60] and trained on national scale data, the scale of data, model size, and compute is still orders of magnitude smaller than the general domain [61].

Foresight-E is designed to be data-driven, rather than relying on predefined parametric models such as exponential hazard–style structures [60, 62]. Although we trained from scratch due to a custom vocabulary and NHSE SDE import restrictions, smaller-scale studies indicate that fine-tuning pretrained LLMs for medical-event prediction is beneficial to performance [63]. If future NHSE SDE policy permits importing pretrained models, initialising from a general model, and adapting in-domain is a promising direction.

Our tokenisation strategy prioritised clinical and representational fidelity. We retained the recorded granularity of clinical codes, all event codes, and encoded exact day-level time gaps, rather than collapsing codes into broader categories, filtering low-occurrence events, or binning time [9]. This maximised diagnostic specificity and allowed prediction of rare events often excluded in other models [52]. However, this enlarged the vocabulary and prediction space and reduced per-token training frequency. Furthermore, there were no out-of-vocabulary codes during training, so when encountered during inference, unseen codes had to be dropped, potentially losing patient information. Future extensions of the tokeniser could leverage hierarchical relations between concepts, as defined by ontologies, building on evidence that such structure can enhance EHR model performance [64].

We jointly encoded absolute time and patient age, enabling the model to capture temporal shifts in care patterns; however, this required careful sequence validity constraints (e.g., preserving demographic and year-start tokens). Sequence truncation was done at the event level during training as a pre-processing step to avoid invalid timelines. However, this is not possible during GPU inference, so a fixed maximum input length was used, meaning some long histories were partially discarded, thereby losing patient information.

Batching was performed per patient, rather than randomly sampling from concatenated EHRs [60, 9]. This prevented spurious cross-patient attention, at the cost of more padding and lower GPU utilisation. Future work could explore intra-document causal masking [65] to enable the concatenation of patients' timelines without information leakage, thereby reducing padding and improving training efficiency.

3.3 Inference

We extended prior probabilistic patient-trajectory approaches [9], applying Foresight-E in a zero-shot setting to forecast $\sim$40,000 medical events, across 30-day and 1-year time horizons relevant to

COVID-19 outcomes. This offers the broad ability to predict differential diagnoses and clinical trajectories, rather than single outcomes, but is computationally intensive, particularly at a population scale. Comparative studies with task-specific fine-tuning are needed to clarify efficiency–flexibility trade-offs.

Our current setup generates 48 trajectories per patient, constrained practically by the NVIDIA A10 GPU's memory. Exploring convergence dynamics and alternative decoding strategies (beam search, top-k, temperature) may yield gains in efficiency, accuracy, and calibration, but would require tuning. Furthermore, quantifying the temporal horizon over which Foresight-E can reliably forecast patient outcomes, both in absolute time and in terms of tokens, requires further study.

Finally, enhancing interpretability through techniques such as attention-weighted visualisation, gradient-based saliency, or counterfactual generation could help clinicians understand the model's forecasts by scrutinising learned associations for known and novel patterns, as well as the presence of spurious correlations, such as shortcut learning. Such methods could support safe adoption in practice by explaining why the model anticipates particular outcomes, as well as identifying potentially modifiable risk factors as targets for intervention to optimise health.

3.4 Evaluation

We evaluated Foresight-E's ability to predict COVID-19 outcomes in two settings designed to reflect real-world deployment challenges.

First, we assessed its ability to predict direct COVID-19 outcomes during the pandemic, a period marked by shifting conditions such as emerging viral variants, changing testing protocols, public health interventions and population immunity. Unlike earlier models [9], Foresight-E explicitly encodes both absolute time and patient age, enabling it to adapt to these temporal shifts. Forecasting future novel threats, outside the current training data and vocabulary, could be supported by continually retraining on the latest batches of routinely collected data.

Second, we sought to predict the indirect effects of COVID-19 and simulate a prospective deployment by predicting events in 2023, one year beyond the training period, on over 1,400 Phecodes, emergency hospitalisation, and all-cause mortality. COVID-19 highlights the challenges of temporal data shifts. Foresight-E was trained during the height of the COVID-19 pandemic, a period of profound healthcare system disruption and excess all-cause mortality, which left an enduring and evolving legacy in the evaluation period of 2023, encompassing the indirect effects the pandemic exerted on individuals, healthcare systems, and society at large [40].

This represents, to our knowledge, the broadest zero-shot evaluation of an EHR foundation model to date. Working at the population level meant facing low prevalence for many acute outcomes, in contrast to models trained solely on high-acuity inpatient cohorts (e.g., MIMIC-IV [66]). While this sparsity poses challenges, it also demonstrates Foresight-E's potential utility for population screening as well as high-risk patient monitoring, by training on both healthy and acutely ill patient timelines. In order to critically evaluate Foresight-E's performance in population screening we additionally reported recall at a fixed 10% false-positive rate, consistent with prior studies [47] and representing a hypothetical operational threshold.

Due to NHSE SDE constraints, external pretrained models could not be imported for direct comparison, and compute limits restricted the number of baseline models trained. Moreover, framing each patient trajectory as thousands of binary prediction tasks provided established metrics (AUROC, AUPRC, Brier score) but did not capture the overall fidelity of generated trajectories, an important direction for future work.

A key limitation of this work is the inability to benchmark against commonly used clinical risk prediction tools, due to a lack of required data (e.g. physiological measurements and test results [67]), limited data duration (e.g. QRisk measures 10-year cardiovascular disease risk [1]) and the fact that where risk scores were recorded the resulting outcomes were then conditioned on resultant clinical decision making [68].

3.5 Future Directions and Considerations

Foresight-E was developed as a research pilot strictly for COVID-19-related research and is not a validated clinical tool. The model is therefore confined to this scope, and any future directions for the methodology are entirely contingent on navigating the significant governance challenges outlined below.

The generative, zero-shot forecasting of models like Foresight-E offers broad potential, including forecasting population health demands, stratifying groups at increased risk of adverse outcomes, and enabling personalised risk prediction to guide preventive interventions. Beyond direct clinical care, potential applications include improving clinical trial efficiency through prognostic enrichment or advancing drug discovery by better modelling disease trajectories. A priority area for research is validating the model's capacity for counterfactual generation, a crucial step toward creating robust digital twins and enabling trustworthy in-silico trials.

However, moving from research to application would require secure, real-time model deployment, a capability beyond current NHSE SDE infrastructure. This gap is particularly critical for large generative models, which can inadvertently memorise and expose sensitive training data [69]. A potential mitigation strategy is to deploy such models within a secure environment behind narrowly scoped APIs. These would provide only predefined, validated outcomes, such as calibrated risk scores, rather than open-ended generative trajectories, thereby constraining vectors for data extraction and mitigating privacy risks [70].

However, these ambitions are secondary to the fundamental governance hurdles. The most significant barrier is that any extension of this work is contingent on securing new data access approvals under a transparent framework with a clear public benefit beyond COVID-19 research. This would require establishing a new social license through deep and sustained engagement with patients, the public, and professional bodies. Furthermore, the path from a research model to a trustworthy clinical tool would necessitate additional rigorous evaluation and a clear route to regulatory approval as a medical device. Addressing these socio-technical challenges is the central prerequisite for future progress in this domain.

4 Conclusion

We have presented Foresight-E, a 243-million-parameter transformer trained on national-scale EHR data from 54.9 million NHS patients and evaluated on its ability to perform zero-shot prediction across $\sim$1.4k medical outcomes for a 6.1-million-patient test set. Within the constraints of COVID-19 governance, we outlined the data integration pipeline, tokenisation strategy, model architecture, training procedure, and inference and evaluation framework.

Although quantitative results are currently withheld pending ongoing discussions, this work demonstrates that it is technically feasible to develop a foundation model for healthcare entirely within existing NHS infrastructure. By combining routinely collected population-scale de-identified EHRs with modern generative modelling, Foresight-E offers a blueprint for zero-shot healthcare AI systems.

Expanding access to broader clinical modalities, extending beyond COVID-restricted datasets, and developing safe deployment pathways could enable models like Foresight-E to support both population-level planning and individualised care. Realising this potential will require not only technical advances but also transparent governance, sustained public and professional engagement, and rigorous evaluation in real-world clinical settings to generate evidence for regulatory approval.

5 Ethics approval

The North East - Newcastle and North Tyneside 2 research ethics committee provided ethical approval for the CVD-COVID-UK/COVID-IMPACT research program (REC No 20/NE/0161) to access, within secure trusted research environments, unconsented, whole-population, de-identified data from EHRs collected as part of patients' routine healthcare.

Patient and public involvement was included in the approvals process and has continued to shape the research through Patient and Public Involvement and Engagement sessions organised via BHF Data Science Centre [71].

References

[1] Julia Hippisley-Cox, Carol Coupland, and Peter Brindle. Development and validation of qrisk3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *bmj*, 357, 2017.

[2] Gregory YH Lip, Robby Nieuwlaat, Ron Pisters, Deirdre A Lane, and Harry JGM Crijns. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*, 137 (2):263–272, 2010.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] 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):7155, 2020.

[7] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

[8] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Baston, Jack Ross, Esther Idowu, et al. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024.

[9] Pawel Renc, Yugang Jia, Anthony E Samir, Jaroslaw Was, Quanzheng Li, David W Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *NPJ Digital Medicine*, 7(1):256, 2024.

[10] Atiquer Rahman Sarkar, Yao-Shun Chuang, Noman Mohammed, and Xiaoqian Jiang. De-identification is not enough: a comparison between de-identified and synthetic clinical notes. *Scientific reports*, 14(1):29669, 2024.

[11] Sheng Zhang, Qin Liu, Naoto Usuyama, Cliff Wong, Tristan Naumann, and Hoifung Poon. Exploring scaling laws for ehr foundation models. *arXiv preprint arXiv:2505.22964*, 2025.

[12] Joseph E Alderman, Joanne Palmer, Elinor Laws, Melissa D McCradden, Johan Ordish, Marzyeh Ghassemi, Stephen R Pfohl, Negar Rostamzadeh, Heather Cole-Lewis, Ben Glocker, et al. Tackling algorithmic bias and promoting transparency in health datasets: the standing together consensus recommendations. *The Lancet Digital Health*, 7(1):e64–e88, 2025.

[13] NHS Digital. Control of patient information (COPI) notice. `https://digital.nhs.uk/coronavirus/coronavirus-covid-19-response-information-governance-hub/control-of-patient-information-copi-notice`, 2022. Accessed: 2025-07-14.

[14] Angela Wood, Rachel Denholm, Sam Hollings, Jennifer Cooper, Samantha Ip, Venexia Walker, Spiros Denaxas, Ashley Akbari, Amitava Banerjee, William Whiteley, et al. Linked electronic health records for research on a nationwide cohort of more than 54 million people in england: data resource. *bmj*, 373, 2021.

[15] Marta Pineda-Moncusí, Freya Allery, Antonella Delmestri, Thomas Bolton, John Nolan, Johan H Thygesen, Alex Handy, Amitava Banerjee, Spiros Denaxas, Christopher Tomlinson, et al. Ethnicity data resource in population-wide health records: completeness, coverage and granularity of diversity. *Scientific Data*, 11(1):221, 2024.

[16] NHS England. Secure Data Environment. https://digital.nhs.uk/services/secure-data-environment-service, 2025. Accessed: 2025-08-13.

[17] Alex Handy, Angela Wood, Cathie Sudlow, Christopher Tomlinson, Frank Kee, Johan H Thygesen, Mohammad Mamouei, Reecha Sofat, Richard Dobson, Samantha Ip, et al. A nationwide deep learning pipeline to predict stroke and covid-19 death in atrial fibrillation. *Medrxiv*, pages 2021–12, 2021.

[18] Freya Allery, Marta Pineda-Moncusí, Christopher Tomlinson, Nikolas Pontikos, Johan H Thygesen, Sara Khalid, and CVD-COVID-UK/COVID-IMPACT Consortium. Towards mitigating health inequity via machine learning: a nationwide cohort study to develop and validate ethnicity-specific models for prediction of cardiovascular disease risk in covid-19 patients. *medRxiv*, pages 2023–09, 2023.

[19] Stephen Armstrong. Nhs england faces investigation over granting foresight access to gp patient data, 2025.

[20] Gary S Collins, Karel GM Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B Reitsma, Maarten Van Smeden, et al. Tripod+ ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*, 385, 2024.

[21] Karel GM Moons, Johanna AA Damen, Tabea Kaul, Lotty Hooft, Constanza Andaur Navarro, Paula Dhiman, Andrew L Beam, Ben Van Calster, Leo Anthony Celi, Spiros Denaxas, et al. Probast+ ai: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *bmj*, 388, 2025.

[22] Johan H Thygesen, Christopher Tomlinson, Sam Hollings, Mehrdad A Mizani, Alex Handy, Ashley Akbari, Amitava Banerjee, Jennifer Cooper, Alvina G Lai, Kezhi Li, et al. Covid-19 trajectories among 57 million adults in england: a cohort study using electronic health records. *The Lancet Digital Health*, 4(7):e542–e557, 2022.

[23] NHS Digital. COVID-19 General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR). https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services/data-set-catalogue/gpes-data-for-pandemic-planning-and-research-gdppr, 2025. Accessed: 2025-08-13.

[24] NHS Digital. Hospital Episode Statistics (HES). https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics, 2025. Accessed: 2025-08-13.

[25] Health Data Research Gateway. Civil Registration - Deaths. https://healthdatagateway.org/en/dataset/877, 2024. Accessed: 2025-08-13.

[26] NHS Digital. COVID-19 Second Generation Surveillance System. https://digital.nhs.uk/services/data-services-for-commissioners/datasets/covid-19-second-generation-surveillance-system, 2023. Accessed: 2025-01-10.

[27] NHS Digital. COVID-19 Vaccination Status. https://digital.nhs.uk/services/data-services-for-commissioners/datasets/covid-19-vaccination-status, 2023. Accessed: 2025-01-10.

[28] World Health Organization. International Statistical Classification of Diseases and Related Health Problems (ICD). https://www.who.int/standards/classifications/classification-of-diseases, 2025. Accessed: 2025-04-14.

[29] NHS Digital. Clinical Classifications. https://digital.nhs.uk/services/terminology-and-classifications/clinical-classifications, 2025. Accessed: 2025-04-14.

[30] NHS Digital. SNOMED CT. https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct, 2025. Accessed: 2025-04-14.

[31] AWS. Amazon EC2 G5 Instances. https://aws.amazon.com/ec2/instance-types/g5/, 2025. Accessed: 2025-04-14.

[32] nvidia. NVIDIA A10 Tensor Core GPU. https://www.nvidia.com/en-gb/data-center/products/a10-gpu/, 2025. Accessed: 2025-04-14.

[33] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[34] Hugging Face. Llama 2. https://huggingface.co/docs/transformers/en/model_doc/llama2, 2023. Accessed: 2025-04-28.

[35] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[36] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

[37] Andrej Karpathy. nanoGPT. https://github.com/karpathy/nanoGPT, 2025. Accessed: 2025-04-15.

[38] Richard D Riley, Gary S Collins, Laura Kirton, Kym IE Snell, Joie Ensor, Rebecca Whittle, Paula Dhiman, Maarten van Smeden, Xiaoxuan Liu, Joseph Alderman, et al. Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches. *bmj*, 388, 2025.

[39] Amitava Banerjee, Laura Pasea, Steve Harris, Arturo Gonzalez-Izquierdo, Ana Torralbo, Laura Shallcross, Mahdad Noursadeghi, Deenan Pillay, Neil Sebire, Chris Holmes, et al. Estimating excess 1-year mortality associated with the covid-19 pandemic according to underlying conditions and age: a population-based cohort study. *The lancet*, 395(10238):1715–1725, 2020.

[40] Simon Ball, Amitava Banerjee, Colin Berry, Jonathan R Boyle, Benjamin Bray, William Bradlow, Afzal Chaudhry, Rikki Crawley, John Danesh, Alastair Denniston, et al. Monitoring indirect impact of covid-19 pandemic on services for cardiovascular diseases in the uk. *Heart*, 106(24):1890–1897, 2020.

[41] Patrick Wu, Aliya Gifford, Xiangrui Meng, Xue Li, Harry Campbell, Tim Varley, Juan Zhao, Robert Carroll, Lisa Bastarache, Joshua C Denny, et al. Mapping icd-10 and icd-10-cm codes to phecodes: workflow development and initial evaluation. *JMIR medical informatics*, 7(4): e14325, 2019.

[42] Mark RJ Junge and Joseph R Dettori. Roc solid: Receiver operator characteristic (roc) curves as a foundation for better diagnostic tests. *Global spine journal*, 8(4):424–429, 2018.

[43] Mu Zhu. Recall, precision and average precision. *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, 2(30):6, 2004.

[44] Brice Ozenne, Fabien Subtil, and Delphine Maucort-Boulch. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of clinical epidemiology*, 68(8):855–859, 2015.

[45] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.

[46] W Brier Glenn et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[47] Julia Carrasco-Zanini, Maik Pietzner, Jonathan Davitte, Praveen Surendran, Damien C Croteau-Chonka, Chloe Robins, Ana Torralbo, Christopher Tomlinson, Florian Grünschläger, Natalie Fitzpatrick, et al. Proteomic signatures improve risk prediction for common and rare diseases. *Nature medicine*, 30(9):2489–2498, 2024.

[48] scikit learn. LogisticRegression. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html, 2025. Accessed: 2025-08-13.

[49] DMLC XGBoost. XGBoost Documentation. https://xgboost.readthedocs.io/en/stable/index.html, 2022. Accessed: 2025-07-14.

[50] Anmol Arora, Joseph E Alderman, Joanne Palmer, Shaswath Ganapathi, Elinor Laws, Melissa D Mccradden, Lauren Oakden-Rayner, Stephen R Pfohl, Marzyeh Ghassemi, Francis Mckay, et al. The value of standards for health datasets in artificial intelligence-based applications. *Nature medicine*, 29(11):2929–2938, 2023.

[51] Johan H Thygesen, Huayu Zhang, Hanane Issa, Jinge Wu, Tuankasfee Hama, Ana-Caterina Phiho-Gomes, Tudor Groza, Sara Khalid, Thomas R Lumbers, Mevhibe Hocaoglu, et al. Prevalence and demographics of 331 rare diseases and associated covid-19-related mortality among 58 million individuals: a nationwide retrospective observational study. *The Lancet Digital Health*, 7(2):e145–e156, 2025.

[52] The Lancet Rheumatology. Translating ai innovation into clinical practice. *The Lancet Rheumatology*, 7(7):e451, 2025. ISSN 2665-9913. doi: 10.1016/S2665-9913(25)00161-4. URL https://doi.org/10.1016/S2665-9913(25)00161-4. Published July 1, 2025.

[53] Olga Kostopoulou, Christopher Tracey, and Brendan C Delaney. Can decision support combat incompleteness and bias in routine primary care data? *Journal of the American Medical Informatics Association*, 28(7):1461–1467, 2021.

[54] Michael W Sjoding, Robert P Dickson, Theodore J Iwashyna, Steven E Gay, and Thomas S Valley. Racial bias in pulse oximetry measurement. *New England Journal of Medicine*, 383 (25):2477–2478, 2020.

[55] Salwa S Zghebi, David Reeves, Christos Grigoroglou, Brian McMillan, Darren M Ashcroft, Rosa Parisi, and Evangelos Kontopantelis. Clinical code usage in uk general practice: a cohort study exploring 18 conditions over 14 years. *BMJ open*, 12(7):e051456, 2022.

[56] Marta Pineda-Moncusí, Freya Allery, Hoda Abbasizanjani, David Powell, Albert Prats-Uribe, Johan H Thygesen, Angela Wood, Christopher Tomlinson, Amitava Banerjee, Ashley Akbari, et al. Ethnic disparities in covid-19 mortality and cardiovascular disease in england and wales between 2020-2022. *Nature Communications*, 16(1):6059, 2025.

[57] Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Ré, Sanmi Koyejo, and Nigam H Shah. Context clues: Evaluating long context models for clinical prediction tasks on ehrs. *arXiv preprint arXiv:2412.16178*, 2024.

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[59] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[60] Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *medRxiv*, pages 2024–06, 2024.

[61] Meta. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025. Accessed: 2025-07-14.

[62] Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. Motor: a time-to-event foundation model for structured medical records. *arXiv preprint arXiv:2301.03150*, 2023.

[63] Zeljko Kraljevic, Joshua Au Yeung, Daniel Bean, James Teo, and Richard J Dobson. Large language models for medical forecasting–foresight 2. *arXiv preprint arXiv:2412.10848*, 2024.

[64] Xiaorui Su, Shvat Messica, Yepeng Huang, Ruth Johnson, Lukas Fesser, Shanghua Gao, Faryad Sahneh, and Marinka Zitnik. Multimodal medical code tokenizer. *arXiv preprint arXiv:2502.04397*, 2025.

[65] Yu Zhao, Yuanbin Qu, Konrad Staniszewski, Szymon Tworkowski, Wei Liu, Piotr Miłoś, Yuxiang Wu, and Pasquale Minervini. Analysing the impact of sequence composition on language model pre-training. *arXiv preprint arXiv:2402.13991*, 2024.

[66] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[67] Stephen R Knight, Antonia Ho, Riinu Pius, Iain Buchan, Gail Carson, Thomas M Drake, Jake Dunning, Cameron J Fairfield, Carrol Gamble, Christopher A Green, et al. Risk stratification of patients admitted to hospital with covid-19 using the isaric who clinical characterisation protocol: development and validation of the 4c mortality score. *bmj*, 370, 2020.

[68] Hugh Logan Ellis, Edward Palmer, James T Teo, Martin Whyte, Kenneth Rockwood, and Zina Ibrahim. The early warning paradox. *npj Digital Medicine*, 8(1):81, 2025.

[69] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[70] Emily Jefferson, James Liley, Maeve Malone, Smarti Reel, Alba Crespi-Boixader, Xaroula Kerasidou, Francesco Tava, Andrew McCarthy, Richard Preen, Alberto Blanco-Justicia, et al. Graimatter green paper: Recommendations for disclosure control of trained machine learning (ml) models from trusted research environments (tres). *arXiv preprint arXiv:2211.01656*, 2022.

[71] British Heart Foundation. CVD-COVID-UK / COVID-IMPACT. `https://bhfdatasciencecentre.org/areas/cvd-covid-uk-covid-impact/`, 2025. Accessed: 2025-01-10.

[72] British Heart Foundation. Home- British Heart Foundation. `https://bhfdatasciencecentre.org/`, 2025. Accessed: 2025-07-20.

[73] NHS Digital. Advisory Group for Data (AGD). `https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/advisory-group-for-data`, 2025. Accessed: 2025-07-20.

[74] NHS Digital. Data Access Request Service (DARS) products and services. `https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services`, 2024. Accessed: 2025-07-20.

[75] Databricks. Databricks Runtime 14.3 LTS. `https://docs.databricks.com/aws/en/release-notes/runtime/14.3lts`, 2024. Accessed: 2025-04-29.

A Data availability

The data used in this study are available in the NHSE SDE service for England, but as restrictions apply, they are not publicly available [16].

The CVD-COVID-UK/COVID-IMPACT programme, led by the BHF Data Science Centre [72], received approval to access data in the NHSE SDE service for England from the Independent Group Advising on the Release of Data (IGARD) [73] via an application made in the Data Access Request Service (DARS) Online system (ref. DARS-NIC-381078-Y9C5K) [74].

The CVD-COVID-UK/COVID-IMPACT Approvals & Oversight Board [71] subsequently granted approval to this project to access the data within the NHSE SDE service for England. The de-identified data used in this study were made available to accredited researchers only. Those wishing to gain access to the data should contact bhfdsc@hdruk.ac.uk in the first instance.

B Code availability

All data preparation, model training, and evaluation code will be released on GitHub (pending regulatory review and NHSE SDE output review) at: `https://github.com/BHFDSC/CCU078_01_Foresight-SDE`. The repository will include a `requirements.txt` specifying all package versions.

Due to data restrictions, trained model weights and artefacts are only accessible to a subset of approved consortium researchers on a dedicated Foresight cluster within the NHSE SDE. Those wishing to gain access to the data should contact bhfdsc@hdruk.ac.uk in the first instance.

All analyses were executed within the NHSE SDE [16] using Databricks Runtime 14.3 LTS for ML [75]; training/evaluation used an AWS g5.48xlarge instance with eight NVIDIA A10 GPUs [31, 32].

C Datasets

Table 1: Summary of datasets used in the study, including 4 separate secondary care datasets.

Domain	Datasets	Citation
Primary care	General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR)	[23]
Secondary care	Hospital Episode Statistics (HES): Outpatients (OP), Accident & Emergency (A&E), Admitted Patient Care (APC) and Critical Care (CC)	[24]
Mortality	Office for National Statistics (ONS) Civil Registration of Deaths	[25]
COVID-19 testing	UK Health Security Agency (UK HSA), formerly Public Health England (PHE), COVID-19 Second Generation Surveillance System (SGSS)	[26]
COVID-19 vaccination	NHS England COVID-19 Vaccination Status	[27]

D Endpoint to Token Mapping

Table 2: Mapping of evaluation endpoints and associated figures to timeline tokens. An example phecode definition is shown.

Endpoint	Tokens
Mortality	DEATH
Hospitalisation	APC
Emergency Hospitalisation	APC_ADMIMETH_21, APC_ADMIMETH_22, APC_ADMIMETH_23, APC_ADMIMETH_24, APC_ADMIMETH_25, APC_ADMIMETH_28, APC_ADMIMETH_2A, APC_ADMIMETH_2B, APC_ADMIMETH_2C, APC_ADMIMETH_2D
Phecode 8.0 - Intestinal Infection	ICD10_A000, ICD10_A009, ICD10_A011, ICD10_A012, ICD10_A013, ICD10_A014, ICD10_A059, ICD10_A060, ICD10_A062, ICD10_A063, ICD10_A064, ICD10_A065, ICD10_A067, ICD10_A068, ICD10_A069, ICD10_A079

E TRIPOD+AI Checklist

Table 3: Assessment of Foresight against the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)+AI Checklist [20]. The results section is excluded as not included in this paper.

Item	Checklist item	Section
Title		
Title	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	Title
Abstract		
Title	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	Title
Background	Provide a brief explanation of the healthcare context and rationale for developing or evaluating the performance of all models	Abstract
Objectives	Specify the study objectives, including whether the study describes model development, evaluation, or both	Abstract
Methods	Describe the sources of data	Abstract
	Describe the eligibility criteria and setting where the data were collected	Abstract
	Specify the outcome to be predicted by the model, including time horizon of predictions in case of prognostic models	Abstract
	Specify the type of model, a summary of the model-building steps, and the method for internal validation	Abstract
	Specify the measures used to assess model performance (eg, discrimination, calibration, clinical utility)	Abstract
Results	Report the number of participants and outcome events	Pending
	Summarise the predictors in the final model	Pending
	Report model performance estimates (with confidence intervals)	Pending
Discussion	Give an overall interpretation of the main results	Pending
Registration	Give the registration number and name of the registry or repository	None
Introduction		
Background	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	1
	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (eg, healthcare professionals, patients, public)	1

	Describe any known health inequalities between sociodemographic groups	1
Objectives	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	1
Methods		
Data	Describe the sources of data separately for the development and evaluation datasets (eg, randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data	2.1
	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	2.1
Participants	Specify key elements of the study setting (eg, primary care, secondary care, general population) including the number and location of centres	2.1
	Describe the eligibility criteria for study participants	2.1
	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	2.2
Data preparation	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	2.1, 2.2
Outcome	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups	2.6
	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	N/A
	Report any actions to blind assessment of the outcome to be predicted	None
Predictors	Describe the choice of initial predictors (eg, literature, previous models, all available predictors) and any pre-selection of predictors before model building	2.1, 2.2
	Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors)	2.1, 2.3
	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	N/A
Sample size	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation	2.1
Missing data	Describe how missing data were handled. Provide reasons for omitting any data	2.3

Analytical methods	Describe how the data were used (eg, for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements	2.1
	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardisation)	2.1, 2.2, 2.3
	Specify the type of model, rationale†, all model-building steps, including any hyperparameter tuning, and method for internal validation	2.4
	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (eg, hospitals, countries). See TRIPOD-Cluster for additional considerations	None
	Specify all measures and plots used (and their rationale) to evaluate model performance (eg, discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	2.6.1, 2.7
	Specify all measures and plots used (and their rationale) to evaluate model performance (eg, discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	2.6.1
	Describe any model updating (eg, recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	None
	For model evaluation, describe how the model predictions were calculated (eg, formula, code, object, application programming interface)	2.5
Class imbalance	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	None
Fairness	Describe any approaches that were used to address model fairness and their rationale	2.1, 2.4.1
Model output	Specify the output of the prediction model (eg, probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified	2.5, 2.6.1
Training versus evaluation	Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors	2.1, 2.3,2.6.1
Ethical approval	Name the institutional research board or ethics committee that approved the study and describe the participant informed consent or the ethics committee waiver of informed consent	5
Open Science		
Funding	Give the source of funding and the role of the funders for the present study	Blinded for review
Conflicts of interest	Declare any conflicts of interest and financial disclosures for all authors	Blinded for review

Protocol	Indicate where the study protocol can be accessed or state that a protocol was not prepared	Currently not publicly released
Registration	Provide registration information for the study, including register name and registration number, or state that the study was not registered	Not registered
Data sharing	Provide details of the availability of the study data	A
Code sharing	Provide details of the availability of the analytical code	B
Patient and public involvement		
Patient and public involvement	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement	5
Discussion		
Interpretation	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies	Pending
Limitations	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalisability	3.1
Usability of the model in the context of current care	Describe how poor quality or unavailable input data (eg, predictor values) should be assessed and handled when implementing the prediction model	N/A
	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users	3.5
	Discuss any next steps for future research, with a specific view to applicability and generalisability of the model	3

F PROBAST+AI Assessment

Table 4: Assessment of Foresight-E against Risk Of Bias ASsessment Tool (PROBAST) + AI tool [21]

Item	Description	Section/ Comment
Step 1: PICOTS guidance		
Population	Define the target population (e.g., patients) in whom the assessed prediction models are to be applied. The target population not only directs search strings and in/exclusion criteria of prediction models or prediction model studies in case of a systematic literature review, but also directs the applicability assessment.	2.1
Index Model	Define the targeted prediction models to be assessed, which may be a single prediction model (the index model) of which the predictive accuracy is meta-analysed across multiple external evaluation studies of that index model but may also address multiple prediction models (developed or evaluated) for the targeted population, outcome or setting, depending on the assessor's or prediction model review focus.	2.4
Comparator model(s)	Define the other prediction models whose predictive ability is compared to that of the index model.	2.7
Outcome(s)	Define the outcomes or endpoints that are predicted by the index (and possibly comparator) prediction models in the target population.	2.6
Timing	Define the moment or time-point (e.g., in the patient work-up) at which the prediction with the prediction models is made (i.e., the start point or T0 of the use of the models).	2.6
	Define the time or follow-up period in which the outcomes are being predicted by the prediction models in the targeted population (prediction horizon).	2.6
Setting and intended use of the prediction model	Define the healthcare setting or context to which the index prediction models apply. The prediction ability of models may change across healthcare settings or contexts.	1
Step 2: Classify the type of prediction model assessment		
Development only	Prediction model development only, i.e., without evaluation of its performance.	✓
Evaluation only	External validation of one or more existing models in new data	
Combination	Prediction model development combined in the same study(publication) with the evaluation of its apparent performance, internal validation performance, or external validation performance.	
Step 3: Assess quality and applicability or risk of bias and applicability		
Participants and data sources	Describe the sources of data and criteria for participant selection	2.1
	Were appropriate data sources used?	Yes
	Was an appropriate study design used?	Yes
	Did the in- and exclusions of study participants result in a representative dataset?	Yes
	Concern regarding quality of selection of participants and data sources	Low
	Concern that the (data of the) included participants do not match the review question or the assessor's intended use of the prediction model	Low

Predictors	List and describe predictors included in the final prediction model, how they were defined and assessed, and their timing of assessment	2.1, 2.3
	Were predictors defined and assessed in a similar way for all participants?	Yes
	Was any pre-processing of predictors similar for all participants?	Yes
	Were predictor assessments made without knowledge of outcome data?	Yes
	Were the predictors included in the model available at the time the model was intended to be used?	Yes
	Concern regarding the quality of the predictors or their assessment	Low
	Concern that the definition, pre-processing, assessment, or timing of assessment of the predictors in the model do not match the review question or the assessor's intended use	Low
Outcome	Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination	2.6
	At what time point was the outcome determined? If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome?	2.6
	Were outcomes defined and assessed appropriately?	Yes
	Were outcomes defined and assessed in a similar way for all participants?	Yes
	Were outcome assessments made without use or knowledge of predictor data?	Yes
	Was the time interval between predictor assessment and outcome assessment appropriate?	Yes
	Concern regarding quality of the outcome or its determination	Low
	Concern that the outcome, its definition, assessment, or timing of assessment do not match the review question or the assessor's intended use	Low
Analysis	Describe the numbers of participants, number of candidate predictors, number of outcome events	2.1, 2.6
	Describe how the prediction model was developed (e.g., with respect to modelling technique, predictor selection, and classification or risk group definition)	2.4
	Describe the performance measures of the prediction model, e.g., (re)calibration, discrimination, (re)classification, net benefit, and whether they were adjusted for optimism	2.6.1
	Describe missing data on predictors and outcomes as well as methods used for handling these missing data	2.3
	Was there evidence that the sample size was reasonable?	Yes
	Were continuous and categorical predictors handled appropriately?	Yes
	Were participants with missing or censored data handled appropriately in the analysis?	N/A
	If methods to address class imbalance were used, was the model or the model predictions recalibrated?	N/A
	Were methods used to address potential model overfitting?	Yes
	Concern regarding quality of the analysis	Low
Step 4: Assess the overall concerns regarding quality, risk of bias and applicability of the prediction model		
Overall concern regarding quality of the prediction model development	Low concern regarding quality- If all four domains were rated low concern regarding quality.	✓

	High concern regarding quality- If at least one domain was rated high concern regarding quality. Unclear concern regarding quality- If at least one domain was rated unclear concern regarding quality and no domains were rated high concern.	
Overall concern regarding applicability of the prediction model development	Low concern for applicability- If all three domains were rated low concern for applicability.	✓
	High concern for applicability- If at least one domain was rated high concern for applicability. Unclear concern for applicability- If at least one domain was rated unclear concern for applicability and no domains were rated high concern.	