Learning the natural history of human disease with generative transformers

https://doi.org/10.1038/s41586-025-09529-3

Received: 18 May 2024

Accepted: 13 August 2025

Published online: 17 September 2025

Open access

Check for updates

Artem Shmatko[1,2,3,13], Alexander Wolfgang Jung[2,4,5,6,13], Kumar Gaurav[2,13], Søren Brunak[4,7], Laust Hvas Mortensen[5,7,8], Ewan Birney[2]✉, Tom Fitzgerald[2]✉ & Moritz Gerstung[1,2,9,10,11,12]✉

Decision-making in healthcare relies on understanding patients' past and current health states to predict and, ultimately, change their future course[1–3]. Artificial intelligence (AI) methods promise to aid this task by learning patterns of disease progression from large corpora of health records[4,5]. However, their potential has not been fully investigated at scale. Here we modify the GPT[6] (generative pretrained transformer) architecture to model the progression and competing nature of human diseases. We train this model, Delphi-2M, on data from 0.4 million UK Biobank participants and validate it using external data from 1.9 million Danish individuals with no change in parameters. Delphi-2M predicts the rates of more than 1,000 diseases, conditional on each individual's past disease history, with accuracy comparable to that of existing single-disease models. Delphi-2M's generative nature also enables sampling of synthetic future health trajectories, providing meaningful estimates of potential disease burden for up to 20 years, and enabling the training of AI models that have never seen actual data. Explainable AI methods[7] provide insights into Delphi-2M's predictions, revealing clusters of co-morbidities within and across disease chapters and their time-dependent consequences on future health, but also highlight biases learnt from training data. In summary, transformer-based models appear to be well suited for predictive and generative health-related tasks, are applicable to population-scale datasets and provide insights into temporal dependencies between disease events, potentially improving the understanding of personalized health risks and informing precision medicine approaches.

The progression of human disease across age is characterized by periods of health, episodes of acute illness and also chronic debilitation, often manifesting as clusters of co-morbidity. Patterns of multimorbidity affect individuals unevenly and have been associated with lifestyle, heritable traits and socioeconomic status[1–3]. Understanding each individual's multi-morbidity risks is important to tailor healthcare decisions, motivate lifestyle changes or direct entrance into screening programs, as is the case for cancer[8,9]. Critically, health cannot only be understood by the presentation of individual diagnoses but, rather, in the context of an individual's co-morbidities and their evolution over time. While a wide range of prediction algorithms exist for specific diseases, from cardiovascular disease to cancer[10–12], few algorithms are capable of predicting the full spectrum of human disease, which recognizes more than 1,000 diagnoses at the top level of the International Classification of Diseases, Tenth Revision (ICD-10) coding system.

Learning and predicting patterns of disease progression is also important in populations that are ageing and that exhibit shifts in their underlying demographic's morbidities. For example, it has been predicted that, globally, the number of cancer diagnoses will increase 77% by 2050 (ref. 13) or that, in the UK, the number of working-age individuals with major illnesses, including depression, asthma, diabetes, cardiovascular disease, cancer or dementia, will increase from 3 to 3.7 million by 2040 (ref. 14). Modelling the expected burden of disease is therefore critical for healthcare and economic planning and, moreover, the continual tracking of disease occurrence along with its likely future prevalence within population groups promotes a more informed healthcare system.

Recent developments in AI may help to address some methodological limitations of multi-morbidity modelling, which have so far proved difficult to overcome[15]. Aside from the great number of diagnoses, these include challenges in modelling temporal dependencies among previous events, the integration of potentially diverse prognostically relevant data and the statistical calibration of predictions. Large language models (LLMs)[16–19]—a subfield of AI that enables chatbots such as ChatGPT[20,21]—model language as a sequence of word fragments (tokens). Generated token by token, the new text is based on all preceding text

[1]Division of AI in Oncology, German Cancer Research Centre DKFZ, Heidelberg, Germany. [2]European Molecular Biology Laboratory, European Bioinformatics Institute EMBL-EBI, Hinxton, UK. [3]Faculty of Biosciences, Heidelberg University, Heidelberg, Germany. [4]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [5]Statistics Denmark, Copenhagen, Denmark. [6]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland. [7]Department of Public Health, University of Copenhagen, Copenhagen, Denmark. [8]ROCKWOOL Foundation, Copenhagen, Denmark. [9]Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany. [10]Robert Bosch Center for Tumor Diseases, Stuttgart, Germany. [11]Medical Faculty, Eberhard-Karls-University, Tübingen, Germany. [12]University Hospital Tübingen, Tübingen, Germany. [13]These authors contributed equally: Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav. ✉e-mail: birney@ebi.ac.uk; tomas@ebi.ac.uk; moritz.gerstung@dkfz.de

and, with enough training, the statistical dependencies among these tokens prove sufficient to produce context-aware and even conversational text, which is often indistinguishable from that of a human counterpart.

The analogy between LLMs and disease progression modelling, which also entails recognizing past events and exploiting their mutual dependencies to predict the future sequence of morbidity, has recently inspired a series of new AI models. For example, BERT-based models[22–25] have been developed for specific prediction tasks. Transformer models trained on electronic health records have been used for predicting diagnoses such as pancreatic cancer[26], self-harm[25] and stroke[24], as well as non-clinical parameters such as self-esteem[27]. However, despite promising proofs of concept[4,28,29], the potential for comprehensive and generative multi-morbidity modelling has not yet been fully assessed.

Here we demonstrate that attention-based transformer models, similar to LLMs, can be extended to learn lifetime health trajectories and accurately predict future disease rates for more than 1,000 diseases simultaneously on the basis of previous health diagnoses, lifestyle factors and further informative data. Our extended model, termed Delphi-2M, was trained on data from the UK Biobank, a population-scale research cohort, and validated on Danish population registries. The vocabulary of the model includes ICD-10 top-level diagnostic codes, as well as sex, body mass, smoking, alcohol consumption and death. Delphi provides individual-level predictions of multi-disease incidences and models future health trajectories at any point throughout an individual's life course. Moreover, the internal model of Delphi offers insights into how past data influence the rates of subsequent diseases. We further assess biases and fairness across demographic subgroups and discuss Delphi's potential as a framework for healthcare modelling.

A transformer model for health records

A person's health trajectory can be represented by a sequence of diagnoses using top-level ICD-10 codes recorded at the age of first diagnosis as well as death. Furthermore, 'no event' padding tokens were randomly added at an average rate of 1 per 5 years to eliminate long intervals without other inputs, which are especially frequent for younger ages and during which the baseline disease risk can change substantially (Extended Data Fig. 1). Together, these data comprise 1,258 distinct states—tokens in LLM terminology. Additional information includes sex, body mass index (BMI) and indicators of smoking and alcohol consumption, which are used as input information but not predicted by the model (Fig. 1a).

Training data comprised 402,799 (80%) participants of the UK Biobank recorded before the 1 July 2020. Data for the remaining 100,639 (20%) participants were used for validation and hyperparameter optimization, while all records for 471,057 (94%) participants still alive on 1 July 2020 were used for longitudinal testing up until 1 July 2022 (Fig. 1b). Additional external testing was conducted on the Danish disease registry data, which covered 1.93 million Danish nationals and spanned the period from 1978 to 2018.

To model disease history data, which, in contrast to text, occurs on a continuous time axis, we extended the GPT-2 architecture[6] (Fig. 1c). Transformer models map their inputs into an embedding space, where information is successively aggregated to enable autoregressive predictions. The first change therefore replaces GPT's positional encoding, a mapping that identifies each text token's discrete position, with an encoding of continuous age using sine and cosine basis functions[16]. Standard GPT models only predict the next token using a multinomial probability model. Thus, the second extension is the addition of another output head to also predict the time to the next token using an exponential waiting time model (Methods). Third, GPT's causal attention masks, which ensure that the model accesses only information from past events, are amended to additionally mask tokens recorded at the same time. Padding, lifestyle and sex tokens use a similar encoding but do not enter the likelihoods, as the model is deliberately not trained to predict them.

We term this model Delphi (Delphi large predictive health inference). This architecture enables one to provide the model with a partial health trajectory (prompt in LLM terminology) to calculate the subsequent rate (per day) for each of the 1,256 disease tokens plus death. Furthermore, the next token and the time to this event can be sampled on the basis of these rates. Iteratively, this procedure samples entire health trajectories (Fig. 1d).

A systematic screen of architecture hyperparameters (embedding dimensionality, number of layers, heads) confirms the reported empirical scaling laws[30], which state that model performance increases with the number of datapoints and, up to a limit defined by the available data, as the number of parameters increases (Fig. 1e). The screen indicates that, for the UK Biobank dataset, optimal Delphi models have around 2 million parameters. One of the models within the optimal range has an internal embedding dimensionality of 120, 12 layers and 12 heads, amounting to a total of 2.2 million parameters. Results based on this model parameterization are discussed throughout the rest of the paper. We note that qualitatively similar results are obtained from other parameter choices (Extended Data Fig. 2 and Supplementary Fig. 1).

An ablation analysis shows how Delphi-2M architectural modifications contribute to a better age- and sex-stratified cross-entropy compared with a standard GPT model (Fig. 1f, Supplementary Table 1 and Supplementary Fig. 2). A good, albeit slightly inferior, classification performance at different ages may already be achieved by adding regular 'no event' padding tokens to the input data with GPT models alone. However, a key distinguishing feature of Delphi compared with basic GPT models is its ability to calculate the absolute rates of tokens, which provide consistent estimates of inter-event times (Fig. 1g). This property also implies that the rates may be interpreted as the incidences of tokens.

Modelling multi-disease incidences

Delphi-2M's accuracy in predicting diverse disease outcomes in the validation cohort is compared to the sex and age-stratified incidence as an epidemiological baseline. As can be seen in the ten examples shown in Fig. 2a, the incidence curves are very varied, with some diseases, such as chickenpox, peaking in infancy, while others, such as asthma or depression, are relatively flat and with most rising exponentially in old age. Moreover, there are noticeable differences between the sexes, which are obvious for breast cancer but also pronounced for diabetes, depression, acute myocardial infarction and death. Delphi-2M's predictions are updated for each individual when new inputs are recorded. The predictions largely follow the sex- and age-stratified incidence curves but also indicate events or periods when the individual risk remains below or rises above the population average. For some diseases, such as asthma or arthrosis, the spread is narrow, indicating a limited ability to predict beyond the sex- and age-incidence trend. Yet for other diseases, including septicaemia, and also death, the spread is wide, indicating predictable inter-individual differences in disease rates.

Delphi's ability to predict the next diagnosis token across the spectrum of human disease is confirmed by the average age-stratified area under the receiver operating characteristic curve (AUC), which averages at values of approximately 0.76 in the internal validation data (Fig. 2b and Supplementary Table 2). For 97% of diagnoses, the AUC was greater than 0.5, indicating that the vast majority followed patterns with at least partial predictability. These patterns were found to be true across the different chapters of the ICD-10 spectrum, which define broad groups of disease for both sexes (Fig. 2c,d). Among the most confidently predicted next events is death, with an age-stratified AUC of 0.97 in both sexes. Importantly, calibration analyses in 5-year age brackets show that the predicted rates closely match the observed number of cases, showing that the models'

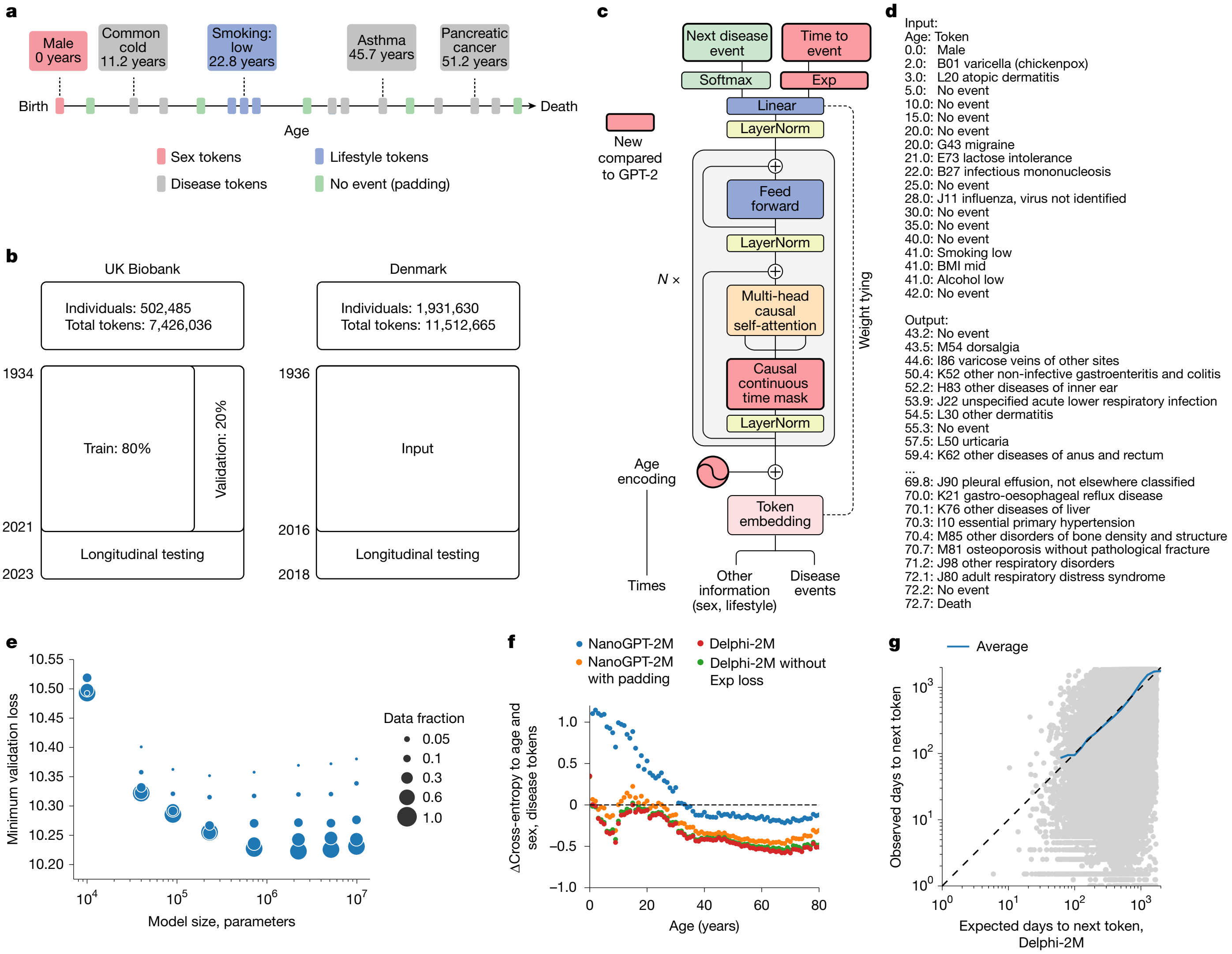

Fig. 1 | Delphi, a modified GPT architecture, models health trajectories. **a**, Schematic of health trajectories based on ICD-10 diagnoses, lifestyle and healthy padding tokens, each recorded at a distinct age. **b**, Training, validation and testing data derived from the UK Biobank (left) and Danish disease registries (right). **c**, The Delphi model architecture. The red elements indicate changes compared with the underlying GPT-2 model. '*N*×' denotes applying the transformer block sequentially *N* times. **d**, Example model input (prompt) and output (samples) comprising (age:token) pairs. **e**, Scaling laws of Delphi, showing the optimal validation loss as a function of model parameters for different training data sizes. **f**, Ablation results measured by the cross-entropy differences relative to an age- and sex-based baseline (*y* axis) for different ages (*x* axis). **g**, The accuracy of predicted time to event. The observed (*y* axis) and expected (*x* axis) time to events are shown for each next token prediction (grey dots). The blue line shows the average across consecutive bins of the *x* axis.

rates of the next tokens are consistently estimated (Extended Data Fig. 3).

Next-event predictions are often the consequence of acute illness or diagnostic refinements that accrue over the course of a few weeks or months, which may be undesirable for prognostication. Delphi-2M's average AUC values decrease from an average of 0.76 to 0.70 after 10 years, indicating that its predictions are also relevant for long-term prognostication (Fig. 2e and Supplementary Fig. 3). Similar results were observed in longitudinal test data, which also show no substantial shift in diagnostic patterns throughout the Biobank's follow-up (Supplementary Fig. 4).

The performance of Delphi was similar to routinely used clinical risk scores for cardiovascular disease and dementia, and better than those used for death. For diabetes, the performance of Delphi was worse compared with the use of a single marker, HbA1c, which is used clinically for risk prediction and diagnosis of diabetes (Fig. 2f, Supplementary Fig. 4c and Supplementary Table 3). This was the case for next-event predictions, as well as prediction horizons up to 24 months. Delphi-2M's AUC values were also generally higher than those of a recent machine learning algorithm that calculates the risks of a similarly broad spectrum of ICD-10 diagnoses using 67 different biomarkers available through the UK Biobank[31], even though for many diagnoses, such as diabetes, biomarkers remain indispensable (Fig. 2e and Extended Data Fig. 4), marking potential for future modifications of Delphi that additionally use data beyond health records (Extended Data Fig. 5). For most cases, Delphi-2M's multi-disease predictions match or exceed current risk models for individual disease outcomes and offer the great advantage of enabling the simultaneous assessment of more than 1,000 diseases and their timing at any given time, while also surpassing multi-disease models in quality.

Sampling future disease trajectories

One of the most promising features of generative models is the ability to sample disease trajectories, conditional on data recorded up to a