Learning the natural history of human disease with generative transformers

Artem Shmatko[1,2*], Alexander Wolfgang Jung[2,3*], Kumar Gaurav[2*], Søren Brunak[3], Laust Mortensen[4], Ewan Birney[2#], Tom Fitzgerald[2#] and Moritz Gerstung[1,2,5,6,7#]

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. Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
4. Statistics Denmark, Copenhagen, Denmark
5. Robert Bosch Center for Tumor Diseases, Stuttgart, Germany
6. Medical Faculty, Eberhard-Karls-University, Tübingen, Germany
7. University Hospital Tübingen, Tübingen, Germany

* These authors contributed equally
# Correspondence to: Ewan Birney, birney@ebi.ac.uk; Tom Fitzgerald, tomas@ebi.ac.uk; and Moritz Gerstung, moritz.gerstung@dkfz.de

Abstract

Decision-making in healthcare relies on the ability to understand patients' past and current health state to predict, and ultimately change, their future course. Artificial intelligence (AI) methods promise to aid this task by learning patterns of disease progression from large corpora of health records to predict detailed outcomes for an individual. However, the potential of AI has not yet been fully investigated at scale yet.

Here we modify the GPT (generative pretrained transformer) architecture to model the temporal progression and competing nature of human diseases in a population scale cohort. We train this model, termed Delphi-2M, on data from 0.4 million participants of the UK Biobank 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 different ICD-10 coded diseases and death, conditional on each individual's past disease history, age, sex and baseline lifestyle information, and with accuracy comparable to existing single-disease models. Delphi-2M's generative nature also enables sampling future health trajectories at any point within an individual's life course with outcomes across the entire disease spectrum. Sampled health trajectories provide meaningful estimates of future disease burden for up to 20 years and enable training AI models which have never seen actual data.

Explainable AI methods provide insights into Delphi-2M's predictions, revealing temporal clusters of co-morbidities within and across different disease chapters and their time-dependent consequences on the future health course. These analyses, however, also reveal that biases underlying the available training data, which in the case of the UK Biobank stem from distinct healthcare sources, are learned and highlighted.

In summary GPT-based models appear well suited for predictive and generative health related tasks, are applicable to population scale health data sets and provide insights into the temporal dependencies of past events that shape future health impacting our ability to obtain an instantaneous view of personalised health state.

Introduction

The progression of human disease across age is characterised by periods of health, episodes of acute illness and also chronic debilitation, often manifesting as clusters of co-morbidity. Patterns of multi-morbidity 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 enterance into screening programs as is the case for cancer[4,5]. 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[6–8], few algorithms are capable of predicting the full spectrum of human disease, which recognises more than 1,000 diagnoses at the top level of the ICD-10 (international classification of diseases, tenth revision) coding system.

Learning and predicting patterns of disease progression is also important in populations that are ageing and which 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[9] or that in the UK the number of working-age persons with major illnesses, including depression, asthma, diabetes, cardiovascular disease, cancer or dementia, will increase from 3 to 3.7 million by 2040[10]. Modelling the expected burden of disease is thus 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 health system.

Recent developments in artificial intelligence (AI) may help overcome some methodological limitations of multi-morbidity modeling which have so far proved difficult to overcome[11]. Aside from the great number of diagnoses, these include challenges in modeling temporal dependencies among previous events, the inclusion of potentially diverse prognostically relevant data and also the statistical calibration of predictions. Large language models (LLMs)[12–15], a subfield of AI that enables chatbots such as ChatGPT[16,17], model language as a sequence of word fragments (so-called tokens). Generated token by token, the new text is based on all preceding text - 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 large language and disease progression modeling – which also entails recognition of past events and exploitation of their mutual dependencies to predict the future sequence of morbidity – has recently inspired a series of new AI models. For example, BERT-based models[18–21] have been developed for specific prediction tasks. Transformer models trained on electronic health records have been used for predicting diseases such as pancreatic cancer[22], self-harm[21] and stroke[20], as well as non-clinical parameters such as self-esteem[23]. However, despite promising proof of concepts[24], the potential for comprehensive and generative modeling of multi-morbidity has not been fully assessed yet.

Here, we demonstrate that attention-based transformer models, similar to LLMs, can be extended to achieve the task of learning lifetime health trajectories and accurately predict

future disease rates of more than 1,000 diseases simultaneously based on prior health diagnosis, lifestyle factors and further informative data. Our extended model, termed Delphi-2M is trained on data from 400,000 participants of the UK Biobank, a population scale research cohort. The vocabulary of the model includes 1,256 ICD10 top level codes, as well as sex, body mass, smoking, alcohol consumption and death. The analyses not only reveal that Delphi-2M provides accurate instantaneous multistate rates but also highlights how elements of past data influence the subsequent rates of disease and enables sampling of future courses of health and disease at any point throughout an individual's life course. The predictions remain accurate even if evaluated on an external dataset from Denmark without any additional finetuning. Delphi can provide competitive individual level prediction of multiple outcomes simultaneously and can model the overall expected evolution of a population from a specific health history starting point. In addition, the internal model of Delphi is partly interpretable and can provide a framework for future research into healthcare trajectories.

Results

A transformer model for health trajectories

A person's health trajectory can be represented by a sequence of diagnoses using by top level ICD10 codes recorded at the age of first diagnosis as well as death. We choose to further include "no event" padding indicators, which were randomly added at an average rate of 1/5 years to ensure basic temporal continuity of inputs. Together, these data comprise 1,258 distinct states (or tokens in LLM terminology), each paired with the corresponding age, measured from birth, the sequence and timing of which are auto-regressively predicted. Further health-related information includes sex, BMI, as well as smoking and alcohol consumption indicators, which are added at the age when the corresponding information becomes available, but are not themselves part of predictions (**Figure 1a**).

Training data comprised 400,000 (80%) participants of the UK Biobank recorded before the 1st of July 2020. Test data sets were (i) the remaining 102,485 (20%) participants and (ii) data for 471,057 (94%) participants still alive on the 1st of July 2020 for longitudinal validation up until the 1st of July 2022 (**Figure 1b**). Additional external validation was carried out on a test set based on Danish disease registry data covering 1.93M Danish nationals aged 50-80 years old, with data available between 1978 and 2018.

In order to model disease history data, which, unlike text, occurs on a continuous time axis, we extended the GPT-2 architecture[25] (**Figure 1c**). Transformer models map their inputs into a lower dimensional embedding space, in which information is successively aggregated to facilitate auto-regressive predictions. The first change thus replaces GPT's positional encoding, a mapping which identifies the discrete position of each text token, with an encoding of continuous age using sine and cosine basis functions[12]. Standard GPT models only classify the next token using a multinomial probability model. Therefore, 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**). As the rates of competing exponentially

distributed waiting times define the logit probabilities of the next event, this parametrisation naturally extends GPT's existing classification step (**Figure 1g**). Third, GPT's so-called causal attention masks, which ensure that the model only accesses information from past events, are amended to resolve ties of tokens recorded at the same time (we note that such temporal ordering does not imply the model is capable of identifying causal mechanisms in the sense of modifiable features). 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 allows to provide the model with a partial health trajectory (prompt in LLM terminology) in order 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 based on these rates. Iteratively, this procedure samples entire health trajectories (**Figure 1d**).

A systematic screen of architecture hyperparameters (embedding dimensionality, number of layers, heads) confirms the reported empirical scaling laws[26] which state that model performance increases with the number of data points and, up to a limit defined by the available data, as the number of parameters increase (**Figure 1e; Extended Data Figure 1**). The screen indicates that for the UK Biobank dataset, optimal Delphi models have around 2M parameters. One of the models within the optimal range, contains 120 internal embedding dimensions, 12 layers and 12 heads amounting to a total of 2.2M parameters. Results based on this model parameterisation are discussed throughout the rest of the paper. We note that qualitatively similar results are obtained from other parameter choices.

An ablation analysis shows how Delphi-2M's architectural modifications contribute to a better age-and-sex-stratified cross-entropy compared to a standard GPT model (**Figure 1f**; **Table 1; Extended Data Figure 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 to basic GPT models is its ability to calculate the absolute rates of tokens, which in a statistically accurate sense represent inter-event times (**Figure 1g**). This property also implies that the rates may be interpreted as the incidences of tokens, as recorded in the training data.

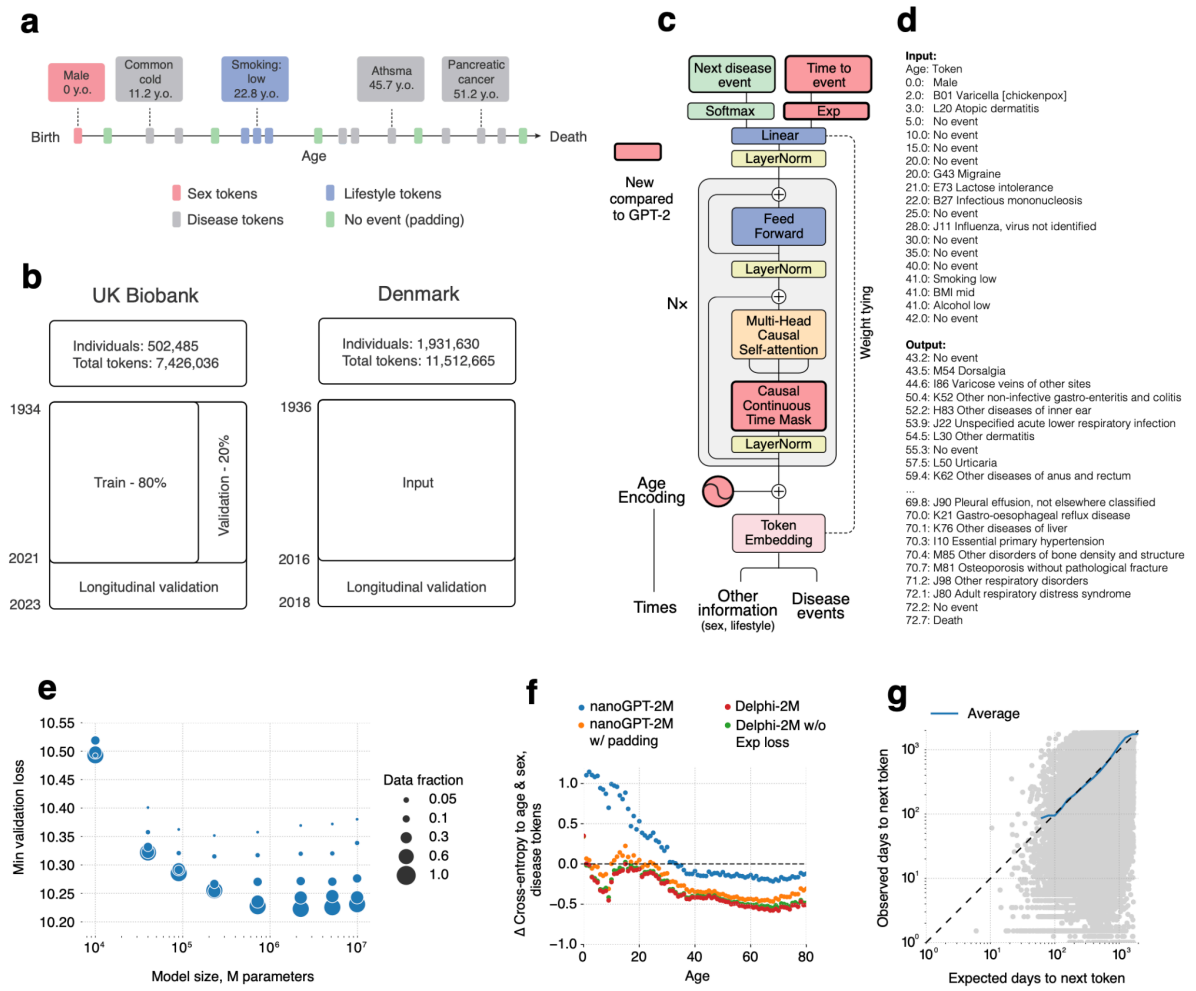
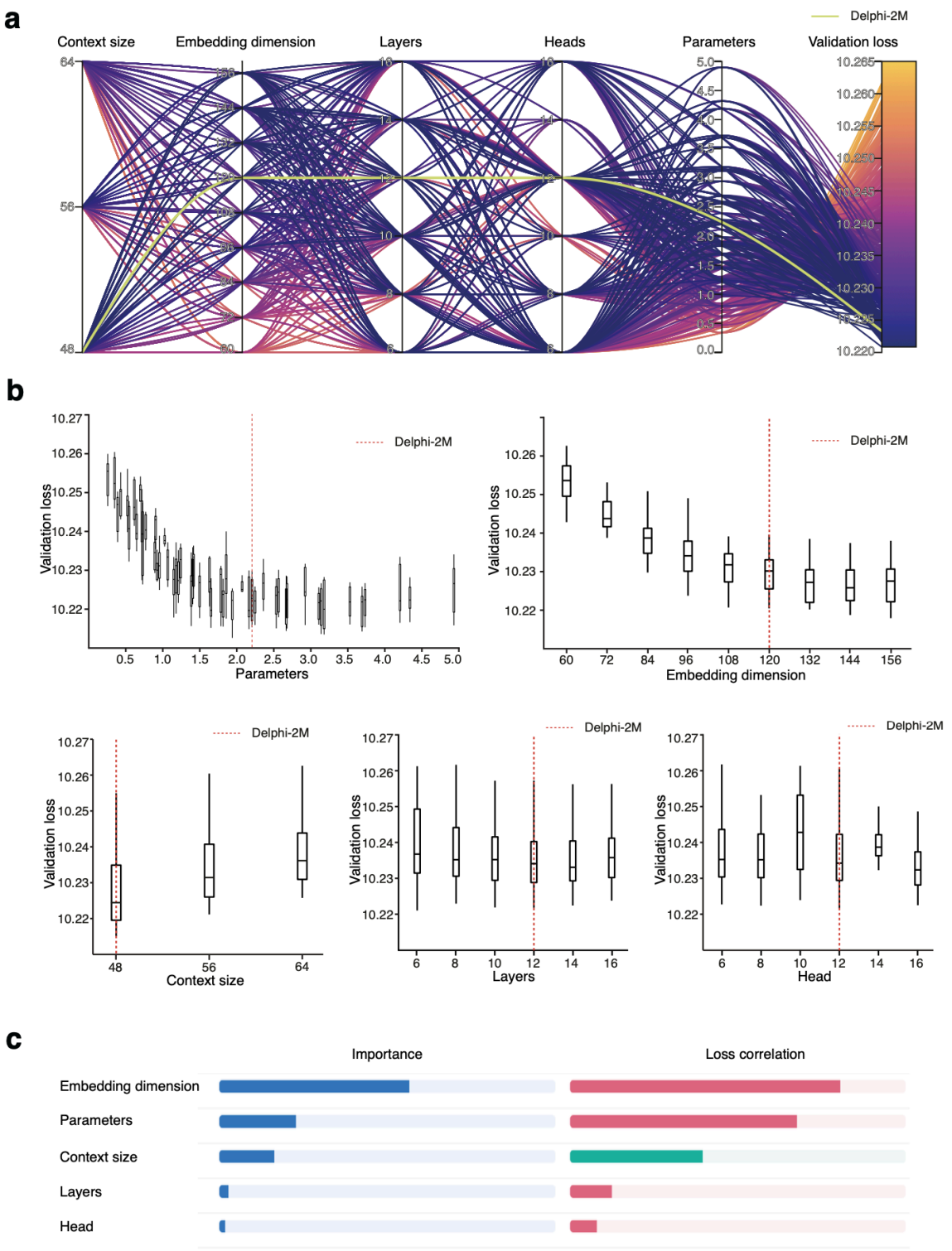


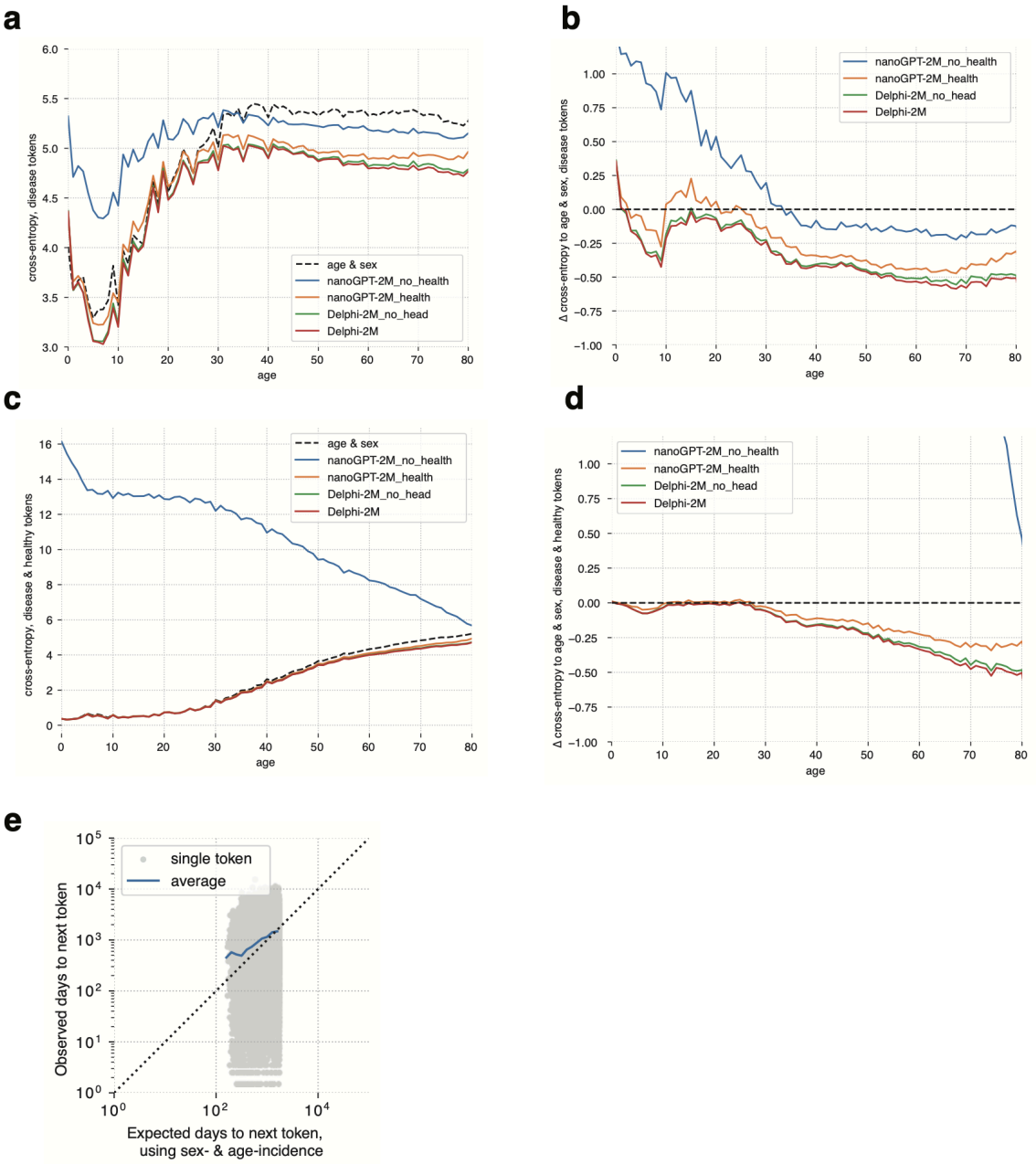
Figure 1. Delphi, a modified GPT architecture, models health trajectories. a, Schematic representation of health trajectories based on ICD-10 disease tokens, lifestyle and healthy padding. **b,** Training and validation (test) data derived from the UK Biobank. **c,** Modifications to the GPT-2 architecture. **d,** Example model input (prompt) and output. **e,** Scaling laws of Delphi. **f,** Ablation results measured by the cross-entropy differences relative to an age and sex based baseline. **g,** Accuracy of predicted time-to-event.



Extended Data Figure 1. Parameter screen. a, Validation cross-entropy (leftmost axis) for models trained with different architectural hyperparameter values (other axes). **b**, Same as **a,** taking a closer look at each parameter individually. The boxplots feature median as the center line, the box from the first to the third quartile and the whiskers for 1.5x IQR. **c,** Importance of different hyperparameters (via Weights & Biases Sweeps).

Loss	Age and Sex	GPT	+ healthy tokens	+ age enc. & masks	+ time loss
Next token class (cross-entropy)	0 (baseline)	–0.206	–0.478	–0.567	**–0.574**
Time to next token (exponential)	0 (baseline)	N/A	N/A	N/A	**–0.19**

Table 1. Ablation results for Delphi-2M. Top. Change in cross-entropy of disease token predictions relative to a baseline derived from age (in years) and sex incidence. **Bottom.** Change in exponential loss for the waiting time until the next token.



Extended Data Figure 2. Ablation study. a, Average cross entropy loss for disease tokens. **b**, cross-entropy loss for disease tokens relative to sex- and age-incidence baseline. **c,** Average cross entropy loss for disease and healthy tokens. **d**, Average cross-entropy loss for disease and healthy tokens relative to sex- and age-incidence baseline. **e**, Time to next token for sex- and age-incidence.

Delphi-2M accurately models reported 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 **Figure 2a**, the incidence curves are very varied, with some diseases such as chickenpox peaking in infancy, whereas 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 noticeable 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 measurable inter-individual differences of disease rates.

Delphi's ability to predict the next diagnosis token across the spectrum of human disease is confirmed by the average age-stratified AUCs (area under the receiver operating characteristic), which averages at values of approximately 0.8 in the internal validation split (**Figure 2b; Supplementary Table 1**). For 97% of recurrent diagnoses the AUC was greater than 0.5 indicating that the vast majority followed patterns with at least partially predictability. These patterns were found to be true across the different chapters of the ICD-10 spectrum, which define broad groups of disease. Among the most confidently predicted events in the short term 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' rates of the next tokens are statistically accurate (**Extended Data Figure 3**).

Delphi-2M's ability to utilise prior health information to accurately model the rates of the next possible outcomes across a wide spectrum of diseases is critical for predicting occurrences over the short term and also for generative sampling of health trajectories (see next section). Short-term predictions are often the consequence of acute events or diagnostic refinements that accrue over the course a few weeks or months, which may be undesirable for prognostication. After excluding inputs within the year prior to diagnosis Delphi-2M's AUCs still exceed those of the age and sex baseline by approximately 0.1 across most chapters indicating that its predictions are also relevant for prognostication (**Extended Data Figure 3a,b**). The same analysis, a longitudinal validation using UK Biobank data from 2020-2022 that was not part of model training, also shows that there was no substantial shift in diagnostic patterns throughout Biobank's follow-up.

Delphi-2M's disease predictions on the longitudinal validation data in the UK Biobank are comparable to clinical risk models like the Framingham[7] risk score for cardiovascular diseases, the UKBDRS[27] risk score for dementia and the Charlson[28] comorbidity index for mortality (**Figure 2d**). Delphi-2M's performance also appears stable when evaluated over varying prediction horizons ranging from 1-24 months, albeit with slightly decreasing performance over longer time periods (**Extended Data Figure 3c**). This shows that Delphi-2Ms' multi-disease predictions match current risk models for individual disease

outcomes, but offer the great advantage of being enabling the simultaneous assessment of more than 1,000 diseases at any given time.

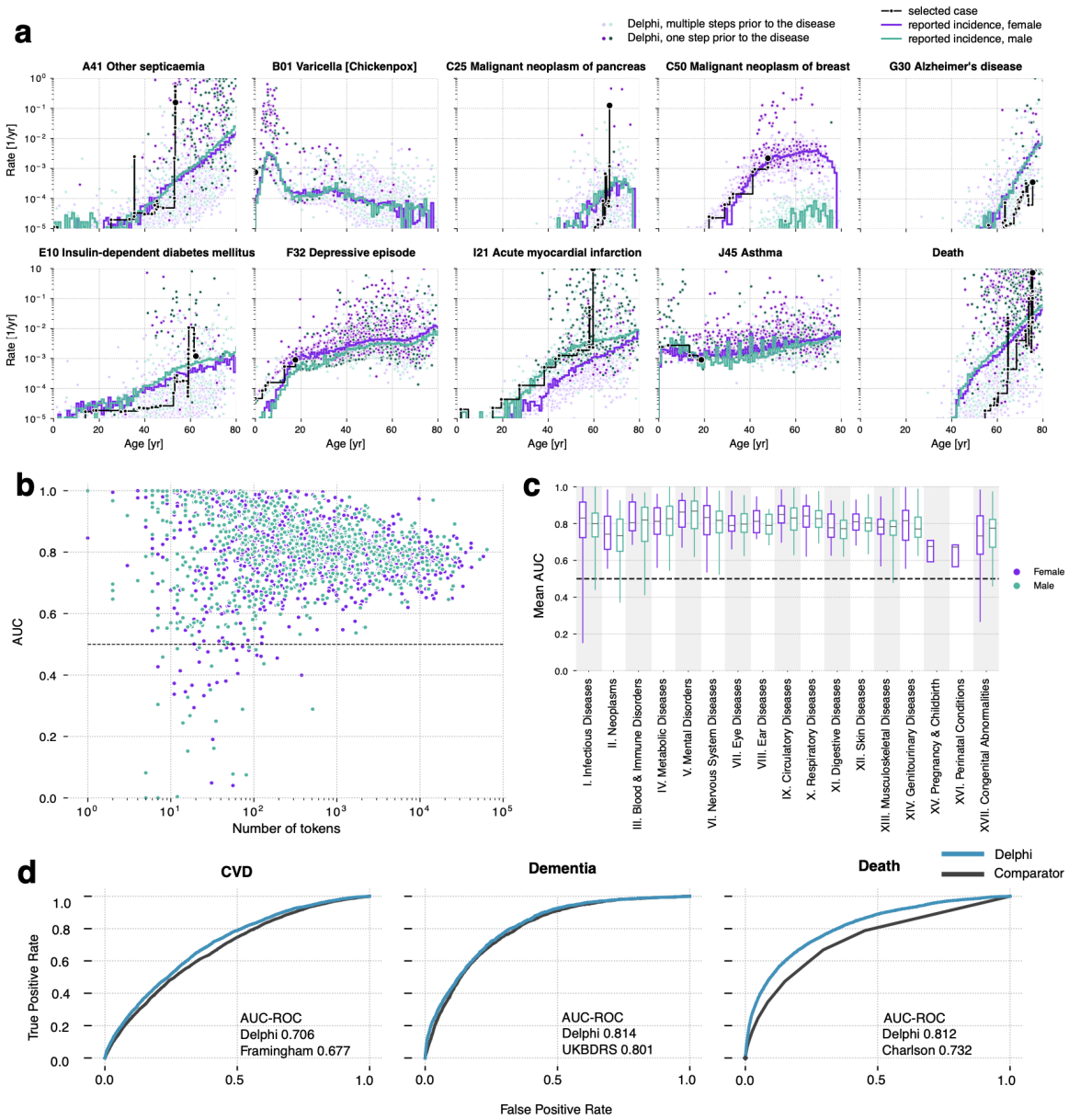
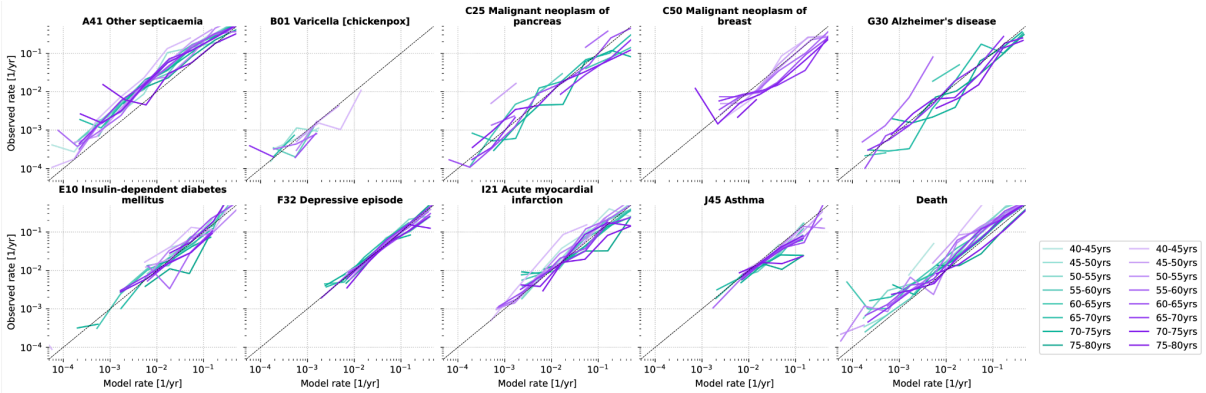
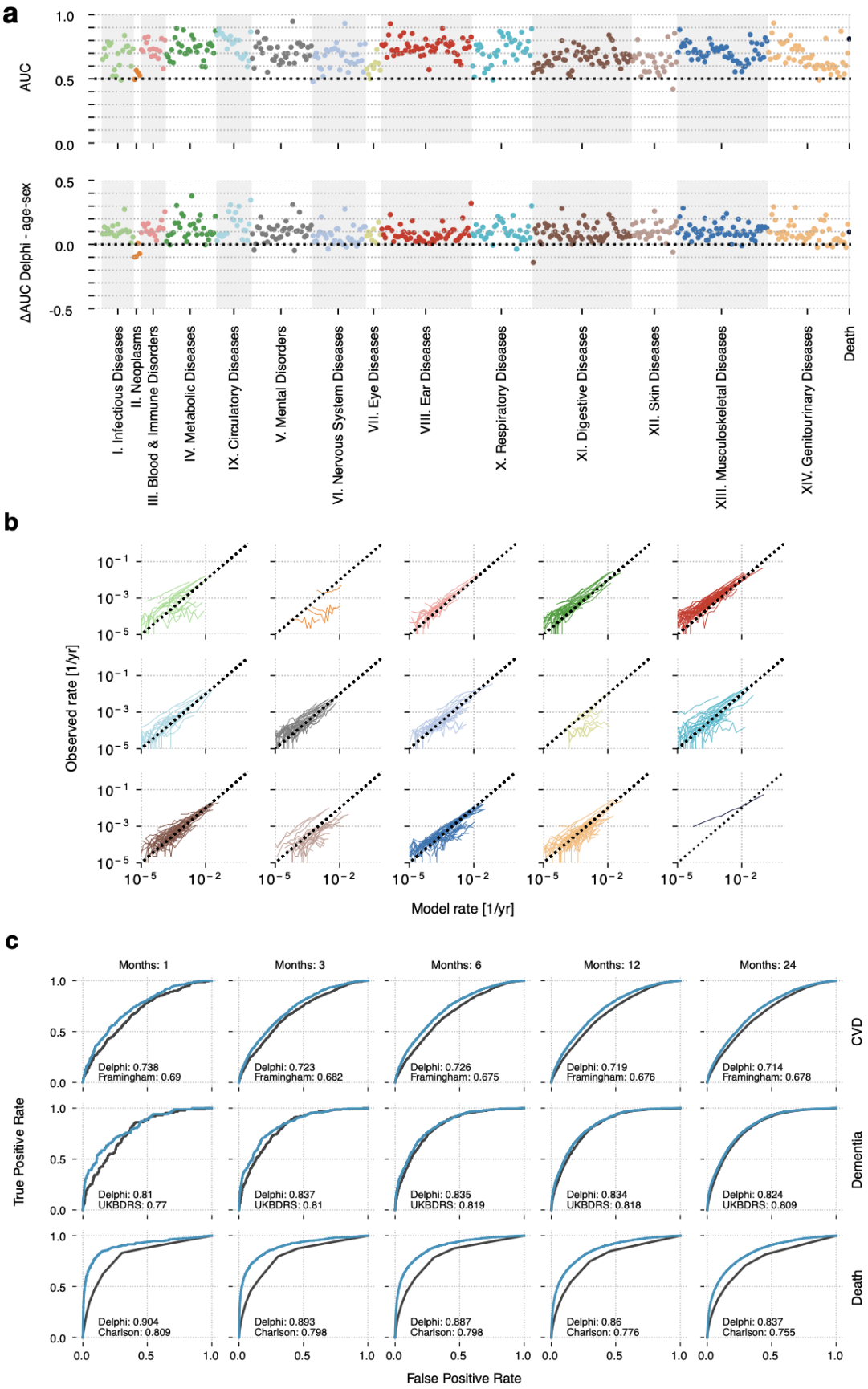


Figure 2. Delphi-2M accurately models the rates of a wide range of diseases. **a**, Predicted disease rates follow observed age-rate curves. **b**, Average AUC across 5 year age groups ranging from 40 to 80 years of age as a function of training occurrences. Shown are data for n=906 diagnoses for males and n=957 diagnoses for females for which sufficiently many events were recorded in the validation data to evaluate AUCs. **c**, Same as **b**, aggregated by ICD-10 chapter. The boxplots feature median as the center line, the box from the first to the third quartile and the whiskers for 0.025 and 0.975 quantiles. **d**, Comparison of AUC-ROC curves for Delphi and established individual disease risk models.



Extended Data Figure 3. Delphi-2M's instantaneous predictions are calibrated. Shown are predicted rates for 10 selected diseases averaged in bins of powers of 10 and evaluated for age groups of 5 years and sex (x-axis) versus observed rates in the same strata (y-axis). Delphi-2M's rates are well-calibrated when compared to empirically estimated from the dataset (observed) disease rates.


a
AUC
ΔAUC Delphi - age-sex
I. Infectious Diseases
II. Neoplasms
III. Blood & Immune Disorders
IV. Metabolic Diseases
IX. Circulatory Diseases
V. Mental Disorders
VI. Nervous System Diseases
VII. Eye Diseases
VIII. Ear Diseases
X. Respiratory Diseases
XI. Digestive Diseases
XII. Skin Diseases
XIII. Musculoskeletal Diseases
XIV. Genitourinary Diseases
Death
b
Observed rate [1/yr]
Model rate [1/yr]
c
Months: 1
Months: 3
Months: 6
Months: 12
Months: 24
CVD
Dementia
Death
True Positive Rate
False Positive Rate
Delphi: 0.738
Framingham: 0.69
Delphi: 0.723
Framingham: 0.682
Delphi: 0.726
Framingham: 0.675
Delphi: 0.719
Framingham: 0.676
Delphi: 0.714
Framingham: 0.678
Delphi: 0.81
UKBDRS: 0.77
Delphi: 0.837
UKBDRS: 0.81
Delphi: 0.835
UKBDRS: 0.819
Delphi: 0.834
UKBDRS: 0.818
Delphi: 0.824
UKBDRS: 0.809
Delphi: 0.904
Charlson: 0.809
Delphi: 0.893
Charlson: 0.798
Delphi: 0.887
Charlson: 0.798
Delphi: 0.86
Charlson: 0.776
Delphi: 0.837
Charlson: 0.755

Extended Data Figure 4. Longitudinal validation on UK Biobank data. a, AUC results of Delphi for each token with at least 25 occurrences coloured by the respective ICD-10 chapter. Predictions are based on UKB adults (50-80 yrs of age at cutoff) with data up until 30. June. 2020 and evaluated on incidence between 1. July. 2021 - 1. July. 2022. Below are the respective differences of AUC values between Delphi and a UK Biobank age-sex baseline. **b,** Calibration plots split by each ICD-10 chapter. Each line represents an ICD10 disease evaluated for each decile of the Delphi rate and compared against the observed rate in the population. **c,** Comparison of Delphi against established risk models (rows) for cardiovascular disease and death with a varying data gap between inputs and outcomes (columns). Shown are the ROC curves for Delphi-2M and the corresponding alternative risk prediction model.

Generative modeling of future disease trajectories

One of the most promising features of generative models is to sample disease trajectories, conditional on data recorded up to a certain point. This is a property which few conventional epidemiological models possess. Reassuringly, trajectories sampled from birth, only provided with the sex token, closely mirror the empirical age incidences and other population-level statistics for both males and females. At the same time, samples appear unique and do not copy any trajectories from the training dataset (**Extended Data Figure 5,6**).

In order to systematically assess the influence of medical histories on future health, we sampled health trajectories for each participant from the UK Biobank validation cohort based on data available until the age of 60 (**Figure 3a**). This provides the opportunity to compare 100,000 sampled and observed trajectories. When evaluated at the population level, the overall incidences at age 70 are well recapitulated, showing that the overall distributions are well preserved by iterative sampling **(Figure 3b**). This is further confirmed by the cross-entropy loss of sampled trajectories, which is, on average, indistinguishable from the observed data, but drops when the preceding disease histories are shuffled between participants (**Extended Data Figure 5b**).

In the first year of sampling, there are, on average 17% disease tokens that are correctly predicted, which drops to less than 14% 20 years later. These figures compare to values of 12-13% of correctly predicted disease tokens using sex and age alone, confirming that the conditional generation helps make more accurate predictions of future events (**Figure 3c**). Assessing the cumulative incidences of selected diseases among high and low risk groups reveals that Delphi-2M can predict the disease burden across a period of 20 years (**Figure 3d**). For the selected diseases the highest and lowest risk groups differ approximately 20-fold, which is largely reflected by Delphi-2M's sampled trajectories. While the precision is lower than that of short-term predictions these analyses show that Delphi-2M's conditional samples provide meaningful extrapolations for future health courses. This is not trivial as many of the imminent risks can change over time as more diseases accrue.

The use of synthetic data has been proposed to help overcome issues with privacy in biomedical modeling if such data sets do not reveal characteristics specific to any one person. Turning back to data sampled from birth, ie without any condition except for sex, our assessment shows that the generated trajectories do not exhibit any greater similarity to the training data than those from the validation cohort (**Extended Data Figure 6**). While partially overlapping disease trajectories may be found in terms of absolute disease tokens, the extent of overlap would be expected based on the observed incidences and co-morbidity patterns. Also, matching tokens occur on average 10 years apart underscoring that Delphi-2M does not memorise specific trajectories. This is also in keeping with its statistical rates, which at values between 10^{-4} to 10^{-2} per year are usually well below levels that would imply determinism (**Figure 2a**). Another piece of evidence that speaks to the quality of the generated synthetic trajectories is that they can be used to train Delphi-2M from scratch. Remarkably, such a recursively trained model retains many features of the original Delphi-2M model, for instance, it surpasses the age-sex baseline in terms of quality of next-token prediction (**Extended Data Figure 6g**).

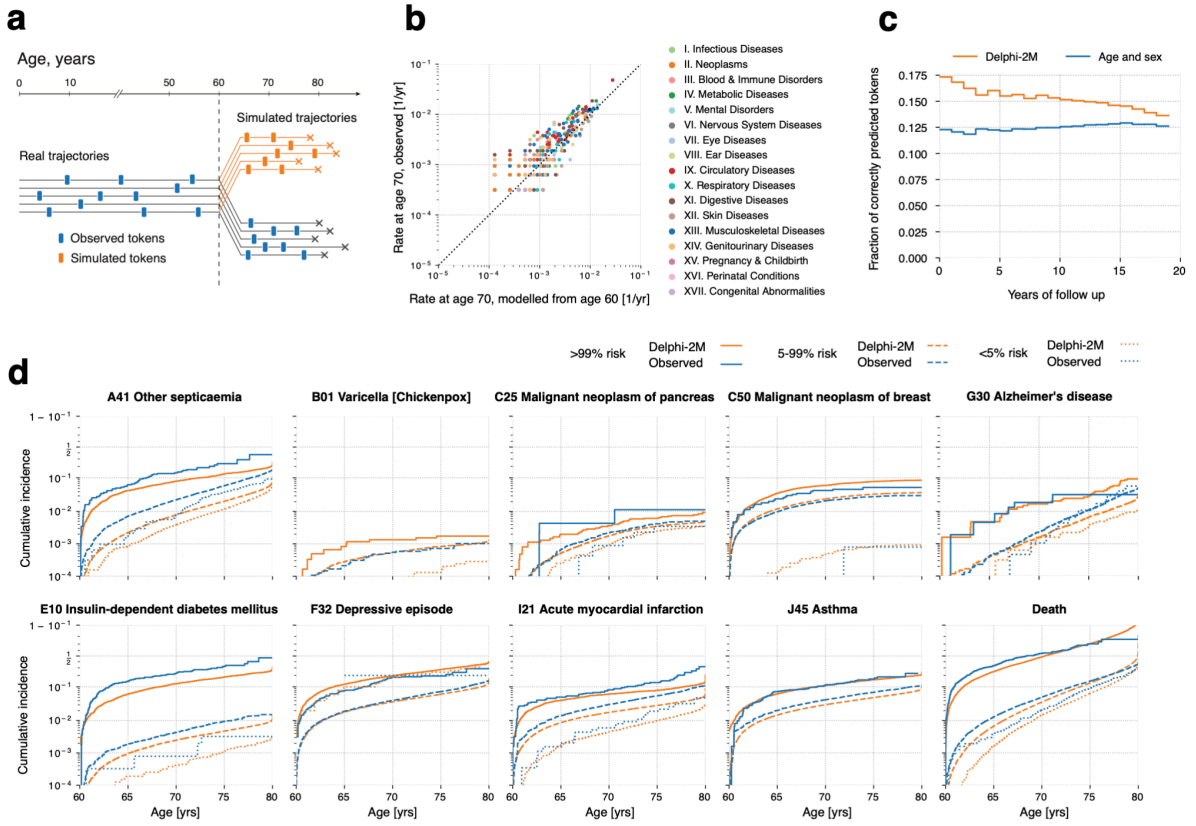
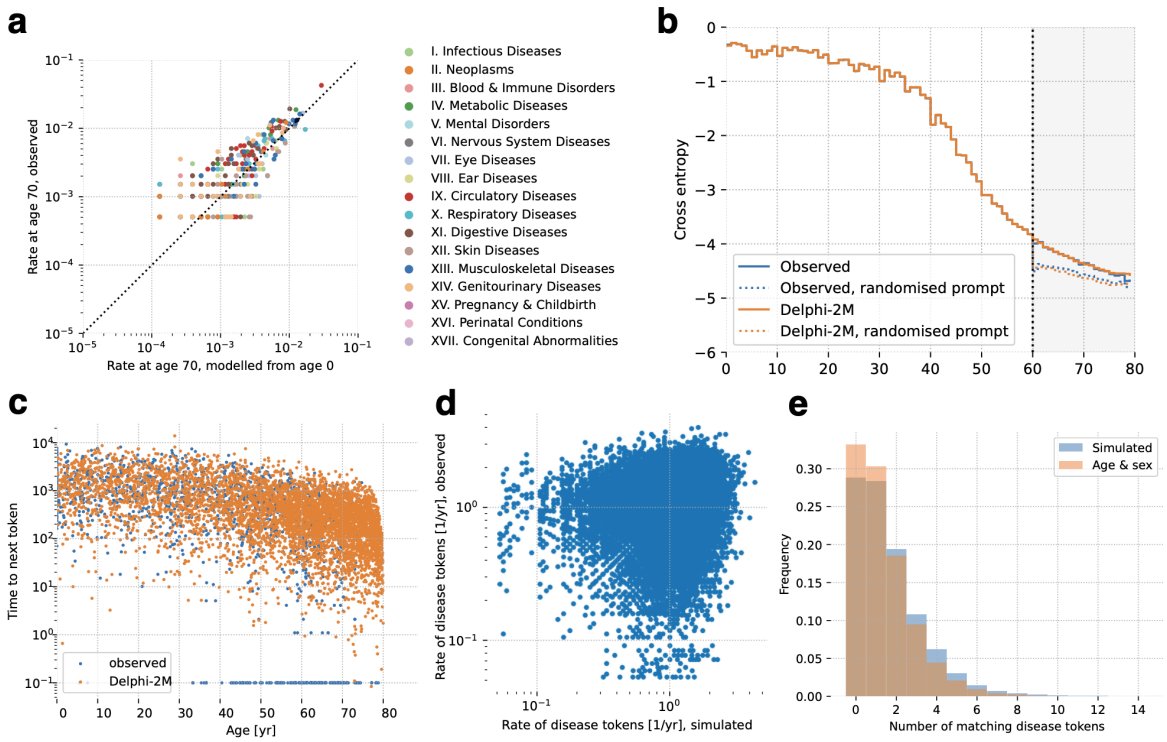
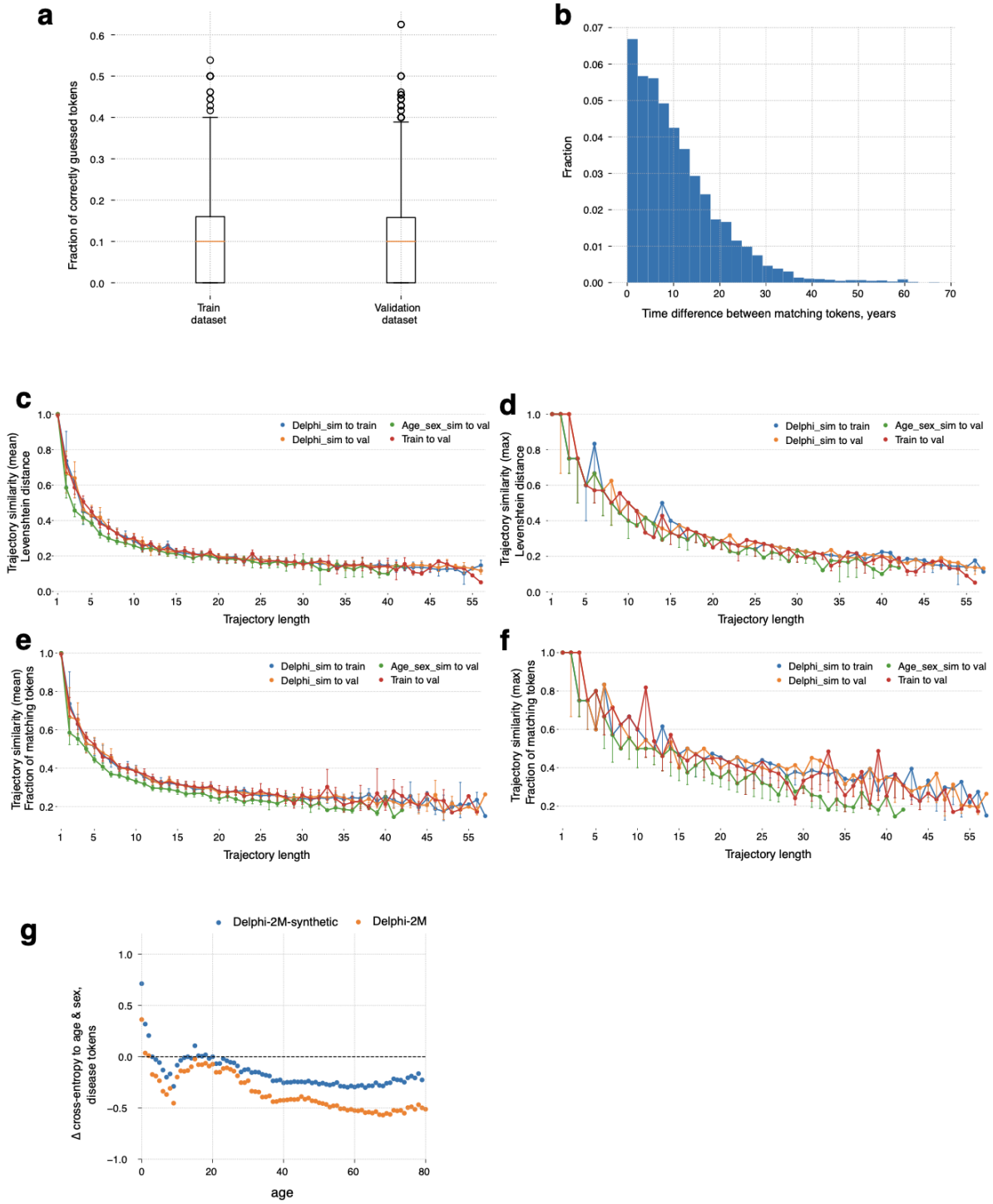


Figure 3. Generative modelling with Delphi informs future outcomes. a, Schematic representation of the experiment design. Delphi is used to continue real trajectories, the proposed continuations are compared to ground truth. **b,** Delphi-modelled disease rates match observed rates. **c,** Delphi predictions are more accurate compared to the age & sex baseline. **d,** Delphi accurately stratifies trajectories into low-, mid- and high-risk groups.



Extended Data Figure 5. Sampling health trajectories for long-term predictions. a, Delphi-modelled disease rates match observed rates. **b,** Breaking the trajectory structure by shuffling causes performance drop. **c,** Delphi time to the next event matches the observed time across all ages. **d,** Delphi disease rate (number of simulated tokens/15yr follow up from 60-75) versus the observed rate for each participant in the same period. **e,** Distribution of number of currently predicted tokens.



Extended Data Figure 6. Delphi-sampled trajectories are unique and allow for training models without accessing original data. a, When completing the partial trajectories from the train and validation subsets of the UKB dataset, the fractions of correctly guessed tokens are the same, indicanding that there is no train set overfit. The boxplots feature median as the center line, the box from the first to the third quartile, the whiskers for 1.5x IQR and the outliers. **b,** When sampling from scratch, even the trajectories that are similar to real ones token-wise have major shifts time-wise. **c-f,** In terms of fraction of matching tokens are normalised Levensthein distance, Delphi-sampled trajectories are closer to real ones compated to age-sex sampled. **g,** Delphi-2M-synthetic, a model

that is trained on trajectories sampled from Delphi-2M, still performs better than age-sex epidemiological baseline.

Explainable AI provides insights into disease progression

Insights into how Delphi-2M utilises past information to predict future disease rates can be obtained by assessing the structure of the disease embeddings. GPT models linearly map inputs into a lower dimensional embedding space, in which the temporal sequence of events is iteratively aggregated to produce a state from which predictions of each next token are derived (**Figure 1c**). Delphi-2M's specific implementation uses the same mapping from this final embedding state to the token risk, which guides the interpretation as the embedding matrix reflects the observed structure of co-morbidity risks.

As shown by the UMAP representation of Delphi-2M's embedding matrix in **Figure 4a**, disease codes cluster closely by the underlying chapter, a property the model has no direct knowledge of, and which purely reflects co-occurrence patterns in the data. Yet there are also noticeable exceptions, for example, cancers and precancers of the female reproductive tract. Another noteworthy cluster involves the two types of diabetes, retinal disorders and the neuropathies caused by them. Diseases with high acute mortality such as myocardial infarction or septicaemia are clustering with death. As participants age they progress through the embedding space, albeit often in a discontinuous fashion marked by jumps caused by acquired disease, and tend to attract certain health trajectories causing momentary shifts in the future risk of surrounding diseases (**Extended Data Figure 7**).

To gain further insights into how individual tokens influence future risks SHAP (SHapley Additive exPlanations)[29] values measure the influence of each token from the input trajectory on model predictions by a systematic assessment of subsampled data for individual prediction. As illustrated with the example of the proband's partial trajectory shown in **Figure 4b**, this analysis reveals that a series of disease diagnoses of the digestive tract (Chapter XI) elevated their pancreatic cancer risk 19-fold. The subsequent pancreatic cancer diagnosis, in turn, increased the rate of mortality almost ten thousand fold.

SHAP analysis of data from 100,000 individuals of the validation cohort reveals the mutual dependencies by which each disease, sex and lifestyle tokens influences the rate of disease tokens within the next 5 years (**Figure 4c, left)**. Effects mostly increase the rates of other diseases and are often found among diseases of the same chapters, underscoring that the recorded patterns of co-morbidities often cluster within specific ICD-10 disease chapters. Of note such patterns often appeared symmetrical, indicating similar predicted effect sizes of one disease token influencing another as vice versa (**Extended Data Figure 8**). Particular clusters spanning entire disease chapters are visible for chapters V. Mental Disorders and XV. Pregnancy & Childbirth. Moreover mortality is highly associated with diagnostic tokens from a broad range of chapters, including neoplasms.

Similar patterns are observed for dependencies after 10 years, even though the strength of associations is greatly attenuated (**Figure 4c, right)**. The cluster relating diseases of chapter XV. Pregnancy & Childbirth within 5 years is entirely absent after 10 years, which is expected given that pregnancy-associated diseases occur within a finite period. Dependencies among mental disorders, however, remain apparent, similar to the effects of neoplasms on mortality.

These observations are noteworthy as the quantification of temporal dependencies on past events poses a particular challenge for conventional epidemiological models whereas Delphi-2M's GPT model uses attention based weights which are updated with every new input, including the 'no other event' paddings.

To further illustrate Delphi-2M's capabilities of modeling temporal dependencies we note that for some diseases such as cancers, the influence on mortality decays with a half-life of several years reflecting the sustained risks of recurrence or impacts of treatment (**Figure 4d**). For septicaemia, however, the influence on mortality is much more short-lived and drops sharply, effectively recovering to values close to the population average. This inference agrees with traditional Nelson-Aalen analyses of the hazard rates (**Extended Data Figure 8f**). This behaviour is also reflected by Delphi-2M's attention maps, which show that cancer tokens are attended to for long periods, while that of septicaemia, myocardial infarction and many other diseases tends to be short-lived (**Extended Data Figure 9**).

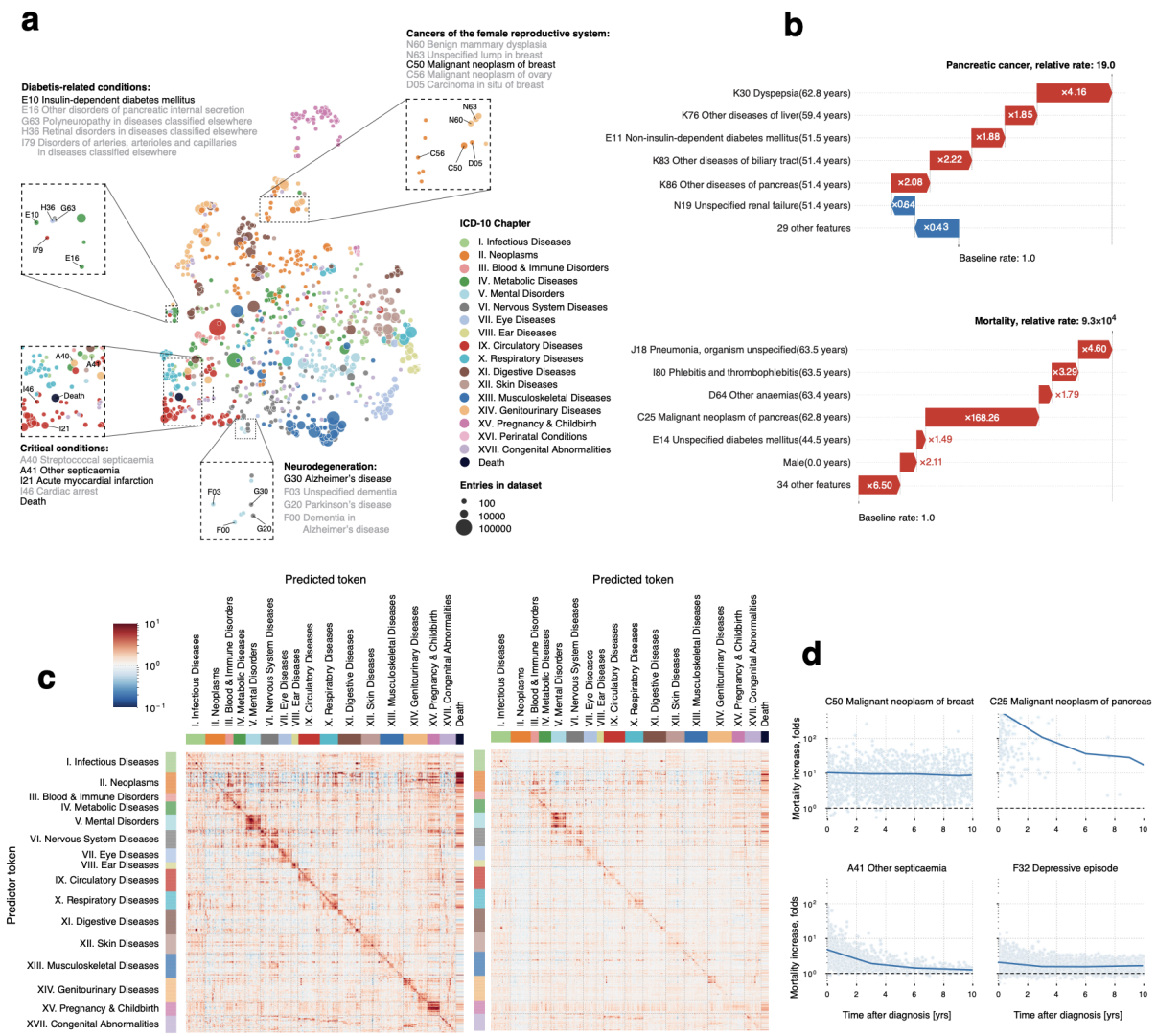
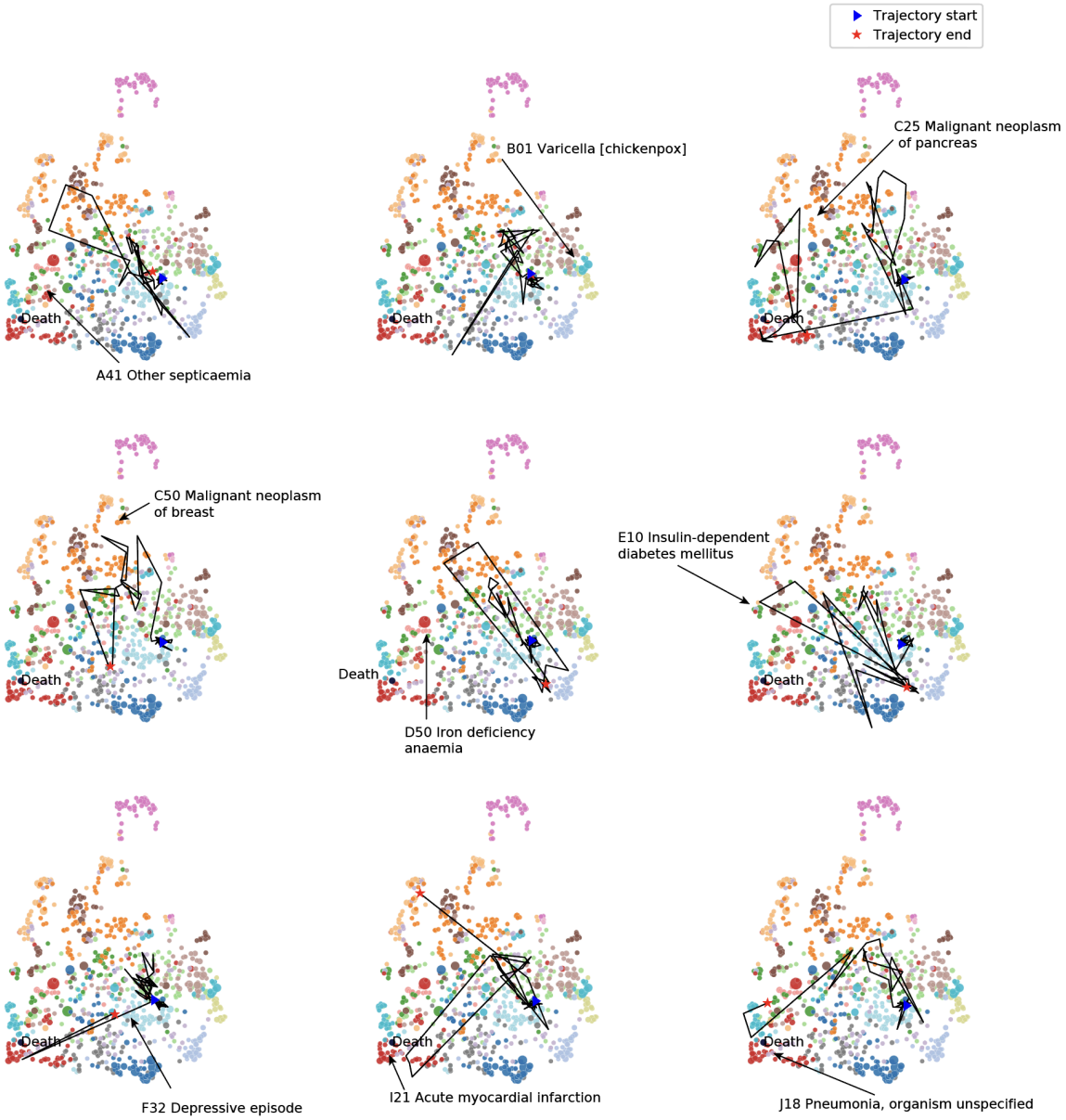
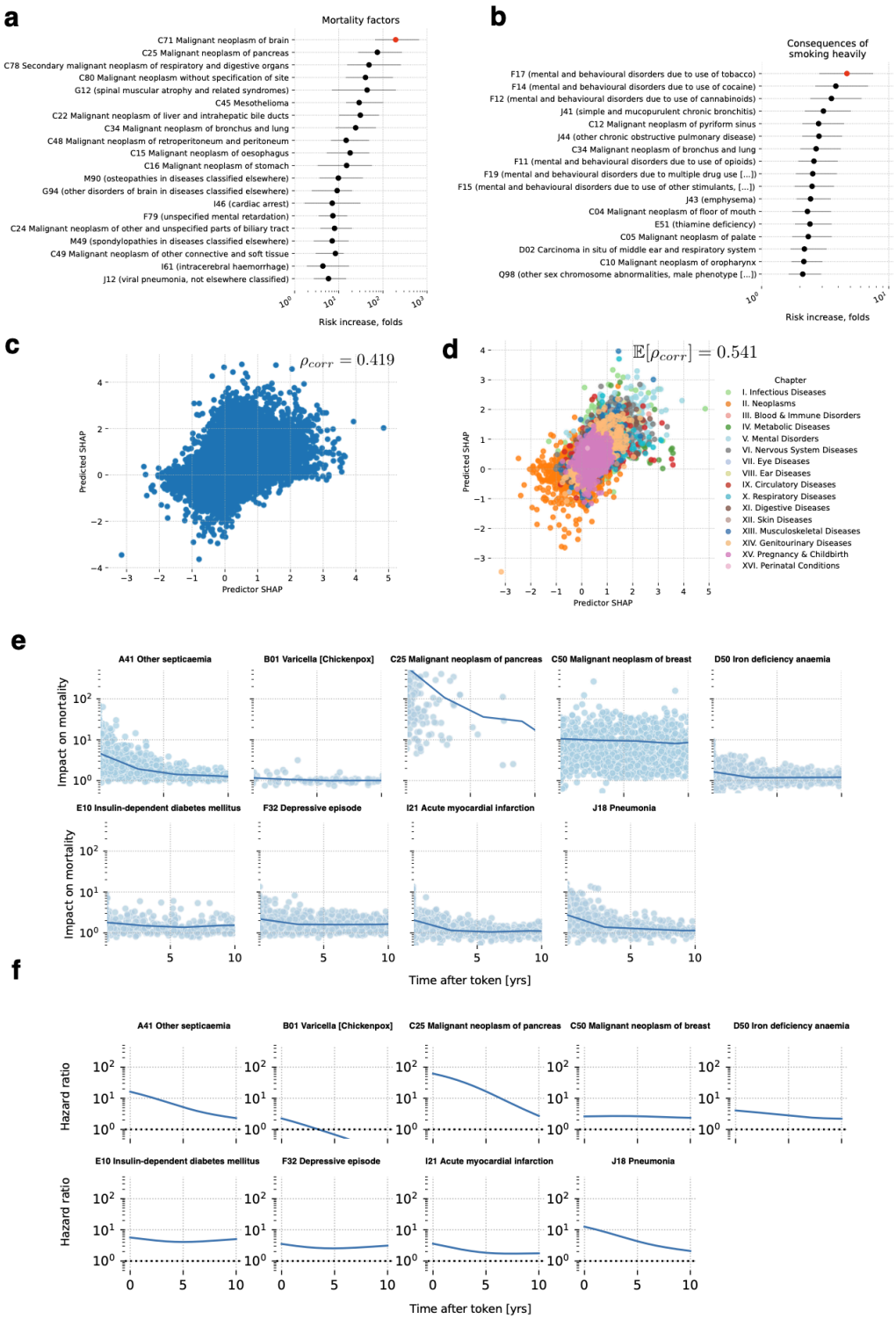


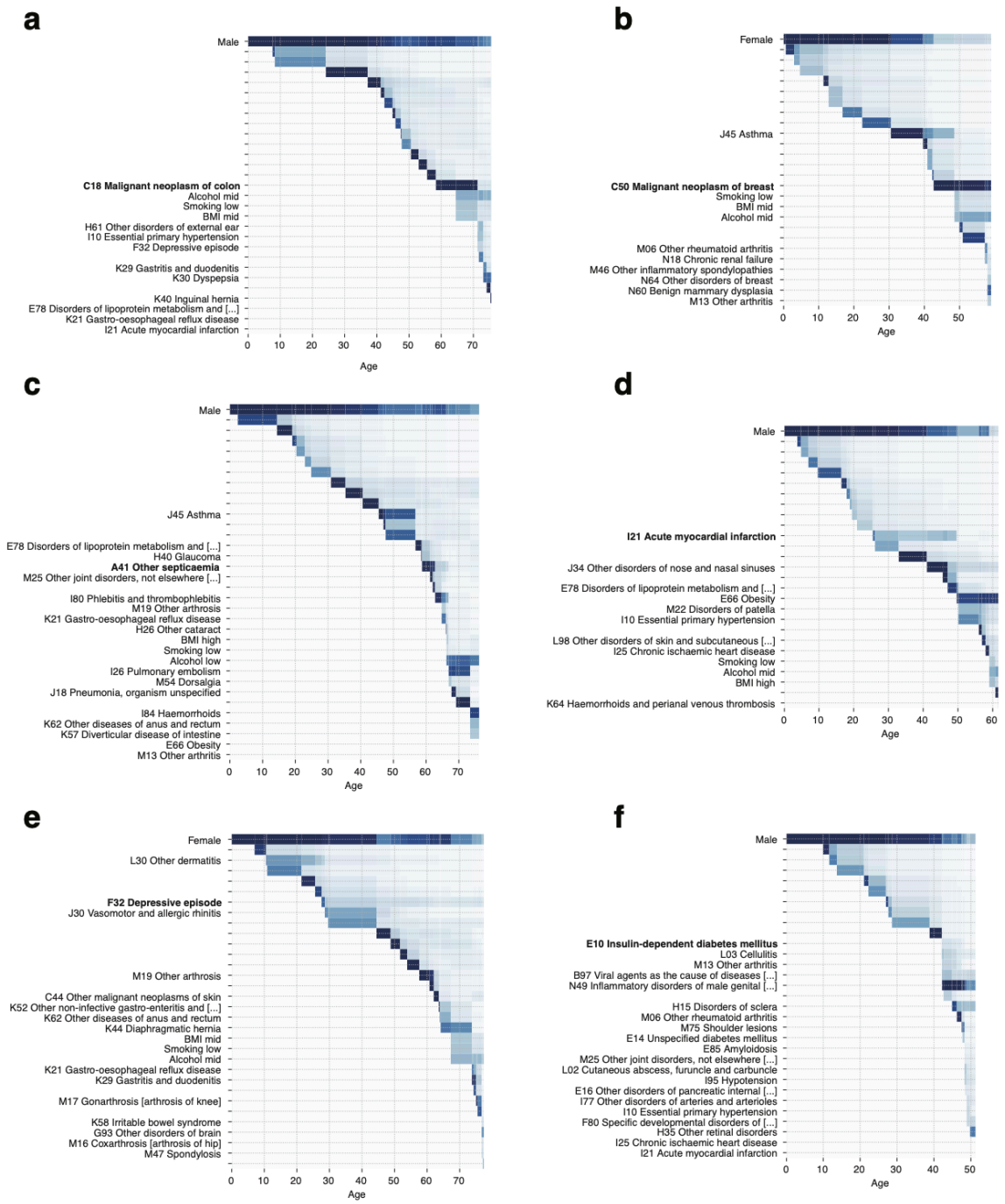
Figure 4. Explainable AI offers insights into disease progression. a, UMAP projection of token embeddings. Selected diseases are shown in the zoomed-in areas. Colours define disease chapters. **b,** SHAP-explained token risk contributions for individual trajectories. The top shows the risk of pancreatic cancer immediately prior to diagnosis at age 68.2, which was found to be 19x increased. The bottom panel shows the SHAP-based estimates of contributions to estimated mortality at age 63.5, which was greatly increased, in large parts due to the preceding diagnosis of pancreatic cancer. **c,** Heatmap of average rate increase that is by each of n=778 disease tokens with more than 5 occurrences and grouped by chapter (y-axis) on the same set of tokens plus death (x-axis). Red colours indicate a risk increase, whereas blue is a decrease. **d,** Relative effect (SHAP value) on mortality (y-axis) as a function of time after diagnosis (x-axis) for selected diseases.



Extended Data Figure 7. Embedding representation of health trajectories. Shown are ten health trajectories, each one representative of one of the 9 diseases indicated with an arrow. Shown is the last layer embedding at each time step projected into the weight embedding space with UMAP. In many cases the trajectory traverses the labelled diagnosis in close proximity, thereby temporarily raising the likely of surrounding diseases.



Extended Data Figure 8. Analysis of estimated influences among diagnoses. a, Strongest SHAP predictors of increased mortality risk. **b,** Tokens with largest risk increase, associated with having "Smoking - high" lifestyle token. **c,** Scatter plot of pairs "SHAP value of X predicting Y" and vice versa for all possible pairs. **d,** Same as c, but for pairs within ICD-10 chapters. **e,** SHAP values for mortality (y-axis) as a function of time after diagnosis of 9 different diseases (x-axis). **f,** Nelson-Aalen analysis of the hazard rates for the same diseases shown in **e** confirming shape and magnitude of SHAP effects.



Extended Data Figure 9. Attention maps indicate temporal effects of prior diagnoses. a-f. Attention for 6 selected trajectories representative of the token highlighted on the y-axis. Shown is the maximum attention across the 12 layers and 12 heads utilised by Delphi-2M. Darker values indicate higher attention (range 0-1). Typically the last (token, age) pair is most strongly attended to after which attention drops. Sex and selected other inputs are attended for longer periods. Rows without labels are “No other event” tokens.

External validation using Danish disease registries

The previous sections showed that Delphi-2M accurately reproduces health trajectories and the underlying complex dependencies of health and disease tokens as recorded in the UK Biobank. However, there is a risk that such patterns are specific to the particular training data and do not generalise to data sets from other populations. While the UK Biobank cohort has been recruited according to epidemiological best practices, due to the age of recruitment and the volunteer nature, the demographics are not entirely representative of the general population of the UK[30]. Furthermore, there may be differences in the availability of health-related data and how it is recorded between different healthcare systems.

To assess whether Delphi-2M's inference generalises to data from another country, we performed an external validation using Danish registry data. For this purpose, we transferred Delphi-2M with the learned weights from the UK Biobank training and evaluated predictions on the Danish data, hence, no retraining or any adjustments have been made. We use a cross-section of the entire Danish population aged 50-80 years on 1. Jan 2016 and evaluate Delphi-2M's predictions on the incidence of the corresponding population between 1. Jan 2017 and 1. Jan 2018, therefore, enforcing a 1-year data gap to also assess prognostication. Data are collected for 1.93M individuals (51% females and 49% males) with 11.51M disease tokens, recorded between 1978 and 2016. Predictions are evaluated on 0.96M disease tokens across 771 ICD10 codes (each with at least 25 cases).

The average ROC-AUC when Delphi-2M is applied to Danish data was 0.67 (s.d.=0.09) with similar performance across most chapters. The AUC difference to age-sex baselines were on average 0.07 (s.d.=0.08) points higher for Delphi-2M with 83% of all diseases showing a signal above the baseline (**Figure 5a**). When compared to the longitudinal validation in the UK Biobank, AUC values are similar with a correlation of 0.78 (95% CI: 0.71–0.79) (**Figure 5b, Extended Data Figure 4**). Further, the AUC differences over an age-sex baseline remain comparable across the two cohorts with a correlation of 0.54 (95% CI: 0.48–0.61) (**Figure 5c**).

Comparing the observed rate of individual diseases with the rate predicted by Delphi-2M across each decile shows good overall calibration (**Figure 5d**). However, there is a noticeable underestimation, especially for mortality, which we will discuss in the next section. Overall, Delphi-2M's predictions transfer well across two healthcare systems, without any reconfiguration, and show considerable prognostic performance across most diseases.

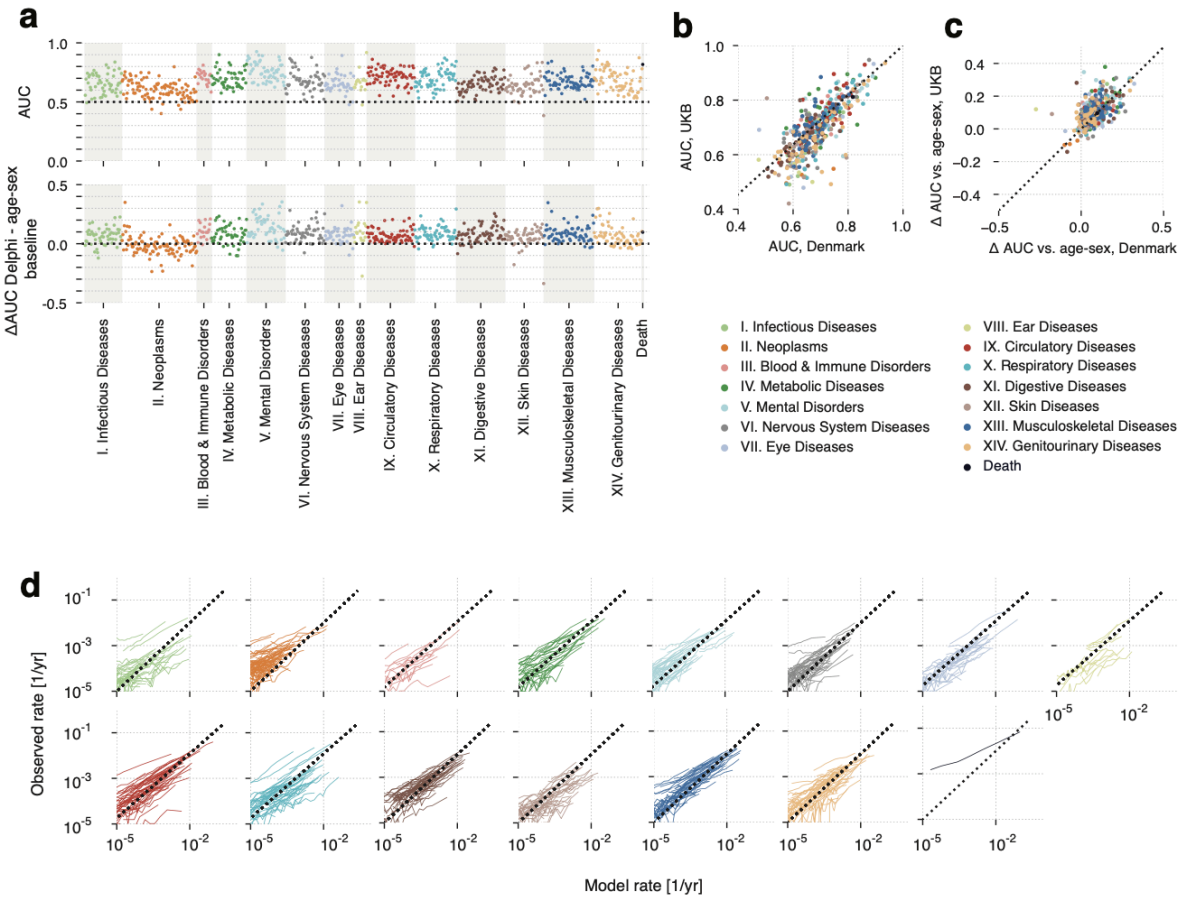


Figure 5. External validation using Danish registries. a, ROC-AUC results of Delphi for each token with at least 25 occurrences coloured by the respective ICD-10 chapter. Predictions are based on Danish adults (50-80 yrs. of age) with data up until 1. Jan. 2016 and evaluated on incidence between 1. Jan. 2017 - 1. Jan. 2018. Below are the respective differences of AUC values between Delphi and an age-sex baseline based on Danish data from 2010-2016. **b**, Comparison between AUC values in the UKB longitudinal validation and the Danish validation. Correlation is measured as pearson's rho. **c,** Comparison between delta AUC values in the UKB longitudinal validation and the Danish validation with their respective age-sex baselines. Correlation is measured as pearson's roh. **d,** Calibration plots on the Danish data similar to (a) split by each ICD10 chapter. Each line represents an ICD10 disease evaluated for each decentile of the Delphi rate and compared against the observed rate in the population.

Limitations, biases and fairness

While AI models can learn diverse and complex patterns, they may also replicate biases that exist in the underlying training data. Delphi-2M is no exception to this and its reduced accuracy on Danish registries suggests a range of biases in the UK Biobank. In particular, we identify and discuss the consequences of the following biases underlying UK Biobank training data: Recruitment bias, sampling/immortality bias, and health data availability bias. Similar limitations will exist to some degree in most other data sources and are important considerations when training AI models on healthcare data.

Epidemiological cohorts often only approximately reflect the general population due to differential response rates in different subpopulations. The UK Biobank comprises more white British citizens than other ethnicities and participants tend to be, on average, more affluent and educated than the general population[30]. While Delphi-2M accurately replicates the observed rates of disease tokens in UK Biobank training and validation data, there are more pronounced differences to those recorded in Danish registries, even though a correlation of 0.7 clearly remains (**Figure 6a**). On average incidence differed by a factor of 3 with a slight tendency of underreporting in the UK Biobank in line with a healthy volunteer bias.

Other biases arise from the use of retrospective data in a cohort. In the context of the UK Biobank, most individuals have been recruited between 40-50 years of age. This creates a type of immortality bias as no deaths are recorded prior to recruitment. This has direct implications on estimated mortality (**Figure 6b**), and indirectly also on the incidences of diseases associated with high mortality, such as cancers, as only survivors are included in the UK Biobank. For time-dependent analyses, the jump of mortality to non-zero values at recruitment can also lead to false attribution of the apparent increase to unrelated variables recorded at the time of recruitment.

Furthermore, the heterogeneous sources of UK Biobank's first occurrences data and their missingness patterns introduce artificial correlations. The first occurrence data has been collated from self-reports, primary care, hospital admissions, cancer and death registries where each source contributes characteristic disease tokens, but may be missing in a subset of individuals or cover only specific periods in time (**Figure 6c,d, Extended Data Figure 11**). Self-reporting and GP records contain mostly common diseases, with self-reporting covering 78% of participants, while GP records were only available for just 45% of participants. Data from hospital records covered 86% participants and contributed more aggressive disease tokens, such as myocardial infarction or septicaemia, similar to the death registry. Lastly, cancers were provided by the national cancer registry.

Delphi-2M can learn patterns of missingness and use them to predict future rates accordingly – even though this concerns data availability rather than true incidence. Accordingly the predicted rates of diseases which are specific to hospital records are on average 10x higher in individuals which have a history including other hospital records (**Figure 6e**). Septicaemia, for example, is diagnosed in 93% of cases in a hospital setting and is predicted to occur at 8x greater rates in individuals with any other hospital data. Similarly, the predicted rates for primary care linked diseases are on average 3x higher in participants with other primary care data (**Figure 6f**). For example, acute sinusitis is almost

exclusively (99%) found in primary care data and is predicted to occur at 3.5x greater rate in individuals with any linked primary care data. While some of these associations may reflect true diagnostic pathways linked to a particular setting, it nevertheless appears that many of these associations are artefacts stemming from the incomplete aggregated nature of UK Biobank's data.

Another consideration of healthcare models is whether predictions are made with similar accuracy in different population groups. The predicted rates of observed diseases are largely comparable between the two sexes, however it is lower for individuals self-reporting a "White" ethnic background compared to any other ethnic group (**Extended Data Figure 11a**). This is likely a consequence of participants of "White" ethnic background having lower numbers of disease tokens across time, which reduces the ability to predict future disease burden (**Extended Data Figure 11c**). Similarly, Delphi-2M's predicted rate of disease is highest in the top quintile of Townsend deprivation, which mirrors the average number of tokens per individual in this group (**Extended Data Figure 11b,d**). Together these observations indicate that Delphi-2M's predictions reflect the differential burden of disease seen across different population subgroups.

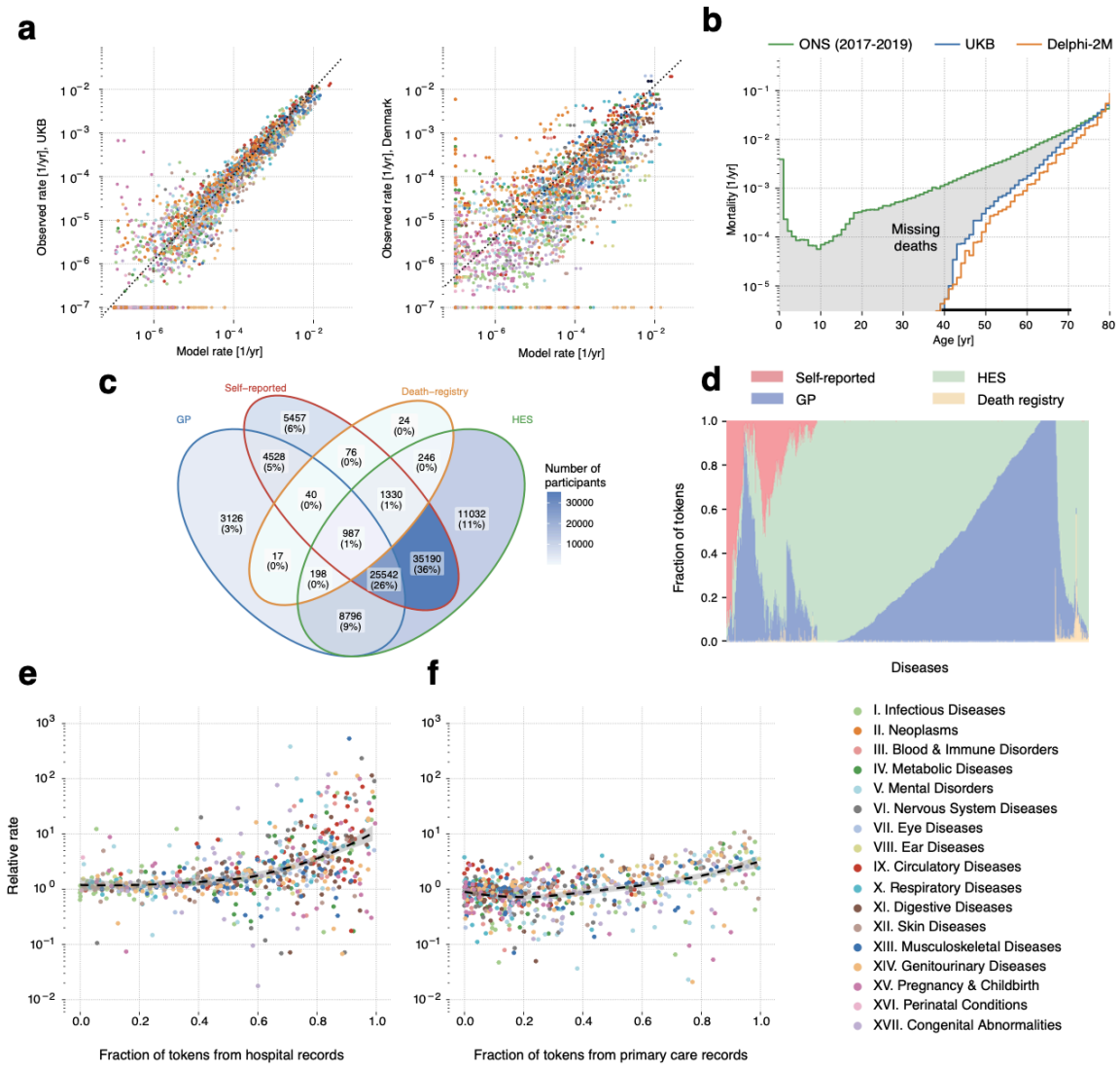
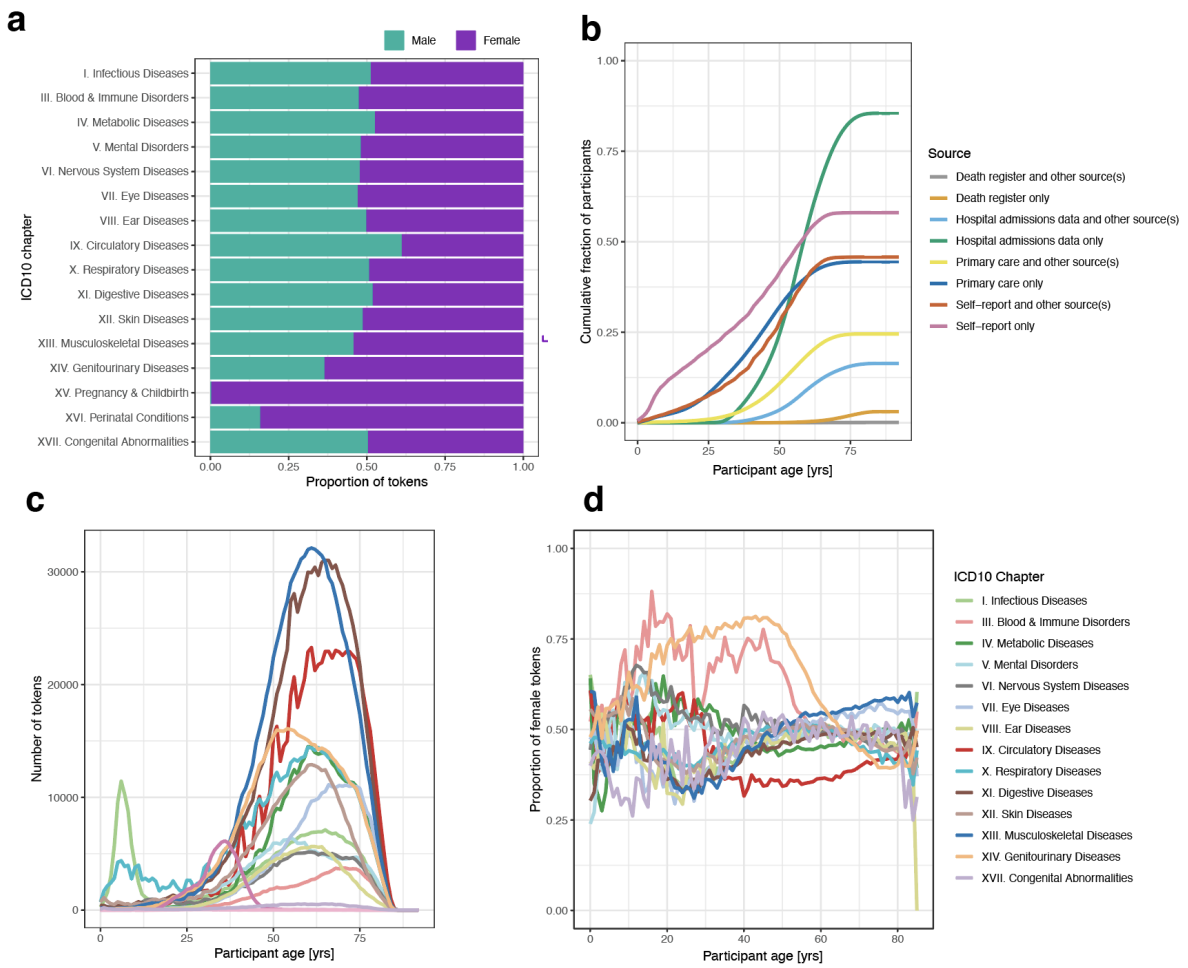
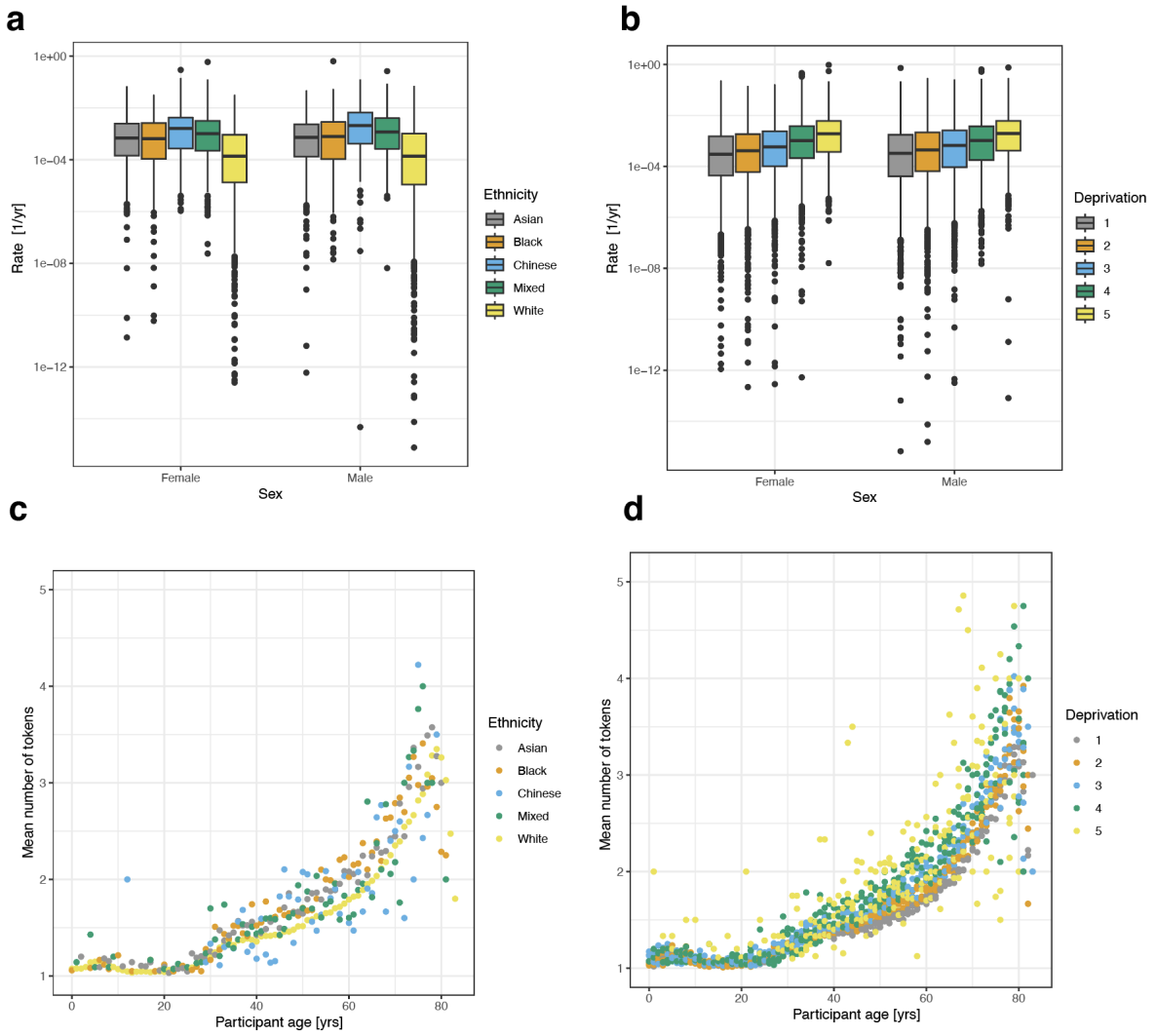


Figure 6. Epidemiological biases in UK Biobank data reflected by Delphi. **a,** Model predicted and observed yearly disease incidence at age 70 in UK Biobank (left) and Danish population registries (right). Colours denote disease chapters as defined as part of panel **f**. **b,** Yearly mortality estimates by Delphi-2M, observed rates in the UK Biobank and ONS national estimates across the entire British population. As only living individuals between 40-70yrs of age (black line) were recruited to the UK Biobank many deaths are missing compared to the ONS population estimate (grey shaded area). **c,** Venn diagram of disease data availability in UK Biobank validation cohort (n=100,000). **d,** Data source per disease token (token position sorted). **e,** Token exclusivity (hospital records) against hospital-records missingness bias, coloured by ICD10 chapter, the overall trend is shown in black using a non parametric (loess) curve with 95% confidence intervals shown in grey. **f,** Token exclusivity (primary care) against primary care missingness bias, coloured by ICD10 chapter, the overall trend is shown in black using a non parametric (loess) curve with 95% confidence intervals shown in grey.



Extended Data Figure 10. ICD10 disease token biases. **a,** The proportion of the total disease token set scaled by the number of samples comparing female and male samples across ICD10 chapters. **b,** Participant age against the cumulative sum of the fraction of participants with disease tokens across 8 different primary data sources. **c,** Participant age against the total number of disease tokens across ICD10 chapters. **d,** Participant age against the proportion of tokens in females across ICD10 chapter (excluding chapters P and O).



Extended Data Figure 11. Fairness and limitations. **a,** Rate per year separated by sex and ethnic background. **b,** Rate per year separated by sex and Townsend deprivation index bins (increasing for greater deprivation index values). The boxplots in a and b feature median as the center line, the box from the first to the third quartile, the whiskers for 1.5x IQR and the outliers. **c,** Participant age against the average number of disease tokens separated by ethnic background. **d,** Participant age against the average number of disease tokens separated by Townsend deprivation index bins.

Discussion

Here we presented Delphi-2M, a GPT-based model of multi-disease progression. Delphi-2M extends the GPT large language model to account for the temporal nature of health trajectory data. Analogous to LLMs, which learn the grammar and contextual logic of language from large bodies of text, Delphi-2M inferred the patterns of multi-disease progression when trained on data of more than 1,000 diseases and baseline health information recorded for 400k UK Biobank participants. For this training corpus a, model size of approximately 2M parameters appeared optimal with bigger data sets poised to yield even greater accuracy with larger models.

A detailed assessment of Delphi-2M's predictions shows that they faithfully recapitulate the patterns of disease occurrence at the population scale as recorded in UK Biobank. Such patterns are largely dominated by age and sex, but also prior health data. For the majority of diseases Delphi-2M's multi-disease state model predicted future rates at comparable accuracy as established single-disease risk models. Furthermore, Delphi-2M achieved very good performance when applied to data from Danish disease registries without additional finetuning, demonstrating that models are largely applicable across national healthcare systems.

Delphi-2M is uniquely capable of sampling future disease trajectories, which enables estimating cumulative disease burdens over periods of up to 20 years, conditional on prior health information. It is worth noting that Delphi's predictions are generally strongly influenced by statistical chance and compatible with a range of outcomes for a given individual. The ability to generate synthetic data may also help create data sets that preserve the statistical co-occurrence patterns without revealing any specific data, which could facilitate the development of further AI models with a decreased risk of revealing personal information.

Delphi-2M's underlying GPT-2 model offers insights into the modes of disease progression. The ability to cluster disease risks may be useful for genomic association studies that focus on comorbidities or are stratified by the risks derived from health trajectories. Delphi-2M's capability to quantify the temporal influence of previous health data revealed that cancers increase mortality in a sustained fashion, while the effects of myocardial infarction or septicaemia regressed within 5 years. Similar analyses also revealed clusters of persisting co-morbidities such as mental health conditions, which informs healthcare planning.

There are, however, also a range of limitations that need to be considered. A detailed analysis of UK Biobank's first occurrence data revealed several biases, reflected by Delphi-2M. In addition to a healthy volunteer bias and sampling/immortality bias of participants before recruitment, the diverse nature of health data sources had an impact on Delphi's prediction as UK-Biobank-specific patterns of missingness are exploited to infer rates of disease. These findings underscore that care must be taken when AI models are used for inference and applied for prediction in heterogeneous healthcare data sets.

A further promising feature of Delphi-2M's implementation is the relative simplicity with which additional data layers may be incorporated in future extensions. Examples of this are the simple additive nature of age encoding or how lifestyle and sex tokens are superimposed on the embeddings of disease history information. Immediate refinements of Delphi-2M may incorporate additional lifestyle data, prescription records and blood tests, both usually available in a general healthcare setting. Further multi-modal extensions could include genomic data, richer metabolomic information, or data from wearables can be added to Delphi-2M's embedding layer similar to how lifestyle tokens are currently incorporated. Such multi-modal extensions will shed light on how different layers of biology interface in shaping health outcomes and thus help address long-standing questions in biomedicine. Lastly, while ICD-10 provides a predefined tokenisation of diseases, LLMs have been shown to also conceptualise natural language making it plausible to expect that future models may derive similar meaning directly from free text records.

These considerations illustrate the wide range of applications of generative models for biomedical research and, ultimately, also for healthcare. The ability to model instantaneous and cumulative risks harbour large potential for health economics and healthcare itself. Risk models are important to efficient cancer screening and to enable preventative medicine. The ability to generate health trajectories may be used to project future health burdens, which will help allocate resources that can be scarce in ageing populations. Future model extensions that also incorporate treatments and outcomes may be used to directly assess treatment effects with respect to prior health data and other information including genomics with utility for the individual and society. Together these findings highlight the considerable healthcare benefits from the era of AI.

Methods

Data

UK Biobank

Cohort: The UK Biobank is a cohort-based prospective study comprising approximately 500,000 individuals from various demographic backgrounds recruited across the UK between 2006 and 2010. At the time of recruitment, individuals were between the ages of 37 and 73[31].

Disease first occurrence data. The main data source for health related outcomes is built on the first occurrence data assembled in category 1712 from the UK Biobank. These data include ICD10-level 3 codes (e.g. E11: Type 2 diabetes mellitus) for diseases in chapters I-XVII excluding chapter II - neoplasms (1256 in total) plus death. The data is pre-assembled by UK Biobank and includes the first reported occurrence of a disease in either the linked primary care data (cat. 3000), inpatient hospital admissions (cat. 2000), death registry (fields: 40001, 40002) or self-reported data through questionnaires (field: 20002).

Information on neoplasms was not included in category 1712 by the UK Biobank, hence we included the data ourselves through addition of the linked cancer registry data in fields 40005 and 40006 (subset to first occurrence and mapped to ICD10 level 3 codes).

Lifestyle and demographics. We extract information on the self-reported sex of participants as recorded in field 31 (indicators for female and male), a physical assessment of body mass index at recruitment from field 21001 which we split into 3 indicators encoding bmi < 22 kg/m^2, bmi > 28 kg/m^2, and otherwise, as well as smoking behaviour from field 1239 with indicators for smoker (UKB coding: 1), occasionally smoking (2) and never smoker (0) and alcohol intake frequency from field 1558 with indicators for daily (1), moderately (2, 3) and limited (4, 5, 6).
Further, information that we extracted and used for stratification to assess model performance in subgroups but were not part of the data for model training include self-reported ethnic background as available in field 21000 participants grouped into 5 level groups (White, Mixed, Asian or Asian British, Black or Black British and Chinese) and an index of multiple deprivation as available in field 26410. The index combines information across seven domains including Income, Employment Derivation, Health and Disability, Education Skills and Training, Barriers to Housing and Services, Living Environment, and Crime.
Additionally, we extract information on diastolic blood pressure (field: 4079), systolic blood pressure (field: 4080), high density lipoprotein - hdl (field: 30760), total cholesterol (field: 30690) and hypertensive medication use (field: 20003 - any mentioned of: enalapril, lisinopril, perindopril, ramipril, candesartan, irbesartan, losartan, valsartan, olmesartan, amlodipine, felodipine, nifedipine, diltiazem, verapamil, indapamide, bendroflumethiazide, atenolol, bisoprolol, doxazosin, amiloride, spironolactone) which are used for the Framingham score evaluation. Further, information on years of education (field: 845), dementia of parents (fields: 20107, 20110) as well as living alone (field: 709) were extracted for the UKBDRS evaluation.

Danish registries

Cohort. Exploring comorbidities and health-related factors is uniquely facilitated by Denmark's comprehensive registries, which gather up to 40 years of interconnected data from across the entire population. All of the used registries are linkable through a unique personal identification number provided in the Central Person Registry along with information on sex and date of birth. Further, we utilise the Danish National Patient Registry[32] (LPR) a nationwide longitudinal register with data on hospital admissions across all of Denmark since 1977 along with the Danish Register of Causes of Death[33] since 1970, to extract information on an individual's acquired diagnoses throughout their lifetime. Our current data extract covers information up until around 2019 when reporting to the LPR was updated to LPR3. Further, we restrict our cohort to individuals aged 50-80 on 1st of January 2016, to obtain a similar age range as in the UK Biobank. The 1st of January 2016 was chosen as the cutoff point as it is the latest time point for which we can guarantee reliable coverage across the entire population over the entire prediction horizon.

Feature adjustments. In order to obtain a dataset that resembles the UK Biobank data we only retain the first occurrence of an individual's diagnosis and transform all codes to ICD10-level 3 codes. Diagnoses before 1995 are reported in ICD8 and have been converted

to ICD10 codes using published mappings[34]. Codes that may be present in the Danish register but were not in the UK Biobank are removed. Information on lifestyle is not available, hence indicators for BMI, smoking and alcohol intake have been treated as absent.

Data splits

UK Biobank. The models were trained on UK Biobank data for 402,786 (80%) individuals using data from birth until 30th of June 2020. For horizontal validation, data contains the remaining 100,636 (20%) individuals for the same period. Internal longitudinal validation was carried out using data for all individuals still alive by the cutoff date (471,057) and evaluated on incidence from 1st of July 2021 to 1st of July 2022, therefore enforcing a 1-year data gap between predictions and evaluation. Horizontal validation assesses how well the model generalised to different individuals from the same cohort. Longitudinal validation investigates whether the model's performance changes over time and if it can be used for prognostic purposes.

Denmark. External longitudinal validation was conducted on the Danish registries. All individuals residing in Denmark aged 50-80 years on the 1st of January 2016 were included. Predictions are based on the available data up to this point and subsequently evaluated on incidence from 1st of January 2017 - 1st of January 2018 similar to the internal longitudinal validation.

Model architecture

GPT model. Delphi's architecture is based on GPT-2[25], as implemented in https://github.com/karpathy/nanoGPT. The basic GPT model uses standard transformer blocks with causal self-attention. A standard lookup table embedding layer with positional encoding was used to obtain the embeddings. The embedding and casual self-attention layers are followed by layer normalisation and a fully connected feedforward network. Transformer layers, consisting of causal self-attention and feedforward blocks, are repeated multiple times before the final linear projection that yields the logits of the token predictions. The residual connections within a transformer layer are identical to those in the original GPT implementation. Here we also use a weight tying of the token embeddings and final layer weights, which has the advantage of reducing the number of parameters and that input and output embeddings can be similarly interpreted.

Data representation and padding tokens. Each data point consists of pairs (token, age) recording the token value and the proband's age, measured in days from birth, at which the token was recorded. The token vocabulary consists of n=1257 different ICD-10 level 3 disease tokens, plus n=9 tokens for alcohol, smoking and BMI, each represented by 3 different levels, as well as n=2 tokens for sex and n=1 no-event padding token as well as n=1 additional, non-informative padding token at the beginning or end of the input sequences.

No-event padding tokens were added to the data with a constant rate of 1/5 years by uniformly sampling 20 tokens from the range of (0, 36525) and interleaving those with the data tokens after intersecting with the data range for each person.

Sex tokens were presented at birth. Smoking, alcohol and lifestyle were recorded at the enrollment into the UK Biobank. As this specific time also coincided with the end of immortal time bias (probands had to be alive when they were recruited), smoking alcohol and BMI tokens times were randomised by -20 to +40 years from this point in time to break an otherwise confounding correlation leading to a sudden jump in mortality rates (and possibly other diseases with high mortality such as cancers) associated with the recording of these tokens. This is likely also to diminish the true effect of these tokens.

Age encoding. Delphi replaces GPT's positional encoding with an encoding based on the age values. Following the logic frequently used for positional encodings, age is represented by sine and cosine functions of different frequencies, where the lowest frequency is given by 1/(100*365), corresponding to a linear and constant basis function each. These functions are subsequently linearly combined by a trainable linear transformation, which enables the model to share the same basis function across multiple encoding dimensions. Another advantage of using age encoding is that Delphi can handle token inputs of arbitrary length, as no parameters are associated with token positions.

Causal self-attention. Standard causal self-attention enables the GPT model to attend to all preceding tokens. For sequential data, these are found to the left of the token sequence. Yet, in the case of time-dependent data, tokens can be recorded at the same time with no specified order. Thus, attention masks were amended to mask positions that occurred at the same time as the predicted token. Non-informative padding tokens were masked for predictions of other tokens.

Exponential waiting time model. The input data to Delphi are bivariate pairs (j, t) of the next token class and the time to the next token. Delphi is motivated by the theory of competing exponentials. Let T_i be the waiting times from the current event to one of $i = 1,...,n$ competing events, where n is the number of predictable tokens. Assuming the T_i are each exponentially distributed waiting times with rates $\exp(\lambda_i)$, The next event being j is equivalent to T_j being the first of the competing waiting times, ie $T_j = \min T_i$, or equivalently $j = \text{argmin } T_i$. It can be shown that the corresponding probability is $P(j = \text{argmin } T_i) = \exp(\lambda_j) / \Sigma_i \exp(\lambda_i)$, which is the softmax function over the vector λ. Conveniently, this definition corresponds to the classical cross-entropy model for classification with λ = logits. Hence Delphi uses a conventional loss term for token classification:

```
loss_j = - log p(j) = -cross_entropy(logits, tokens)
```

Furthermore, in the competing exponential model the time to the next event $T^* = \min\{T_j\}$ is also exponentially distributed with logarithmic rate $\lambda^* = \log(\Sigma_i \exp(\lambda_i))$, which is the logsumexp function of the logarithmic rates λ. The loss function of exponential waiting times T between tokens is simply given by:

```
loss_T = - log p(T) = -(logsumexp(logits) - sum(exp(logits)) * T).
```

These approximations hold as long as the log rates λ_i are constant in time, which is a reasonable assumption over short periods. For this reason, padding tokens were introduced to ensure that waiting times are modelled over a relatively short period, which does not

exceed 5 years in expectation. In line with the tie-braking logic used for causal self-attention, co-occurring events were predicted from the last non-co-occurring token each.

Loss function. The total loss of the model is then given by:

```
loss = loss_j + loss_T.
```

Non-informative padding, as well as sex, alcohol, smoking, and BMI were considered mere input tokens and thus removed from the loss terms above. This was achieved by setting their logits to -Inf and by evaluating the loss terms only on disease and “no event” padding tokens.

Sampling procedure. The next disease event is sampled by sampling the disease token and the time until the next event. The disease token is sampled from the distribution that originates from applying softmax to logits. For sampling the time, samples from all exponential distributions with rates exp(λi) are taken, and the minimum is taken. Logits of non-disease tokens (sex, lifestyle) are discarded from the procedure to sample disease events only.

Model training

Models were trained by stochastic gradient optimisation using the Adam optimiser with standard parameters for 200k iterations. The batch size was 128. After 1000 iterations of warmup, the learning rate was decayed using a cosine scheduler from 6e-4 to 6e-5. 32-bit float precision was used.

Model evaluation

Modelled incidence. In the exponential waiting time definition above, the logits of the model correspond to the log probabilities of daily occurrences of each token. The probability of an event occurring within a year is given by P(T < 365.25) = 1 - exp(exp(lambda) * 365.25).

Age- and sex-stratified incidence. For the training set age- and sex-statified incidences were calculated in annual age brackets. The observed counts were divided by the number of individuals at risk in each age and sex bracket, which was given by the number of probands for each sex minus the cumulative number of deaths to account for censoring.

Model calibration. Calibration curves were calculated based on predicted incidences. To this end, all cases of a given token accruing in five year age bins were identified. Subsequently for all other probands a control data point was randomly selected in the same age band. Predictions were evaluated at the preceding token given that the time difference was less than a year. The predicted incidences were then further grouped log-linearly into risk bins from 10^{-6} to 1 with multiplicative increments of $\log_{10}(5)$. The observed annual incidence was then calculated as the average of cases and control in age bins, divided by 5yr. The procedure was separately executed for each sex.

Relative risks. The relative risk of each prediction was calculated as the ratio of the modelled and yearly incidence to the age- and sex stratified incidence of the same token in the year in which the prediction was made. A summary was calculated as the geometric mean of each relative risk.

Incidence cross-entropy. To compare the distribution of annual incidences of model and observed data a cross-entropy metric was used. Let p_i be the annual occurrence of token i in each year. Hence the age- and sex-based entropy across tokens is is given $H(p,q) = – p * \log(q) – (1-p) * \log(1-q)$. For low incidences p,q the latter term is usually small. The cross-entropy is evaluated across all age groups and sexes.

Generated trajectories. To evaluate the potential of generating disease trajectories two experiments were conducted using data from the validation cohort. First, trajectories were generated from birth using only sex tokens. This was used to assess whether Delphi-2M recapitulates the overall sex-specific incidence patterns. Second, all available data until the age of 60 was used to simulate subsequent trajectories conditional on the previous health information. A single trajectory was evaluated per proband. Trajectories were truncated after the age of 80 as currently little training data was available. Incidence patterns were evaluated as described above.

Model validation

Study design. To validate the predictions of the model we also perform a longitudinal validation, internally for the UK Biobank data and externally on the Danish health registries. This has two advantages (i) we can enforce an explicit cutoff and separate data to avoid any potential time-leakage (ii) we obtain insights into Delphi-2M prognostic capabilities.
As mentioned in the data splits we use two different cutoff dates between the two data sources, mainly due to differential data availability, however, the principle set-up applies to both in the exact same way.
We collate data up to a specific cutoff data for each individual and use Delphi-2M to predict an individual's future rate across all disease tokens. Building on the exponential waiting time representation we obtain rates over a 1 year time frame. The preceding year after the cutoff date is discarded to introduce a data gap. Subsequently the incidence in the next year are used for evaluation.

Algorithms for comparison. As the task of predicting all disease outcomes simultaneously is novel there is no other method to compare against. Therefore, we build a standard epidemiological baseline based on the sex- and age-stratified population rates. These are based on the Nelson-Aalen estimator[35,36], a nonparametric estimator of the cumulative hazard rate, across all diseases. For the UK Biobank the estimators are based on the same training data as Delphi-2M and for the Danish registries we use the entire Danish population in the time period from 2010-2016.
As the UK Biobank contains a wide range of phenotypic measures we also estimate clinically established models for comparison. These include the Framingham score as implemented in R-CVrisk for cardiovascular disease, the Charlson comorbidity index for mortality implemented in R-comorbidity and the UKBDRS risk score for dementia based on our own implementation as reported in the original paper.

The Charlson comorbidity index is based on the same data as Delphi, however, the Charlson comorbidity index is originally based on ICD10-level 4 codes. Therefore, we estimate two versions, (i) that uses the level 3 codes and (ii) a version that maps the level 3 codes to all possible level 4 codes and we retain the version that shows the best performance. Additionally, we evaluate different versions of the comorbidity index, (i) original Charlson[28], (ii) Charlson with weights by Quan et al.[6] (iii) Elixhauser[37] and (iv) the Swiss Elixhauser[38] and again retain the best performing one.
The Framingham score is based on the 2008 version[7] with lab measurements. If covariates are missing, a simple mean imputation within sex has been used. If multiple measurements are available we retain the most recent one. For individuals older than 74, we use their Framingham score at age 74 as the current implementation does not allow for older age groups. Diabetes is classified as a diagnostic code in ICD10: E10-14.The assessed outcomes was cardiovascular disease (CVD) containing a diagnosis in either ischaemic heart disease (ICD10: I20-25), Heart failure or cardiac arrest (ICD10: I46, I50) or stroke (ICD10: I63, I64).
Last, the UKBDRS score is used to evaluate dementia as a diagnosis in (ICD10: F00, F01, F03, G30, G31). We build the risk score following the instructions in the paper as closely as possible. For years of education we impute values based on the reported mean value in the original paper (13.54). Again as we only retain 3rd level ICD10 codes minor adjustments have been made, precisely, depression (ICD10: F32, F33), stroke (ICD10: I60, I61, I63, I64, G45), high cholesterol (ICD10: E77-78) and hypertension (ICD10: I10-15 or hypertensive drug usage as constructed for the Framingham score).

Performance measures. To assess the discriminatory power of the predicted rates, we use the area under the receiver operating curve (ROC-AUC) as implemented in python-scikit-learn. All diseases with at least 25 cases were assessed.

Calibration of predicted rates. Further, we compare the predicted rates from Delphi-2M to the observed incidence to determine calibration of the predicted rates using python-scikit-learn. Delphi-2M predicted rates are split into deciles and for each bin we compare Delphi-2M's average rate against the observed rate within the bin. We include all diseases with at least 25 cases.

Model interpretation

Token embedding UMAP. The low-dimensional representation of token space was constructed by applying the UMAP[39] dimensionality reduction algorithm to the learned token embeddings for Delphi-2M (1270x120 matrix). The cosine metric was used.

SHAP. To evaluate the influence of each token in a trajectory on the next predicted token, we adopted the SHAP[29] methodology. Each trajectory from the validation cohort was augmented by masking one or several tokens and then used for prediction. The change of logits after many such augmentation were aggregated by a PartitionExplainer from the SHAP Python package.

Masking procedure. The number of augmentation for each trajectory was determined by the PartitionExplainer masking algorithm. When masked, tokens were replaced by a "no

event" placeholder that was also used during training. Sex tokens when masked were replaced with the corresponding token of the opposite sex.

SHAP values evaluation. The described procedure was applied to each of 100,636 trajectories in the validation cohort. The predicted token was always the last available token in the trajectory.

Nonparametric hazard ratios. To complement the SHAP analysis and the assessment of Delphi-2M's modelling of time-dependent effects we also performed an evaluation based on the Nelson-Aalen estimator. For a given token, we identify individuals with the token and estimate their corresponding cumulative hazard from the occurrence of the token onwards. Additionally we randomly select 5 age-sex matched individuals for each case and estimate the cumulative hazard in this comparison group. We can then obtain an estimate of the hazard rate by taking the derivative of the cumulative hazard. We apply a gaussian kernel to acquire a smooth estimate. Subsequently, we can take the ratio of the two hazards and obtain a crude nonparametric estimate for the hazard ratio of the token over time.

Code availability

Code for Delphi and accompanying jupyter notebooks are available on github.com/gerstung-lab/delphi. The model's checkpoint will be made available as per UK Biobank's controlled access procedures.

Data availability

UK Biobank data are available under restricted access through a procedure described at http://www.ukbiobank.ac.uk/using-the-resource/. Model weights will be made available through UK Biobank's controlled access system. Danish registry data are available for use in secure, dedicated environments via application to the Danish Patient Safety Authority and the Danish Health Data Authority via https://sundhedsdatastyrelsen.dk/da/english/health_data_and_registers/research_services/apply.

Acknowledgements

We acknowledge the following sources of funding: Novo Nordisk Foundation grant NNF17OC0027594 (K.G., A.W.J., E.B., M.G.), the Robert Bosch Foundation (M.G.) and from the EMBL European Bioinformatics Institute (EMBL-EBI) (E.B., T.F., A.W.J.). We thank Caroline Relton and George Davey Smith for detailed comments on the manuscript.

Ethics approval

The UK Biobank has received approval from the National Information Governance Board for Health and Social Care and the National Health Service North West Centre for Research Ethics Committee 532 (Ref: 11/NW/0382).

This research was conducted using the UK Biobank Resource under project 49978. All investigations were conducted in accordance with the tenets of the Declaration of Helsinki.

The use of the Danish National Patient Registry for validation of the UK Biobank results was conducted under the Danish Data Protection Act. Furthermore, the analysis was conducted under the information security and data confidentiality policies of Statistics Denmark, which is the Danish National Statistical Institution.

Author contributions

M.G. and E.B. conceived the study and supervised the analysis. M.G. developed the model with A.S. and K.G. K.G. prepared data sets and assessed the model parametrisation. A.S. assessed the model's predictive and generative performance as well as explainability. A.W.J. validated the model using longitudinal and external data. L.M. and S.B. prepared validation data. A.S., A.W.J., K.G., T.F., E.B. and M.G. wrote the manuscript. A.S. prepared all figures with contributions by A.W.J, K.G.,T.F. and M.G. All authors approved the manuscript.

Conflicts of interest

A preliminary patent has been filed for modeling time-dependent health data with generative transformers.

References

1. Nyberg, S. T. *et al.* Association of healthy lifestyle with years lived without major chronic diseases. *JAMA Intern. Med.* **180**, 760–768 (2020).
2. Link, B. G. & Phelan, J. Social conditions as fundamental causes of disease. *J. Health Soc. Behav.* **Spec No**, 80–94 (1995).
3. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
4. Mukama, T. *et al.* Risk-Adapted Starting Age of Screening for Relatives of Patients With Breast Cancer. *JAMA Oncol* **6**, 68–74 (2020).
5. Tian, Y. *et al.* Calculating the Starting Age for Screening in Relatives of Patients With Colorectal Cancer Based on Data From Large Nationwide Data Sets. *Gastroenterology* **159**, 159–168.e3 (2020).

6. Quan, H. *et al.* Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am. J. Epidemiol.* **173**, 676–682 (2011).
7. D'Agostino, R. B., Sr *et al.* General cardiovascular risk profile for use in primary care: the Framingham Heart Study: The Framingham heart study. *Circulation* **117**, 743–753 (2008).
8. Hippisley-Cox, J. & Coupland, C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. *BMJ Open* **5**, e007825 (2015).
9. Bray, F. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* (2024) doi:10.3322/caac.21834.
10. Watt, T. *et al.* Health in 2040: projected patterns of illness in England. *The Health Foundation* (2023).
11. Jung, A. W. & Gerstung, M. Bayesian Cox regression for large-scale inference with applications to electronic health records. *aoas* **17**, 1064–1085 (2023).
12. Vaswani, A. *et al.* Attention Is All You Need. *arXiv [cs.CL]* Preprint at http://arxiv.org/abs/1706.03762 (2017).
13. Brown, T. *et al.* Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
14. Gemini Team *et al.* Gemini: A Family of Highly Capable Multimodal Models. *arXiv [cs.CL]* (2023).
15. Touvron, H. *et al.* LLaMA: Open and Efficient Foundation Language Models. *arXiv [cs.CL]* (2023).
16. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *arXiv [cs.CL]* (2022).
17. OpenAI *et al.* GPT-4 Technical Report. *arXiv [cs.CL]* (2023).
18. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized

embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med* **4**, 86 (2021).

19. Li, Y. *et al.* BEHRT: Transformer for electronic health records. *Sci. Rep.* **10**, 7155 (2020).
20. Li, Y. *et al.* Hi-BEHRT: Hierarchical Transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE J. Biomed. Health Inform.* **27**, 1106–1117 (2023).
21. Yang, Z., Mitra, A., Liu, W., Berlowitz, D. & Yu, H. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nat. Commun.* **14**, 7857 (2023).
22. Placido, D. *et al.* A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nat. Med.* **29**, 1113–1122 (2023).
23. Savcisens, G. *et al.* Using sequences of life-events to predict human lives. *Nat. Comput. Sci.* 1–14 (2023).
24. Kraljevic, Z. *et al.* Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digit. Health* **6**, e281–e290 (2024).
25. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners. Preprint at https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.
26. Hoffmann, J. *et al.* Training Compute-Optimal Large Language Models. *arXiv [cs.CL]* (2022).
27. Anatürk, M. *et al.* Development and validation of a dementia risk score in the UK Biobank and Whitehall II cohorts. *BMJ Ment Health* **26**, (2023).
28. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J. Chronic Dis.* **40**, 373–383 (1987).
29. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).

30. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
31. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
32. Schmidt, M. *et al.* The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin. Epidemiol.* **7**, 449–490 (2015).
33. Helweg-Larsen, K. The Danish Register of Causes of Death. *Scand. J. Public Health* **39**, 26–29 (2011).
34. Pedersen, M. K. *et al.* A unidirectional mapping of ICD-8 to ICD-10 codes, for harmonized longitudinal analysis of diseases. *Eur. J. Epidemiol.* **38**, 1043–1052 (2023).
35. Aalen, O. Nonparametric Inference for a Family of Counting Processes. *Ann. Stat.* **6**, 701–726 (1978).
36. Nelson, W. Theory and applications of hazard plotting for censored failure data. *Technometrics* **14**, 945 (1972).
37. van Walraven, C., Austin, P. C., Jennings, A., Quan, H. & Forster, A. J. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. *Med. Care* **47**, 626–633 (2009).
38. Sharma, N., Schwendimann, R., Endrich, O., Ausserhofer, D. & Simon, M. Comparing Charlson and Elixhauser comorbidity indices with different weightings to predict in-hospital mortality: an analysis of national inpatient data. *BMC Health Serv. Res.* **21**, 13 (2021).
39. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).