Article

# A foundation model for continuous glucose monitoring data

https://doi.org/10.1038/s41586-025-09925-9

Received: 20 August 2024

Accepted: 17 November 2025

Published online: 14 January 2026

Check for updates

Guy Lutsker[1,2,3], Gal Sapir[1,4], Smadar Shilo[1,2,5,6], Jordi Merino[7,8], Anastasia Godneva[1,2], Jerry R. Greenfield[9,10,11], Dorit Samocha-Bonet[9,10], Raja Dhir[12], Francisco Gude[13,14], Shie Mannor[3], Eli Meirom[3], Eric P. Xing[15,16], Gal Chechik[3], Hagai Rossman[4,16] ✉ & Eran Segal[1,16] ✉

Continuous glucose monitoring (CGM) generates detailed temporal profiles of glucose dynamics, but its full potential for achieving glucose homeostasis and predicting long-term outcomes remains underutilized. Here we present GluFormer, a generative foundation model for CGM data trained with self-supervised learning on more than 10 million glucose measurements from 10,812 adults mainly without diabetes[1,2]. Using autoregressive prediction, the model learned representations that transferred across 19 external cohorts ($n$ = 6,044) spanning 5 countries, 8 CGM devices and diverse pathophysiological states, including prediabetes, type 1 and type 2 diabetes, gestational diabetes and obesity. These representations provided consistent improvements over baseline blood glucose and HbA1c levels and other CGM-derived measures for forecasting glycaemic parameters[3,4]. In individuals with prediabetes, GluFormer stratified those likely to experience clinically significant increases in HbA1c over a 2-year period, outperforming baseline HbA1c and common CGM metrics. In a cohort of 580 adults with short-term CGM and a median follow-up of 11 years[5], GluFormer identified individuals at elevated risk of diabetes and cardiovascular mortality more effectively than HbA1c. Specifically, 66% of incident diabetes cases and 69% of cardiovascular deaths occurred in the top risk quartile, compared with 7% and 0%, respectively, in the bottom quartile. In clinical trials, baseline CGM representations improved outcome prediction. A multimodal extension of the model that integrates dietary data generated plausible glucose trajectories and predicted individual glycaemic responses to food. Together, these findings indicate that GluFormer provides a generalizable framework for encoding glycaemic patterns and may inform precision medicine approaches for metabolic health.

The emergence of self-supervised learning (SSL) in healthcare marks a shift in medical artificial intelligence (AI) and has enabled the development of foundation models that learn from vast unlabelled datasets and support diverse downstream tasks[6]. Examples include foundation models for retinal imaging[7,8], wearables[9], sleep analysis[10] and pathology[11], all of which demonstrate how SSL can capture complex biomedical signals to improve diagnosis and risk prediction. This convergence of SSL and large-scale biomedical data offers new opportunities for chronic diseases, for which continuous monitoring can guide prevention and treatment[12]. Diabetes exemplifies this need: it currently affects more than 500 million people worldwide, cases are projected to rise to 1.3 billion by 2050 and associated annual global costs could exceed US$900 billion[13]. Type 2 diabetes, the most common form, is largely driven by modifiable behaviours, doubles cardiovascular risk and contributes to multiple comorbidities[14]. Yet predicting early glycaemic dysregulation and variable therapeutic responses remain a central challenge.

CGM provides high-resolution, real-time data on glucose dynamics and is now a cornerstone of diabetes care. Compared with self-monitoring of blood glucose, CGM improves glucose control in both adults and children[15,16], reduces hypoglycaemia and enhances quality of life[15,17]. A recent consensus statement by the American Diabetes Association and the European Association for the Study of Diabetes recommended that CGM-derived metrics should be incorporated

[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [2]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [3]NVIDIA, Tel Aviv, Israel. [4]Pheno.AI, Tel-Aviv, Israel. [5]Faculty of Medical and Health Sciences, Tel Aviv University, Tel-Aviv, Israel. [6]The Jesse and Sara Lea Shafer Institute for Endocrinology and Diabetes, National Center for Childhood Diabetes, Schneider Children's Medical Center of Israel, Petah Tikva, Israel. [7]Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark. [8]Diabetes Unit, Endocrine Division, Massachusetts General Hospital, Boston, MA, USA. [9]Clinical Diabetes, Appetite and Metabolism Laboratory, Garvan Institute of Medical Research, Sydney, New South Wales, Australia. [10]St Vincent's Clinical Campus, School of Clinical Medicine, University of NSW, Sydney, New South Wales, Australia. [11]Department of Endocrinology and Diabetes, St Vincent's Hospital, Sydney, New South Wales, Australia. [12]Swiss Institute of Allergy and Asthma Research (SIAF), University of Zurich, Davos, Switzerland. [13]Department of Medicine, University of Santiago de Compostela, Santiago de Compostela, Spain. [14]Concepción Arenal Primary Care Center, Santiago de Compostela, Spain. [15]Carnegie Mellon University Pittsburgh, Pittsburgh, PA, USA. [16]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. ✉e-mail: hagai.rossman@mbzuai.ac.ae; eran.segal@weizmann.ac.il

into diabetes clinical trials[16]. Moreover, CGM use is expanding beyond diabetes for detecting early dysglycaemia, supporting athletic performance optimization and guiding dietary advice[18,19]. Day-to-day fasting glucose variability observed in adults without diabetes[20] indicates the potential of CGM to refine the assessment of glycaemic status compared with traditional diagnostic criteria. Furthermore, approval by the US Food and Drug Administration of the first over-the-counter CGM device[21] highlights its growing adoption and broad potential applications.

Here we introduce GluFormer, a generative transformer-based foundation model trained in a self-supervised manner on >10 million CGM readings from 10,812 participants in the Human Phenotype Project (HPP)[1,2]. GluFormer learns latent representations of glycaemic patterns and can both generate physiologically plausible glucose trajectories and predict health outcomes. We evaluate its generalizability across 19 external cohorts spanning 5 countries, 8 CGM devices and diverse metabolic states, including prediabetes, type 1 and type 2 diabetes, gestational diabetes and obesity. Despite differences in the training set, GluFormer demonstrates strong out-of-distribution performance in predicting both contemporaneous and long-term glycaemic and cardiometabolic outcomes, including in a cohort with a 11-year follow-up. We demonstrate that GluFormer-derived embeddings provide additional discriminatory capabilities compared with standard CGM-derived metrics such as the glucose management indicator (GMI) or time in range when forecasting deterioration in glycaemic parameters and cardiovascular risk. These results provide support for the potential utility of CGM-based foundation models in risk stratification and personalized treatment planning.

## A CGM foundation model for glycaemic dynamics

GluFormer was pretrained on >10 million glucose measurements from 10,812 participants in the HPP cohort (Supplementary Figs. 11–16). CGM values were tokenized, and the model was trained autoregressively for next-token prediction to enable both continuation of glucose time series and generation of new trajectories (Fig. 1a). The pretrained model also produces latent embeddings that can be used for downstream tasks (Fig. 1b–d).

To assess whether GluFormer representations capture physiologically relevant information, we projected embeddings of HPP recordings using uniform manifold approximation and projection (UMAP)[22]. Colour coding on the basis of fasting plasma glucose or postprandial glucose response revealed clear gradients that corresponded to glycaemic status, which indicated that the latent space may encode meaningful physiological variation. Cosine distance analyses further showed that intra-individual embeddings were significantly more similar than inter-individual embeddings (Mann–Whitney, $P < 0.001$), which suggested that the model captures individual-specific glycaemic signatures.

We next benchmarked GluFormer against conventional models and metrics for predicting HbA1c, a key marker of long-term glycaemic control[3,4]. Transformer representations outperformed convolutional neural networks and multilayer perceptrons (MLPs). Moreover, it showed higher predictive capabilities than GMI, a metric proposed to estimate HbA1c from CGM data[23], and CGM-derived metrics often used in clinical trials, such as mean glucose, time in range and glycaemic variability indices[24,25]. Pretraining with SSL substantially improved accuracy, and frozen embeddings alone provided more informative features than traditional metrics (Supplementary Fig. 2d). These findings indicate that GluFormer learns rich representations of glycaemic physiology beyond established CGM-derived measures.

We next evaluated the ability of GluFormer to generate physiologically plausible CGM traces, an important application of generative models in healthcare, in which synthetic data can help address scarcity and imbalance[26]. For each individual, the model generated CGM trajectories of the last available 24 h based on all previous contexts, which were compared with observed recordings using both visual inspection and quantitative metrics (Methods). To ensure robustness, we averaged three generated series per participant using different random seeds and compared their metrics to the observed curves. The generated series closely matched observed glucose profiles across multiple composite scores, including mean glucose, coefficient of variation (c.v.), GMI, time in range and hypoglycaemia. However, some discrepancies reflected unmodelled personal factors, such as day-to-day dietary changes (Fig. 2a). Pearson's correlations were high across key parameters ($r = 0.98$ for mean glucose, $r = 0.98$ for GMI, $r = 0.89$ for per cent time <70 mg dl$^{-1}$, all $P < 0.001$), which indicated that the model can accurately reproduce clinically relevant features (Fig. 2b). Model performance was improved when longer input histories were used. That is, extending the input time window from 0 to 10 days increased agreement between generated and observed CGMs, with average Pearson's correlations increasing from 0.46 to 0.90 ($P < 0.001$; Supplementary Figs. 3 and 4). This result demonstrates that providing longer sequences of past CGM readings substantially enhances generative accuracy.

Generative fidelity was generalized across diverse studies and individual characteristics, including individuals with obesity, type 1 or 2 diabetes or gestational diabetes and in healthy populations from Israel, Australia, Asia and North America. In these external datasets, generated trajectories consistently correlated with observed values (Pearson's $r > 0.8$, $P < 0.001$). Variability metrics (for example, standard deviation (s.d.) and c.v.) were more challenging to reproduce in some cohorts, a result that underscores the influence of short-term lifestyle factors (Fig. 2c). Collectively, these results show that GluFormer captures both individual-specific and population-level glucose dynamics and produces synthetic series that reflect real physiological processes while enabling embeddings suitable for downstream prediction.

## GluFormer-derived risk score and long-term outcomes

Although the embeddings of GluFormer capture rich, high-dimensional information from CGM, the potential direct use of CGM in clinical risk stratification is less evident. We therefore generated the GluFormer-derived score, which was trained to predict HbA1c trajectories from embeddings to provide a quantitative measure that underlies different glycaemic alterations axes. For this metric, higher scores represent higher glycaemic deviations (Methods).

We evaluated the predictive capabilities of the GluFormer-derived score in 337 HPP participants with prediabetes (baseline HbA1c percentage of 5.7–6.4%). Individuals were stratified into quartiles on the basis of either the GluFormer-derived score or measured HbA1c and followed for 2 years (Fig. 3a). Participants in the top score quartile showed a mean HbA1c increase of +0.18%, whereas those in the bottom quartile decreased by −0.13% ($P < 0.001$, Mann–Whitney). By contrast, quartiles defined by baseline HbA1c did not show significant separation in future HbA1c trajectories. The observed differences between the top and bottom score quartiles may have clinical implications, although further validation in independent studies will be needed.

We next examined whether the score is associated with long-term outcomes in the A Estrada glycation and inflammation study (AEGIS), a longitudinal epidemiological study investigating the links between dysglycaemia, inflammation and cardiometabolic risk in a general population sample in Spain[5]. The cohort included 580 adults with baseline CGM and a median follow-up of 11 years. Participants ranked on the basis of the GluFormer-derived score showed separation in outcomes. Specifically, individuals in the top quartile had a higher incidence of diabetes (log-rank $P = 2.3 \times 10^{-6}$), with the model capturing 65.8% of new-onset cases, whereas stratification by measured HbA1c did not distinguish risk ($P = 0.71$). A similar pattern was observed for

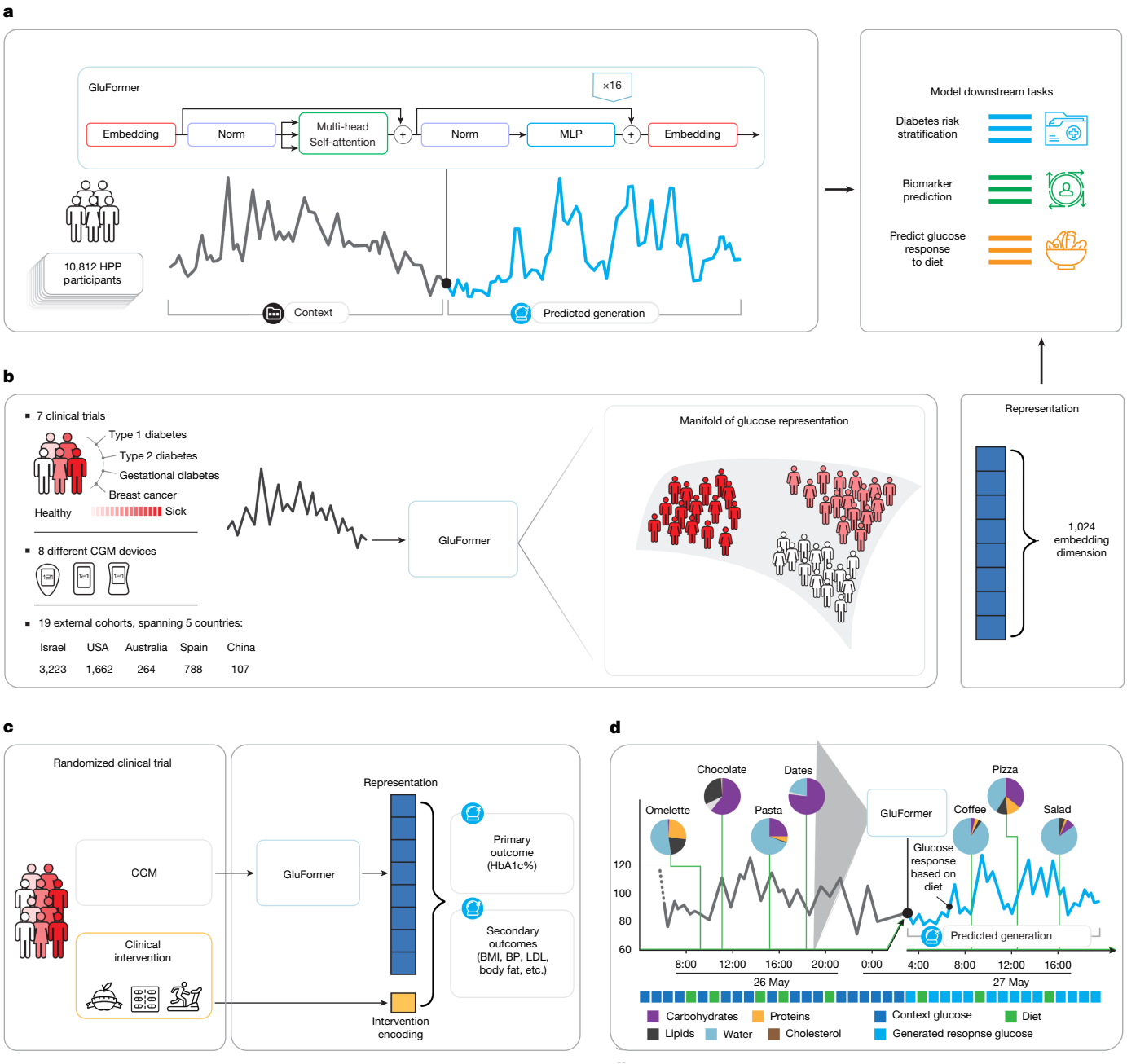


**Fig. 1 | Overview of GluFormer architecture, training pipeline and downstream tasks. a**, We pretrained GluFormer on CGM data from 10,812 individuals in the HPP cohort with the objective of predicting subsequent glucose measurements (next-token prediction). We evaluated its utility on downstream tasks, including generating CGM time series, predicting clinical measures and creating embeddings usable by linear models for medical outcomes. Top row (left), transformer attributes: normalization layer (Norm), multi-head attention layer (Multi-head) and self-attention layer (Self-attention). **b**, Generalization of clinical measure prediction was tested on 19 external datasets, including 7 clinical trials across pathologies (type 1 diabetes, type 2 diabetes, gestational diabetes, among others), using 8 CGM devices from 5 countries. Numbers below each country indicate participant counts. **c**, We forecasted clinical trial outcomes using pre-intervention CGM representations. **d**, Multimodal GluFormer with added date and time inputs is capable of processing both glucose and dietary tokens. BMI, body–mass index; BP, blood pressure; LDL, low-density lipoprotein cholesterol. Schematic in **a** was adapted from ref. 2, Springer Nature America.

cardiovascular-related mortality, with 69.2% of deaths occurring in the top quartile compared with none in the bottom quartile ($P = 0.001$). Measured HbA1c was not significantly associated ($P = 0.25$) (Fig. 3b).

These results suggest that the GluFormer-derived score captures prognostic information from dynamic CGM data that is not captured by baseline HbA1c. In this cohort, the score stratified long-term risk of diabetes and cardiovascular complications, whereas baseline HbA1c did not show significant associations. This finding points to the potential of dynamic CGM signals to provide clinically relevant information beyond conventional glycaemic measures, although validation in larger and more diverse populations will be important.

## Predicting clinical measures across populations

In addition to the changes in glycaemic parameters among people with prediabetes, we investigated the ability of GluFormer to predict glycaemic changes in the HPP cohort, which is mainly composed of individuals with apparent normoglycaemia. Using ridge regression,

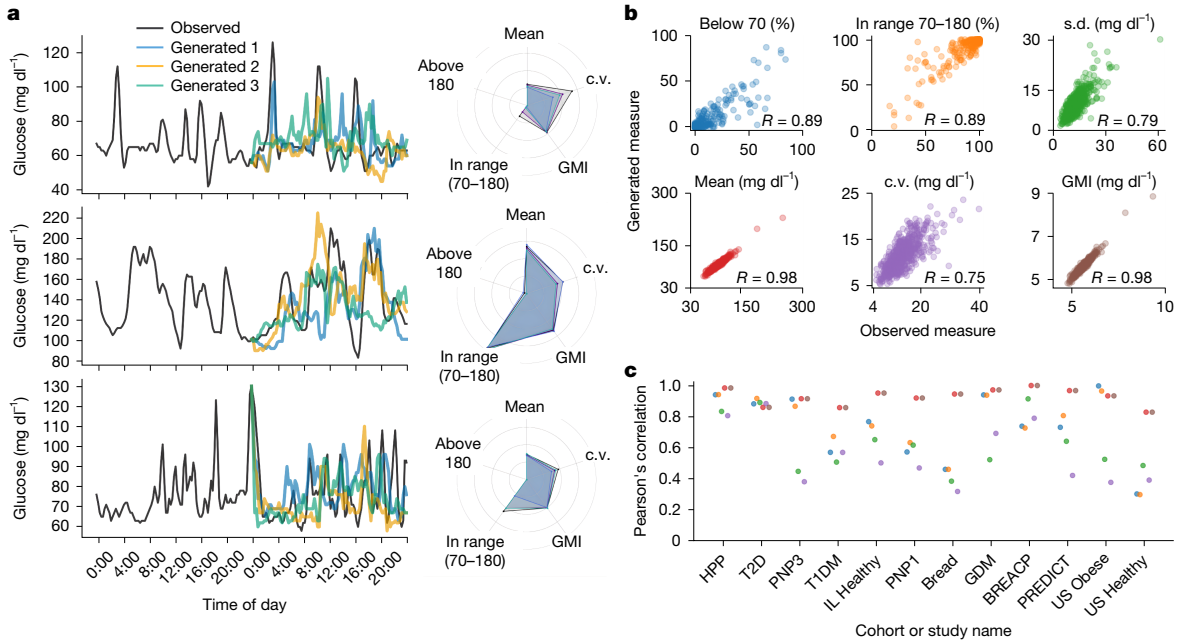


**Fig. 2 | Evaluation of the capabilities of GluFormer in simulating and analysing CGM data. a**, Day-by-day analysis for three participants from the HPP test set. Left, comparisons of observed CGM (black) with three predicted series generated by GluFormer using different seeds (blue, orange and green). The first day of CGM data provided context for predicting the second day, which was compared with true profiles. Right, radar charts of composite scores (mean, c.v., GMI, time in range (70–180) and above 180) for observed and generated data. **b**, Scatter plots of composite scores (iglu) from observed and generated series across HPP participants. For each individual, three synthetic days were generated (excluding the final day for evaluation) and scores were averaged. Each point represents 1 day with the following metrics: below 70, in range 70–180, s.d., mean, c.v. and GMI. Pearson's correlations exceeded 0.75 ($P < 0.001$) across all metrics, thereby validating the ability of GluFormer to reproduce clinically relevant CGM measures. **c**, Correlations of composite scores between observed and generated CGMs across external cohorts (for details of the cohorts used, see Supplementary Table 1). Generations used all available CGM days except the last for evaluation. Metrics are identical to those in **b**. For a more detailed explanation of the analyses, see 'Analysis details' in the Methods.

we compared GluFormer embeddings to two established CGM-derived baselines: the GMI, and composite scores computed using the R package iglu[24,25]. Predictions were focused on HbA1c and fasting glucose at the time of CGM recording and at 2-year and 4-year horizons after baseline measurements. Even among adults without diabetes, CGM has been shown to correlate with broader metabolic traits[27].

At baseline (Fig. 4a), GluFormer showed higher predictive performance, albeit modest, for HbA1c ($r = 0.43$ versus 0.39 for iglu and 0.35 for GMI) and fasting glucose ($r = 0.48$ versus 0.40 for iglu and 0.34 for GMI). At 2 years, the model maintained stronger correlations for both HbA1c ($r = 0.44$ versus 0.40 for iglu and 0.34 for GMI) and fasting glucose ($r = 0.52$ versus 0.37 for iglu and 0.33 for GMI, $P < 0.001$ for fasting glucose comparisons). At 4 years, GluFormer continued to outperform baselines for HbA1c ($r = 0.25$ versus 0.23 for iglu and 0.18 for GMI) and fasting glucose ($r = 0.43$ versus 0.32 for iglu and 0.28 for GMI, $P < 0.001$ for fasting glucose comparisons), whereas overall correlation strength decreased with longer horizons (Fig. 4, right). Although improvements were consistent across all end points, the clinical significance of this advantage may be limited, as noted by the small reductions in prediction error (root mean squared error (RMSE) and mean absolute error (MAE); Supplementary Table 6).

Beyond continuous measures, we assessed discrimination of future diabetes development at these same time points. At baseline, GluFormer achieved a higher receiver operating characteristic area under the curve (ROC AUC) value (0.75) than CGM composite scores (0.69) or GMI (0.66), and decision curve analysis demonstrated enhanced net benefit across a broad range of threshold probabilities (0.2–0.8)[24,25,28,29] (Fig. 4, left and middle). This pattern of enhanced discriminative performance and clinical utility persisted at both 2-year and 4-year horizons, a result that suggests that GluFormer may provide added value for near-term and longer-term risk assessment.

To assess generalizability, we applied GluFormer to additional external cohorts spanning different countries, CGM devices and disease states. Across these populations, embeddings showed significant, although generally modest, associations with disease-relevant measures (Supplementary Fig. 25). For example, in individuals with type 2 diabetes, GluFormer embeddings correlated with creatinine ($r = 0.27$, $P < 0.001$), an indicator of kidney function. In patients who recovered from breast cancer, embeddings correlated with albumin ($r = 0.19$, $P < 0.001$) and creatinine ($r = 0.12$, $P < 0.001$). Among pregnant women with gestational diabetes, correlations were strong, including with haemoglobin ($r = 0.42$, $P < 0.001$) and platelet counts ($r = 0.35$, $P < 0.001$).

Taken together, these findings indicate that GluFormer embeddings are associated with clinically relevant measures both at baseline and up to 4 years in the future, with modest but consistent improvements over existing CGM-derived metrics. The robustness of these associations across geographies, CGM devices and disease states raises the possibility that the latent space of GluFormer encodes information beyond traditional glycaemic metrics, which may reflect broader aspects of metabolic health status.

## Forecasting trial outcomes

Predicting who is more likely to respond to a given therapeutic intervention is one of the main challenges facing current evidence-based medicine approaches. To explore whether CGM-derived embeddings might overcome such obstacles, we used baseline CGM data from several completed studies to generate GluFormer representations and compared their performance with GMI and other CGM-derived metrics.

Across multiple cohorts, GluFormer representations were associated with modest improvements in forecasting trial outcomes, which

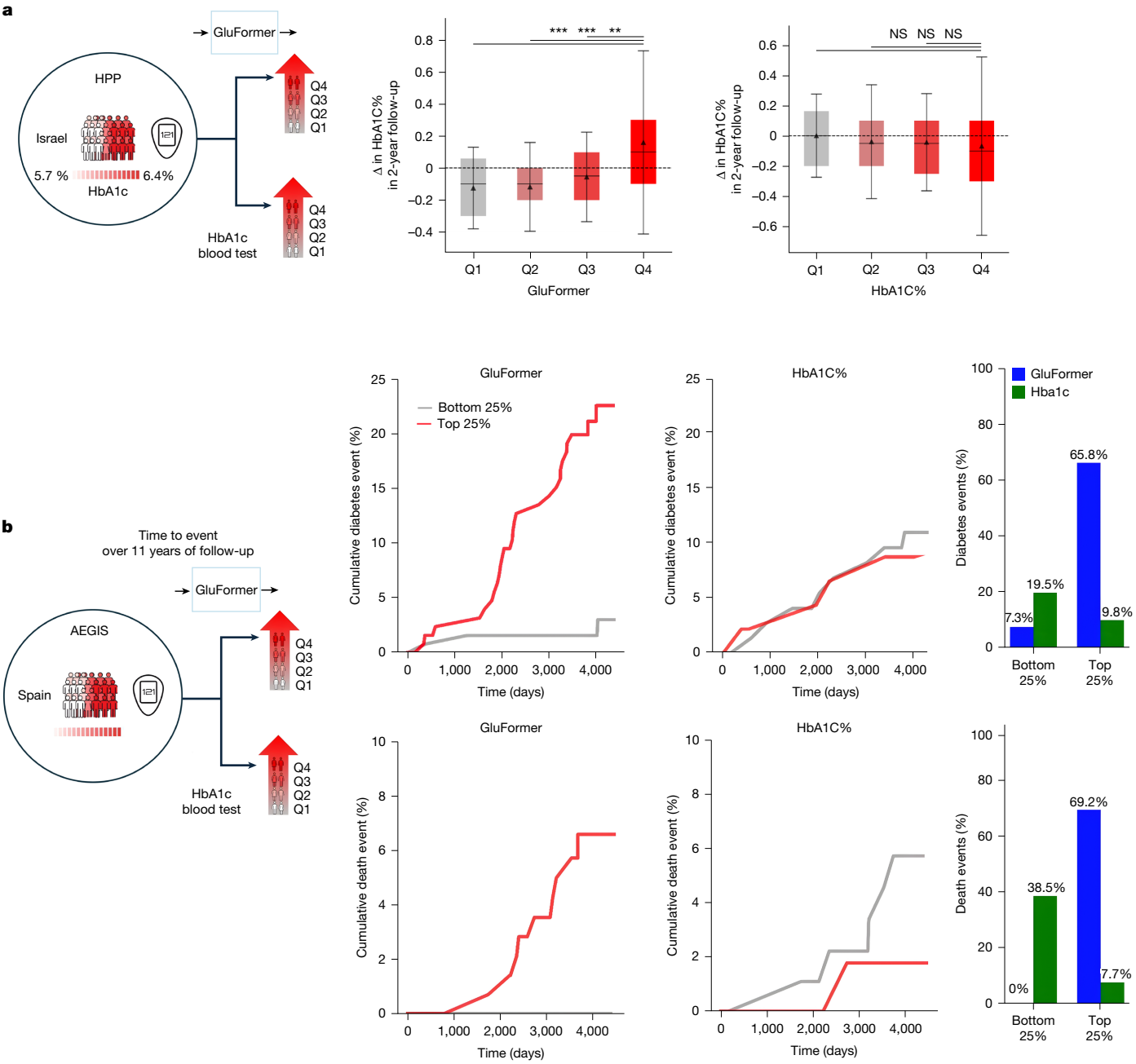


**Fig. 3 | GluFormer-derived score outperforms measured HbA1c for stratifying risk of glycaemic progression and long-term outcomes.** **a**, A prediabetes cohort ($n = 337$; baseline HbA1c of 5.7–6.4%) from the HPP was stratified into quartiles on the basis of their GluFormer-derived score (Q1: 5.17–5.37; Q2: 5.37–5.46; Q3: 5.46–5.60; and Q4: 5.60–6.56) or baseline HbA1c (Q1: 5.70–5.73; Q2: 5.73–5.85; Q3: 5.85–6.00; and Q4: 6.00–6.40). GluFormer quartiles showed clear separation in 2-year HbA1c change, whereby the top quartile increased by +0.18% and the bottom quartile decreased by −0.13% (***$P < 0.001$, two-sided Mann–Whitney $U$-test). Baseline HbA1c quartiles did not show significant differences (NS). Box plots show the median (line), mean (triangle), interquartile range (box) and mean ± 1 s.d. (error bars). **b**, An independent cohort was also analysed: AEGIS cohort ($n = 580$, median follow-up of 11 years). Top, for diabetes incidence (total of 41), the top 25% by GluFormer-derived score (Q1: 4.86–5.22; Q4: 5.62–8.99) had significantly shorter diabetes-free survival than the bottom 25% ($P = 2.3 \times 10^{-6}$, log-rank test). Overall, 65.8% of cases occurred in the top quartile versus 7.3% in the bottom. Stratification by measured HbA1c (Q1: 3.9–5.2; Q4: 5.8–9.9) showed no separation ($P = 0.71$). Bottom, for cardiovascular mortality (total of 13), 69.2% of deaths occurred in the GluFormer top quartile versus 0% in the bottom ($P = 0.001$, log-rank test). Measured HbA1c quartiles were not significant ($P = 0.25$).

were defined as changes in health parameters following intervention (Supplementary Fig. 26). In the PREDICT cohort[30], embeddings showed higher predictive power than GMI for HbA1c, creatinine and waist circumference ($P < 0.001$). In the BREACP study[31], GluFormer improved predictions for HbA1c, body-fat percentage, lymphocyte counts and creatinine ($P < 0.001$). In the PNP3 diet intervention study[32], embeddings improved prediction of changes in HbA1c, low-density lipoprotein cholesterol and glucose levels ($P < 0.001$). These predictions relied only on pre-intervention CGM data and a binary variable indicating the group who received the intervention.

We further observed improved performance relative to GMI when applied to primary outcomes in additional publicly available clinical trials with CGM data (Supplementary Fig. 27). Although effect sizes were variable and generally modest, these results suggest that CGM-derived

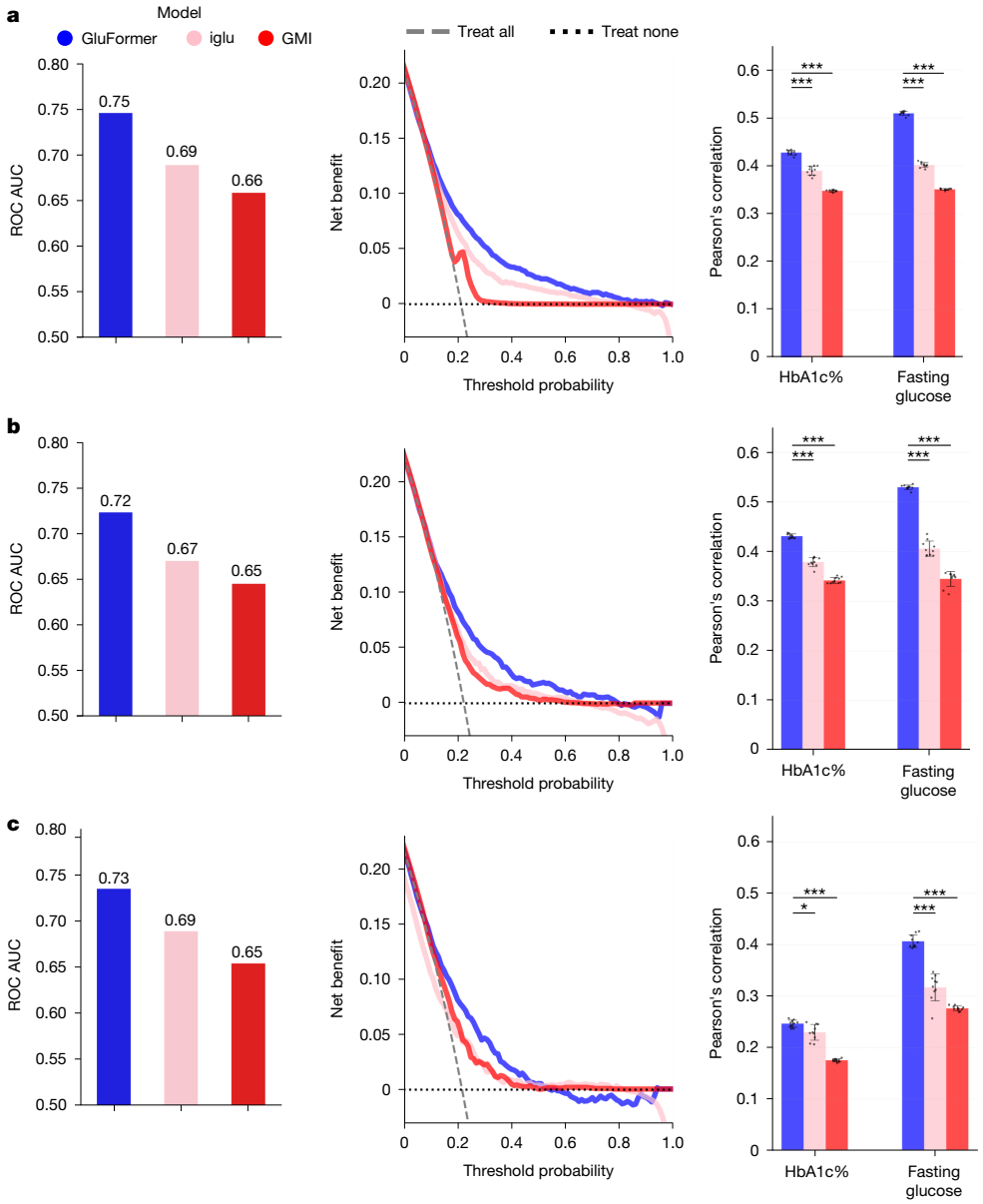


**Fig. 4 | Predictive performance of clinical measures using GluFormer representations versus CGM-derived composite scores and GMI.** Predictions at baseline (at CGM collection) (**a**) and at 2-year (**b**) and 4-year (**c**) horizons from features computed at the corresponding time points. Left, ROC AUC for diabetes prediction for the three horizons. GluFormer consistently exceeds CGM-derived composite scores (iglu) and GMI, which indicated that it has better discrimination. Middle, decision curve analysis (net benefit versus threshold probability). Treat none and treat all strategies are shown. Across horizons, GluFormer generally delivers higher net benefit than iglu and GMI. Right, Pearson's correlations for clinical measures (HbA1c percentage (HbA1c%) and fasting glucose) using ridge regression with GluFormer, iglu and GMI. Each point is the mean of ten runs (different seeds). Error bars show the s.d. Significance was assessed using two-sided Whitney *U*-tests with Benjamini–Hochberg false discovery rate: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. GluFormer outperformed comparators across **a** and **b**, whereas at 4 years (**c**), GluFormer exceeded comparators for all measures except HbA1c.

embeddings may capture aspects of baseline metabolic states relevant to intervention responses. Further validation in prospective trials will be required to determine their potential utility in supporting personalized treatment decisions and trial design.

## Extending to multimodal diet–glucose modelling

We next investigated whether the addition of dietary information could improve the capabilities of GluFormer in modelling glucose responses. We developed a multimodal version of the model that integrates macronutrient content alongside CGM values, tokenizing both glucose and diet events into a synchronized sequence. Training used a next-token prediction strategy, with diet tokens masked from the loss function so that the model learned to predict subsequent glucose values conditioned on both dietary and glycaemic context (Methods).

We compared models trained with and without diet tokens on a CGM-generation task. The incorporation of dietary data improved agreement with observed glucose measurements, with the multimodal model achieving an average correlation of 0.50 compared with 0.22 for the glucose-only model ($P < 0.001$). Across participants, 91% showed improved correlations (Fig. 5a, top) and 92% showed reduced MAE (Fig. 5a, bottom). Visual inspection of generated trajectories confirmed closer alignment with observed CGM curves when dietary data were included, particularly around meal-related glucose excursions (Fig. 5). Robustness was supported by consistent results across multiple random seeds (Supplementary Fig. 8).

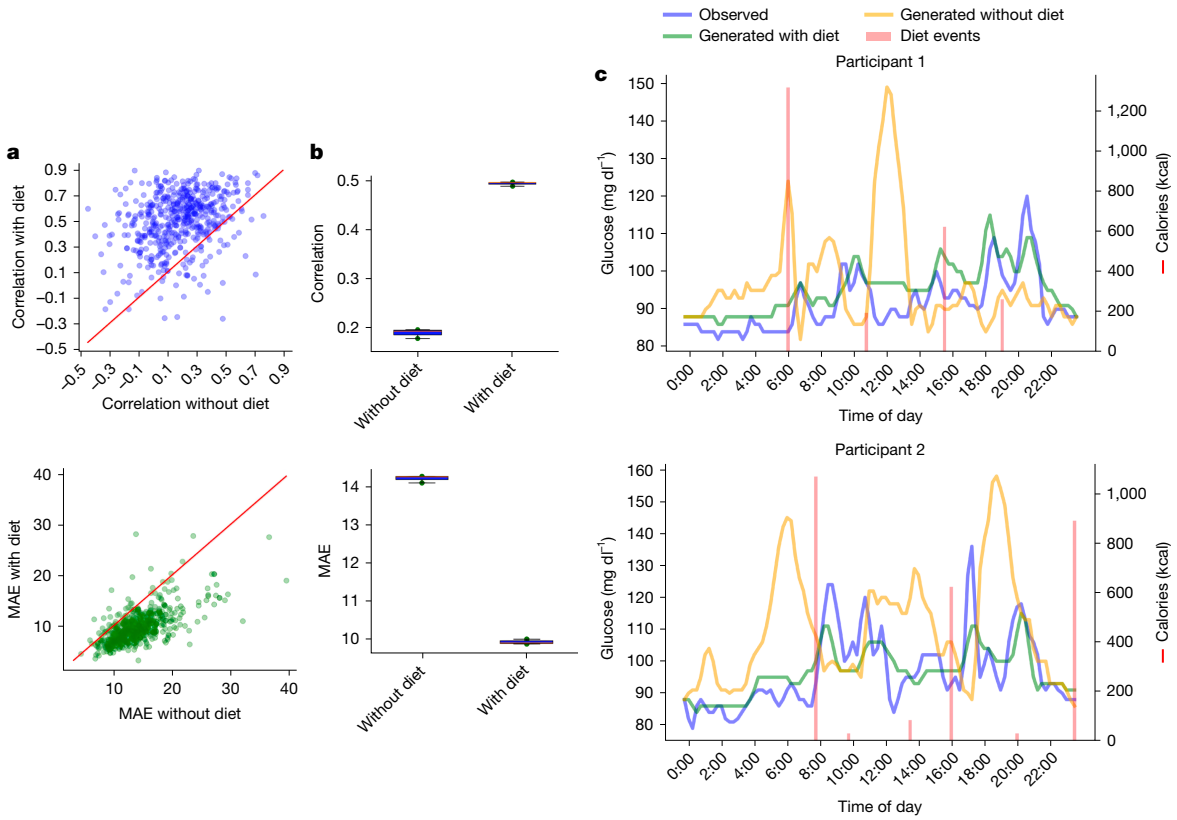


**Fig. 5 | Impact of dietary data on GluFormer model performance.** **a**, Comparison of Pearson's correlation (top) and MAE (bottom) between the original and generated CGM data with and without the inclusion of dietary data. Scatter plots show the improvements in correlation (91%) and MAE (92%) when dietary data are included, as indicated by the majority of points falling above the diagonal line on correlation and below the line on MAE metrics. **b**, Box plots summarizing the overall performance, showing the average correlation (top) and MAE (bottom) across all test participants for five different random seeds (used for generation), with lower MAE and higher correlation for models that included dietary data. In the box plots, the blue box represents the interquartile range of the correlation scores, and whiskers define a single s.d. **c**, Time-series plots demonstrating glucose-level predictions for two example participants. The observed CGM data (blue line) are compared with data generated with dietary tokens (green line) and without dietary tokens (orange line). Red bars indicate times of dietary events, highlighting the improved performance of the model in capturing glucose spikes when dietary information is included.

These results demonstrate that inclusion of dietary data enhances model performance and enables more accurate prediction of postprandial responses. Although exploratory, this approach highlights the potential of multimodal foundation models to incorporate behavioural and nutritional context alongside physiological signals.

## Discussion

This study introduces GluFormer, a foundation model trained on >10 million CGM measurements from 10,812 adults predominantly without diabetes that learns latent glycaemic patterns and supports forecasting tasks. GluFormer showed evidence of generalization across 19 external cohorts that encompass diverse ethnicities, age groups and pathophysiological states.

Compared with standard CGM-derived metrics, GluFormer demonstrated modest, although significant, improvements in predictive accuracy for key clinical outcomes and a greater net benefit across decision thresholds in our analyses. This advantage was most evident for long-term risk stratification. Specifically, in an independent cohort followed up for 11 years, GluFormer identified individuals at increased risk of diabetes and cardiovascular mortality more effectively than HbA1c, capturing 66% of incident diabetes cases and 69% of cardiovascular deaths in the top risk quartile compared with only 7% and 0%, respectively, in the bottom quartile. More generally, pretraining enhanced the predictive capabilities of the model (Supplementary Fig. 2d) and produced improved predictions for various clinical measures relative to traditional CGM-based statistics. The largest gains emerged for broader metabolic outcomes, whereas improvements for glycaemic markers such as HbA1c were more modest. This result may be expected, as conventional CGM-derived metrics are specifically designed to extract glycaemic information. This finding indicates that GluFormer-based representations capture broader aspects of metabolic and cardiovascular risk not fully reflected in existing markers, thereby warranting validation in additional cohorts.

Risk stratification by GluFormer may identify individuals at risk of metabolic deterioration before clinical manifestations appear and therefore enable preventive intervention. This result aligns with findings from landmark studies such as the Diabetes Prevention Program[33], in which early lifestyle modifications reduced diabetes incidence by 58% in high-risk populations. The ability of GluFormer to distinguish future glycaemic trajectories in individuals with prediabetes offers potential clinical value, as this population represents a critical window for intervention. Furthermore, the capacity of the model to simultaneously assess risks for both diabetes and related cardiovascular outcomes could help clinicians develop more comprehensive prevention and care strategies. The performance across different clinical decision thresholds, demonstrated through decision curve analysis, suggests that GluFormer may have practical utility in diverse healthcare settings in which the balance between true positives and false positives must be calibrated according to local resources and intervention capabilities.

Notably, although the GluFormer-derived score significantly stratified long-term risk for diabetes and cardiovascular death, it was only modestly associated with 2-year changes in HbA1c. A potential explanation for this apparent discrepancy is the fact that HbA1c measurement is noisy and influenced by assay variability[34]. It is also possible that individuals alter their lifestyle habits or initiate medications in response to elevated HbA1c values, which could weaken its ability to capture future risk trajectories. To provide a balanced perspective, it is also important to acknowledge the potential limitations of CGM technology. Many devices are calibrated for the wider glycaemic ranges seen in individuals with diabetes, and their application in healthier populations, as in our primary training cohort, may therefore introduce a distribution shift in measurement accuracy. Although CGM data may also be prone to this distribution shift, the GluFormer score is derived from 2 weeks of continuous data, which may enable it to capture more stable, dynamic patterns of glycaemic dysregulation. Taken together, these observations suggest that the model may capture risk-related glycaemic patterns that relate more closely to long-term outcomes than variation in a single biomarker.

Analysis of the latent space of GluFormer revealed how SSL learning can capture biologically meaningful patterns in physiological time-series data. Just as language model embeddings organize semantic relationships, the representations of GluFormer seem to separate distinct aspects of glucose metabolism. That is, it was able to distinguish between postprandial responses (linked mainly to β-cell dysfunction and peripheral (muscle) insulin resistance[35]) and fasting glucose levels (associated mainly with hepatic insulin resistance[36]). This unsupervised organization of metabolic features is consistent with the clinical understanding of diabetes progression, in which postprandial hyperglycaemia typically precedes fasting hyperglycaemia. The ability of the model to encode these temporal dynamics without explicit supervision suggests that it has potential applications in monitoring metabolic dysfunction and disease progression. However, dedicated studies will be needed to link latent dimensions to specific physiological mechanisms.

The choice of next-token prediction as our pretraining strategy proved to be highly effective, compelling the model to leverage past data to predict future glucose measurements and learn complex temporal dependencies. This self-supervised approach not only enables the generation of realistic glucose signals but also supports diverse downstream clinical tasks. Building on insights that discretizing continuous time series into categorical tokens can enhance neural network performance[37–40], GluFormer tokenizes glucose measurements to effectively capture temporal patterns. Extending this framework to a multimodal design that integrates both glucose and nutrient tokens further improved prediction accuracy, particularly around meal times, thereby highlighting the potential of combining temporal and contextual information in biomedical sequence modelling.

Currently, clinicians typically rely on limited metrics such as fasting glucose and HbA1c to assess glycaemic control. Even when CGM is used in certain populations with type 1 or 2 diabetes, clinical focus is often restricted to time in range, hypoglycaemic or hyperglycaemic events, GMI, c.v. and estimated HbA1c[41]. Our unsupervised analysis approach may provide additional predictive signals that are not fully reflected in standard glycaemic metrics used in diabetes care and are therefore an unmet clinical need.

The convergence of increasing CGM data availability, in light of recent approval by the US Food and Drug Administration of over-the-counter devices for nondiabetic use[21], and advancements in AI technology present a substantial opportunity for metabolic health research. GluFormer represents an important step towards using this wealth of information. Moreover, as CGM devices become more affordable and accessible, applications in the wellness realm, such as personalized diet planning for weight loss, are likely to proliferate. The recent global initiative to deliver precision health in diabetes[42] emphasizes the need for a paradigm shift in understanding diabetes heterogeneity. This initiative calls for a redefinition of diabetes subtypes[43], the integration of multiple data sources and the development of new biomarkers. GluFormer aligns closely with these objectives, offering a tool for capturing diabetes heterogeneity across diverse populations. Trained on the HPP dataset, which includes a diverse population that encompasses various ethnicities[2], GluFormer is designed to be applicable to diverse demographics.

Despite the demonstrated performance and broad applicability of GluFormer, several limitations must be acknowledged. First, although our validation methodology ensured participant-level data separation to prevent leakage, the dataset predominantly comprises healthy individuals without diabetes from a limited geographical region, which may restrict the generalizability of the model to populations with rare metabolic conditions or different ethnic backgrounds[44]. However, our models showed good generalizability across 19 studies spanning 5 countries and diverse pathophysiological stages. Our benchmarking choices, particularly the use of GMI as the baseline, although representing current clinical standards, may not capture all possible approaches to CGM data analyses. The varying lengths of CGM recordings and inconsistent availability of contextual information across participants create additional challenges for standardization. The dietary data integrated into the model relies on self-reported logs, which are prone to inaccuracies and omissions, which may affect predictions related to dietary interventions. Moreover, the integration of dietary data required extensive engineering efforts to quantify the nutrient content of each food item, a process that is both expensive and time-consuming, thereby creating barriers to scalability and widespread implementation. Although our model effectively predicted glucose responses to dietary inputs, we acknowledge important limitations, particularly regarding counterfactual prediction modelling that would be necessary for simulating intervention outcomes. Nevertheless, our nutrient-based approach to food representation offers inherent advantages for generalization across dietary patterns and cultural contexts. The success of this multimodal integration suggests promising directions for incorporating additional physiological signals, such as sleep patterns, continuous photoplethysmogram signals and physical activity data, to potentially enable a more comprehensive understanding of health dynamics. This potential motivates the development of a multimodal health model based on continuous signals and tracking. The complexity and interpretability of transformer models, including GluFormer, also pose considerable challenges. These models are often regarded as 'black boxes', making it difficult to understand the reasoning behind their predictions. The computational requirements for training and deploying such models create additional implementation hurdles in clinical settings. Currently, clinical practice relies mainly on simpler metrics, such as HbA1c and fasting glucose levels, and even metrics such as GMI and CGM-based composite scores have not yet been widely adopted. Consequently, although the advanced architecture of GluFormer provides enhanced predictive capabilities for complex tasks, it may still require considerable time before it can be clinically adopted. Furthermore, the model may inherit biases present in the training data, which need to be meticulously managed to ensure accurate and fair predictions across different demographic groups.

In conclusion, GluFormer demonstrated versatility in metabolic health analysis, showing modestly stronger associations than traditional metrics for certain clinical outcomes. Although the representation-learning approach of the model revealed patterns in CGM data that extended beyond conventional glycaemic metrics, further validation studies are needed to establish their full clinical utility. By introducing an analytical approach that captures longitudinal glucose patterns and their associations with health outcomes, this work contributes to ongoing efforts in precision diabetes care and clinical trial optimization. As CGM technology becomes increasingly accessible, approaches such as GluFormer may help advance

the understanding of metabolic health and support more targeted intervention strategies, although substantial work remains to translate these capabilities into clinical practice.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-025-09925-9.


1. Shilo, S. et al. 10 K: a large-scale prospective longitudinal study in Israel. *Eur. J. Epidemiol.* **36**, 1187–1194 (2021).
2. Reicher, L. et al. Deep phenotyping of health–disease continuum in the Human Phenotype Project. *Nat. Med.* **31**, 3191–3203 (2025).
3. Nathan, D. M. et al. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N. Engl. J. Med.* **329**, 977–986 (1993).
4. King, P., Peacock, I. & Donnelly, R. The UK prospective diabetes study (UKPDS): clinical and therapeutic implications for type 2 diabetes. *Br. J. Clin. Pharmacol.* **48**, 643–648 (1999).
5. Gude, F. et al. Glycemic variability and its association with demographics and lifestyles in a general adult population. *J. Diabetes Sci. Technol.* **11**, 780–790 (2017).
6. Saab, K. et al. Capabilities of Gemini models in medicine. Preprint at https://doi.org/10.48550/arxiv.2404.18416 (2024).
7. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
8. Lutsker, G., Rossman, H., Godiva, N. & Segal, E. COMPRER: a multimodal multi-objective pretraining framework for enhanced medical image representation. Preprint at https://doi.org/10.48550/arxiv.2403.09672 (2024).
9. Yuan, H. et al. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *npj Digit. Med.* **7**, 91 (2024).
10. Thapa, R. et al. SleepFM: multi-modal representation learning for sleep across brain activity, ECG and respiratory signals. *Proc. Mach. Learn. Res.* **235**, 48019–48037 (2024).
11. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
12. Krishnan, R., Rajpurkar, P. & Topol, E. J. Self-supervised learning in medicine and healthcare. *Nat. Biomed. Eng.* **6**, 1346–1352 (2022).
13. GBD 2021 Diabetes Collaborators. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. *Lancet* **402**, 203–234 (2023).
14. Rawshani, A. et al. Mortality and cardiovascular disease in type 1 and type 2 diabetes. *N. Engl. J. Med.* **376**, 1407–1418 (2017).
15. Moser, E. G., Crew, L. B. & Garg, S. K. Role of continuous glucose monitoring in diabetes management. *Av. Diabetol.* **26**, 73–78 (2010).
16. Battelino, T. et al. Continuous glucose monitoring and metrics for clinical trials: an international consensus statement. *Lancet Diabetes Endocrinol.* **11**, 42–57 (2023).
17. Kieu, A., King, J., Govender, R. D. & Östlundh, L. The benefits of utilizing continuous glucose monitoring of diabetes mellitus in primary care: a systematic review. *J. Diabetes Sci. Technol.* **17**, 762–774 (2023).
18. Holzer, R., Bloch, W. & Brinkmann, C. Continuous glucose monitoring in healthy adults-possible applications in health care, wellness, and sports. *Sensors* **22**, 2030 (2022).
19. Zahedani, A. D. et al. Digital health application integrating wearable data and behavioral patterns improves metabolic health. *npj Digit. Med.* **6**, 216 (2023).
20. Shilo, S. et al. Continuous glucose monitoring and intrapersonal variability in fasting glucose. *Nat. Med.* **30**, 1424–1431 (2024).
21. U.S. Food & Drug Administration. FDA clears first over-the-counter continuous glucose monitor. *FDA* https://www.fda.gov/news-events/press-announcements/fda-clears-first-over-counter-continuous-glucose-monitor (2024).
22. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://doi.org/10.48550/arxiv.1802.03426 (2018).
23. Bergenstal, R. M. et al. Glucose management indicator (GMI): a new term for estimating A1C from continuous glucose monitoring. *Diabetes Care* **41**, 2275–2280 (2018).
24. Broll, S. et al. Interpreting blood GLUcose data with R package iglu. *PLoS ONE* **16**, e0248560 (2021).
25. Rodbard, D. New and improved methods to characterize glycemic variability using continuous glucose monitoring. *Diabetes Technol. Ther.* **11**, 551–565 (2009).
26. Wang, J. et al. Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nat. Med.* **31**, 609–617 (2025).
27. Keshet, A. et al. CGMap: characterizing continuous glucose monitor data in thousands of non-diabetic individuals. *Cell Metab.* **35**, 758–769 (2023).
28. Vickers, A. J., Van Calster, B. & Steyerberg, E. W. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* **352**, i6 (2016).
29. Van Calster, B. et al. Performance evaluation of predictive AI models to support medical decisions: overview and guidance. Preprint at https://doi.org/10.48550/arxiv.2412.10288 (2024).
30. Htet, T. D. et al. Rationale and design of a randomised controlled trial testing the effect of personalised diet in individuals with pre-diabetes or type 2 diabetes mellitus treated with metformin. *BMJ Open* **10**, e037859 (2020).
31. Rein, M. S. et al. BREAst Cancer Personalised NuTrition (BREACPNT): dietary intervention in breast cancer survivors treated with endocrine therapy—a protocol for a randomised clinical trial. *BMJ Open* **12**, e062498 (2022).
32. Ben-Yacov, O. et al. Personalized postprandial glucose response-targeting diet versus Mediterranean diet for glycemic control in prediabetes. *Diabetes Care* **44**, 1980–1991 (2021).
33. The Diabetes Prevention Program Research Group. The Diabetes Prevention Program. Design and methods for a clinical trial in the prevention of type 2 diabetes. *Diabetes Care* **22**, 623–634 (1999).
34. International Expert Committee. International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes. *Diabetes Care* **32**, 1327–1334 (2009).
35. Cersosimo, E., Solis-Herrera, C., Trautmann, M. E., Malloy, J. & Triplitt, C. L. Assessment of pancreatic β-cell function: review of methods and clinical applications. *Curr. Diabetes Rev.* **10**, 2–42 (2014).
36. Abdul-Ghani, M. A. et al. The relationship between fasting hyperglycemia and insulin secretion in subjects with normal or impaired glucose tolerance. *Am. J. Physiol. Endocrinol. Metab.* **295**, E401–E406 (2008).
37. Ansari, A. F. et al. Chronos: learning the language of time series. *Transact. Mach. Learn. Res.* https://openreview.net/forum?id=gerNCVqqtR (2024).
38. Rabanser, S., Januschowski, T., Flunkert, V., Salinas, D. & Gasthaus, J. The effectiveness of discretization in forecasting: an empirical study on neural time series models. Preprint at https://doi.org/10.48550/arxiv.2005.10111 (2020).
39. van den Oord, A. et al. WaveNet: a generative model for raw audio. In *Proc. 9th ISCA Speech Synthesis Workshop* 125 (2016).
40. van den Oord, A., Kalchbrenner, N. & Kavukcuoglu, K. Pixel recurrent neural networks. In *Proc. 33rd International Conference on Machine Learning* (eds Balcan, M. F. & Weinberger, K. Q.) 1747–1756 (2016).
41. Danne, T. et al. International consensus on use of continuous glucose monitoring. *Diabetes Care* **40**, 1631–1640 (2017).
42. Cefalu, W. T. et al. A global initiative to deliver precision health in diabetes. *Nat. Med.* **30**, 1819–1822 (2024).
43. Ahlqvist, E., Prasad, R. B. & Groop, L. Subtypes of type 2 diabetes determined from clinical parameters. *Diabetes* **69**, 2086–2093 (2020).
44. Xiong, Z. et al. How generalizable are foundation models when applied to different demographic groups and settings? *NEJM AI* https://doi.org/10.1056/AIcs2400497 (2024).

# Methods

## GluFormer

**Data.** The training dataset comprised CGM records from 10,812 participants, each monitored over a 2-week period using a Freestyle Libre Pro 2 device (Abbott), which records glucose levels subcutaneously every 15 min. Only sensor-derived CGM readings were used; manually logged blood glucose values were not included. The Freestyle Libre Pro 2 device measures glucose levels between 40 and 500 mg $dl^{-1}$ (displaying 'lo' or 'hi' for values outside this range). In our cohort, 95% of CGM values ranged from 54 mg $dl^{-1}$ to 171 mg $dl^{-1}$ (see also a previous study[27], which reported 94.3% of values between 70 mg $dl^{-1}$ and 180 mg $dl^{-1}$ in the HPP cohort), with no measurements recorded below the 40 mg $dl^{-1}$ lower detection limit of the device, and the highest values reaching approximately 499 mg $dl^{-1}$ (100th percentile). First-day and last-day readings were excluded to mitigate noise from device calibration errors, which are commonly observed during the first 24 h. Participants with fewer than 100 readings were also removed from the dataset. Missing data points, occurring in less than 0.1% of instances, were addressed through linear interpolation to fill gaps owing to time skips in measurements.

**Tokenization.** We adopted a tokenization approach that transforms continuous glucose measurements into discrete tokens, drawing on compelling evidence that discretization can substantially enhance time-series forecasting performance in neural models[37–40]. Notably, large-scale empirical analyses have demonstrated that binning real-valued observations into categories can provide both stability in training and improved predictive accuracy across different architectures, even though it discards both the continuous and the ordinal nature of the raw data in favour of discrete labels[37]. These benefits are thought to arise because tokenization regularizes the input space, which effectively reduces noise and forces the model to learn robust patterns rather than overfit to minor fluctuations. Moreover, the discrete nature of tokens aligns with the protocols used in large language models, in which a fixed vocabulary simplifies the training pipeline. Indeed, recent developments in foundational time-series models have illustrated the potential of direct analogy to natural language-processing frameworks: models such as Chronos[37] suggest that by treating time-series signals as 'words' in a learned vocabulary, one can leverage the SSL strategies that have revolutionized natural language processing[37]. Likewise, a previous study[38] showed that discretizing continuous values into tokens can outperform sophisticated continuous-valued methods, often because of better generalization and reduced sensitivity to outliers[37]. These insights also echo findings in generative modelling of audio and images, for which discretizing raw data (for example, quantized waveforms or pixel intensities) proved crucial for the success of convolutional and transformer-based architectures[39,40]. In this spirit, we quantized glucose readings between 40 mg $dl^{-1}$ and 500 mg $dl^{-1}$ into 460 evenly spaced intervals on the basis of their empirical distribution in the training dataset. Each resulting token corresponded to one of these discrete bins, which created a well-defined vocabulary of glucose states. Samples with fewer than 1,200 measurements were padded with a special <MASK> token to maintain a constant input size of 1,200 for training and evaluation.

Data division into training, validation and testing sets was participant-based rather than sample-based to ensure consistency in model evaluation. The division was made randomly and was divided into a training set of 80% of participants, a validation set of 10% and a test set of 10%.

**Model architecture.** We used a transformer-based model[45] structured mainly around a decoder mechanism. The architecture was defined using the following parameters: embedding dimension of 1,024, 16 attention heads, 16 transformer layers (or blocks) and a feed-forward dimension of 2,048. The model was designed to process sequences of up to 25,000 tokens in length and was trained with a context length of 1,200 tokens. These choices were made on the basis of a hyperparameter search.

The vocabulary size of the model was 461: 460 values representing glucose values from 40 mg $dl^{-1}$ to 500 mg $dl^{-1}$, and 1 masking token.

The model initially embedded all tokens using the embedding layer (vocabulary size × embedding dimension), such that a single sample was embedded into a matrix of context length × embedding dimension. These then entered the transformer blocks and exited with the same dimensions. These were inputted into the un-embedding layer (same dimension as the embedding matrix) and transformed to a distribution over possible token – context length × vocabulary size.

Each position was trained to predict the next token, and so these distributions were compared with the observed next tokens using cross-entropy to match the observed distribution. We also added positional encoding[45] to model the global positions on the tokens in the sequence.

To create a generative autoregressive model, we used a causal masking technique[45] during training. This approach ensured that the attention mechanism only considered past tokens as context, which effectively prevented access to information from future tokens.

**Pretraining and optimization.** The objective during pretraining was to forecast subsequent tokens using a causal masking technique. This task was structured to enhance predictive accuracy by learning from unmasked tokens to predict the subsequent masked ones, using cross-entropy loss calculated across the distribution of the 460 possible tokens.

**Optimization details.** Model optimization was conducted using the AdamW optimizer[46] with a learning rate of $3 \times 10^{-5}$, with no weight decay and dropout at 0.1, with a batch size of 32 per GPU (effective batch size of $32 \times 8 = 256$). A StepLR scheduler was applied to adjust the learning rate with a decay factor of 0.99 every 10 steps over a total of 100 epochs. Model selection was based on performance metrics from the validation set, which accounted for 10% of the data. The final model was evaluated using the test set, which also represented 10% of the data. Training was performed on 8 NVIDIA A40.

**Producing an output embedding per sample.** Producing a single embedding for each CGM sample offers a compact, learned summary of an individual's glycaemic profile. By compressing long sequences of CGM readings into a single vector, we reduced the dimensionality of the data while retaining meaningful patterns that are critical for downstream tasks, such as disease risk classification or outcome prediction. This representation-based approach parallels methods in other domains (for example, language modelling), in which embeddings succinctly capture the essential features of lengthy input sequences in a way that is both interpretable and readily integrable with traditional machine-learning algorithms.

To represent each CGM sample, we wanted to use the output of the model to obtain the processed version of the sample. Each CGM sample, which contained 1,200 glucose measurements (tokens), was passed through the transformer, which output a 1,024-dimensional vector for each token. We kept these vectors in the high-dimensional representation space rather than convert them back to token probabilities.

For aggregation of vectors, we needed to reduce the 1,200 vectors from each sample into a single representative vector. We evaluated three pooling methods: average pooling; maximum pooling; and minimum pooling. For average pooling, we calculated the mean of the 1,200 vectors, assuming equal contribution from all time points.

For maximum pooling, we selected the maximum value for each of the 1,024 dimensions across the 1,200 vectors, highlighting the most prominent features in the sequence. For minimum pooling, we selected the minimum value for each dimension, capturing the lowest bounds of the glucose measurements. Note that we removed the

<MASK> token before doing any of the pooling methods to remove any non-informative data from the representation.

These methods were evaluated on the basis of their ability to predict clinical outcomes using representations generated from the validation set. Maximum pooling emerged as the most effective method, as it consistently provided better performance in predicting key clinical measures such as HbA1c.

**Initialization of the embedding layer.** Our initial approach for embedding tokens in GluFormer followed standard practice in large language models, wherein the embedding matrix is randomly initialized. We experimented with alternative schemes aimed at leveraging previous knowledge of the scalar glucose values, including techniques that enforced uniform Euclidean or cosine distances among tokens and regularization approaches that maintained these distances throughout training. Despite our expectations that such specialized initializations might produce performance gains—given the known meaning of scalar glucose values—none of these methods consistently improved prediction accuracy or other metrics. Consequently, we reverted to random initialization of the embedding layer for the final model. Nevertheless, we consider that further research is warranted to explore whether specialized embeddings can more effectively encode numeric relationships among glucose measurements.

### UMAP

To visualize the high-dimensional embedding space generated by GluFormer and to explore the clustering of glycaemic profiles, we used UMAP with the umap-learn Python package (v.0.5.6). All representations from the HPP dataset, encompassing training, validation and test subsets, were included to ensure a comprehensive representation of the data distribution. Before dimensionality reduction, representations were standardized to have zero mean and unit variance to facilitate effective distance calculations during the UMAP process. We used the default parameters of UMAP, specifically setting n_neighbors to 15, min_dist to 0.1 and the distance metric to cosine, aligning with the cosine similarity used in the distance analyses.

The standardized representations were then input into the UMAP algorithm to project the 1,024-dimensional vectors into a two-dimensional space, which enabled intuitive visualization of potential clusters and gradients in the data. The choice of cosine as the distance metric ensured consistency with the intra-participant and inter-participant analyses, which preserved the angular relationships between representations. The resulting UMAP projections were subsequently used to generate scatter plots, which were colour-coded on the basis of various clinical and demographic attributes to facilitate the examination of underlying patterns and associations in the embedding space. This dimensionality-reduction technique provided a meaningful visualization framework to interpret the complex relationships captured by representations generated using GluFormer, thereby supporting further downstream analyses and validation of the performance of the model across diverse clinical parameters.

### Clinical measures in UMAP

For HbA1c (Supplementary Fig. 1A), when coloured by independently measured HbA1c levels, the UMAP projection exhibited a gradient from lower to higher values, which indicated that the representations generated using GluFormer effectively captured long-term glycaemic control.

For visceral fat mass (Supplementary Fig. 1B), representations colour-coded by dual-energy X-ray absorptiometry-derived visceral fat mass showed distinct clustering, which suggested that the model encoded information related to body composition.

For hypoglycaemic and hyperglycaemic events (Supplementary Fig. 1C), the distribution of representations based on the percentage of time CGM values above 140 mg $dl^{-1}$ and below 70 mg $dl^{-1}$ displayed clear separations, which reflected the sensitivity of the model to glycaemic variability.

For demographic factors (Supplementary Fig. 1D,E), age-specific and sex-specific colourations further demonstrated that GluFormer representations encapsulated demographic variations, although to a lesser extent compared with clinical metrics.

### Intra-distance and inter-distance analysis

To evaluate the distinctiveness of GluFormer representations at both individual and population levels, we conducted an intra-participant and inter-participant distance analysis using cosine similarity metrics. This analysis aimed to determine whether the model effectively captures unique glycaemic patterns specific to each individual while maintaining broader generalizability across the population.

First, representations for each CGM sample in the test set were generated using the pretrained GluFormer model. Each CGM sample, comprising 1,200 glucose measurements, was processed to obtain a 1,024-dimensional embedding vector through the maximum pooling aggregation method described above. This resulted in an embedding representation for each participant's glycaemic profile on a given day.

**Distance calculation.** For intra-participant distances, for each participant, we computed the cosine distance between all pairs of their representations across different days. Specifically, for a participant with $n$ CGM samples, we calculated $n(n-1)/2$ pairwise cosine distances. This measures the variability of representations in the same individual over time.

For inter-participant distances, to assess the diversity between different individuals, we calculated the cosine distance between representations from distinct participants. Given the large number of participants, we used a stratified sampling approach to ensure computational feasibility while maintaining representativeness. Specifically, for each participant, we randomly selected 100 representations from other participants to compute inter-participant cosine distances.

**Statistical analysis.** To statistically compare the intra-participant and inter-participant distance distributions, we used the Mann–Whitney $U$-test, a nonparametric test suitable for comparing two independent samples without assuming normal distribution. Given the large sample size, the test provides robust significance levels.

A threshold of $P < 0.001$ was set to determine significance, which mitigated the risk of type I errors given the extensive pairwise comparisons.

### Predictive stratification analysis

To evaluate the predictive power of the GluFormer-derived score compared with baseline HbA1c in forecasting future glycaemic trajectories, we conducted a quartile-based stratification analysis in a cohort of 337 individuals with prediabetes. Prediabetes was defined by baseline HbA1c levels ranging from 5.7% to 6.4%, as per established clinical guidelines. Baseline HbA1c measurements were obtained from clinical records at the time of CGM data collection. Using the GluFormer model, we generated predicted HbA1c values on the basis of CGM-derived representations for each participant. Participants were then stratified into quartiles on the basis of either their baseline HbA1c or their GluFormer-predicted HbA1c values. Over a 2-year follow-up period, changes in HbA1c levels were tracked for each quartile. To assess the significance of differences in future HbA1c changes across quartiles, we used the Mann–Whitney $U$-test, setting significance thresholds at $^{*}P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.0001$.

The results of this analysis are visualized in Fig. 3a using box plots, which display the distribution of the 2-year change in HbA1c for each quartile. In these plots, the central horizontal line indicates the median, and the black triangle represents the mean. The box bounds correspond to the interquartile range (25th to 75th percentiles) and the vertical error bars extend to show the mean ± 1 s.d.

### Methodological considerations for evaluating pretrained embeddings

The evaluation of foundation models, such as GluFormer, often involves assessments of the utility of their learned representations (embeddings) on various downstream tasks. A common and methodologically sound practice in this field is to use embeddings derived from data that may have been part of the initial pretraining dataset for these downstream evaluations, provided that certain conditions are met to prevent data leakage and to ensure a fair assessment of generalization. This section outlines the rationale behind this approach as applied in our study.

The core principle lies in the decoupled nature of the self-supervised pretraining phase and the supervised downstream task. GluFormer was pretrained on a large corpus of CGM time-series data from the HPP using a self-supervised, next-token prediction objective. This phase aimed to learn inherent patterns and rich representations from the glucose dynamics themselves, without any exposure to clinical labels, future outcomes (such as HbA1c changes or diabetes diagnosis) or specific task objectives beyond understanding the structure of CGM data.

When evaluating the pretrained GluFormer on a downstream task, such as predicting future HbA1c changes in the prediabetes analysis cohort drawn from the HPP (which includes participants from the original pretraining, validation and test splits of the foundation model), the following critical distinctions ensured the validity of the evaluation:

1. Decoupled tasks and objectives: the self-supervised pretraining task (learning general patterns in the CGM time series) was fundamentally different and separate from the supervised downstream task (for example, using the learned representations to predict a specific clinical outcome, such as 2-year HbA1c change). The information required and the objectives optimized for these were distinct.
2. Frozen representations: crucially, for the downstream evaluation tasks, the pretrained GluFormer model itself is not retrained or fine-tuned. Its learned parameters are fixed or 'frozen'. The evaluation, therefore, tests the inherent informational content and quality of these static, prelearned embeddings when applied to a new, specific problem. The downstream model (for example, a lightweight linear ridge regression model) is trained solely on these fixed embeddings to perform the prediction, not on the raw CGM data, and the weights of the GluFormer model remain unchanged.
3. Temporal separation for longitudinal outcomes: in instances when the downstream task involves predicting a future outcome (for example, HbA1c change at a 2-year follow-up), this temporal separation provides an additional layer of distinction. The GluFormer model, during its pretraining, had no access to this future information. The evaluation assesses whether the patterns learned from baseline CGM data can prospectively inform about future physiological changes.
4. Assessing the generalization power of embeddings: the primary goal of this evaluation strategy was to demonstrate that the unsupervised pretraining phase had successfully captured meaningful, generalizable variance in the CGM data (for example, patterns indicative of underlying metabolic health or dysfunction). Using embeddings derived from participants who were part of the pretraining set to predict a different, previously unseen target variable for a distinct task is a valid and standard method to assess this generalization capability. It analyses whether the learned representations are sufficiently versatile to be useful for various applications beyond the original self-supervised objective.

This approach of evaluating fixed embeddings from a pretrained model on downstream tasks, including those that may use data from the original pretraining cohort for these new tasks, is standard practice in the development and assessment of foundation models across various domains[7,47–51]. These studies similarly leveraged the power of pretrained representations for diverse downstream applications, thereby underscoring the utility and generalizability of the features learned during the initial self-supervised or unsupervised phase.

### Prediction performance of GluFormer versus CGM-derived composite scores and GMI across subgroups

To evaluate the predictive performance of GluFormer in comparison to CGM-derived composite scores (iglu) and the GMI across different clinical subgroups, we conducted subgroup-specific regression analyses followed by correlation assessments. The study cohort was stratified into four distinct subgroups on the basis of baseline glycaemic status: healthy; prediabetic; diabetic; and a combined prediabetic and diabetic group. For each subgroup, we used ridge regression models to predict clinical outcomes, specifically HbA1c and fasting glucose levels, using three sets of input features: GluFormer-derived representations; CGM-derived composite scores (as extracted using the iglu package); and GMI values. For the GluFormer embeddings, using an independent validation set, we observed that an alpha of between 50 and 100 produced the best results. Therefore, we chose an alpha value of 80 for all ridge regressions we did with the GluFormer embeddings. For iglu, we saw the same effect with an alpha value of between 0.1 and 1; therefore, we chose 0.5 for this analysis. For GMI, regularization was not needed as it consists of only one variable. Hence, we set alpha to be 0. Although the ridge regression framework was used for consistency across all feature sets, for the single-variable GMI predictor (with alpha = 0), this is equivalent to standard linear regression, and the regularization aspect is inactive.

The models were trained and validated using a consistent cross-validation framework to ensure comparability across methods. Following model training, we calculated the Pearson's correlation coefficients between the predicted and actual clinical measurements in each subgroup to quantify predictive accuracy.

To determine the significance of the performance of GluFormer relative to CGM-derived composite scores and GMI, we applied the Mann–Whitney *U*-test to compare the distributions of Pearson's correlation coefficients across the different prediction methods in each subgroup.

### Survival analysis (Kaplan–Meier) for long-term outcomes

We used data from the AEGIS study[5], which tracked 580 participants with a median follow-up of 11 years, to determine whether GluFormer representations or standard HbA1c measurements provide stronger prognostic information for future diabetes onset. Among these participants, 42 individuals developed diabetes during follow-up, and we recorded the time (in days) until diagnosis. To evaluate stratification performance, we split participants into top and bottom quartiles on the basis of either baseline blood HbA1c or predicted HbA1c values derived from GluFormer embeddings. For baseline HbA1c, this involved simply ranking all participants by measured HbA1c and taking the highest and lowest 25%. For GluFormer, we used the same ridge regression protocol (with alpha = 80) described in the section 'Predictive modelling of medical measures on HPP participants using representations'. For HbA1c, we used a *k*-fold model on the AEGIS CGM data and aggregated the results into a predicted HbA1c. Participants were then ranked on the basis of these predicted values, again splitting into top and bottom quartiles. We generated Kaplan–Meier curves for each group and performed a log-rank test to compare their time-to-diabetes outcomes, thus testing whether the stratification by GluFormer-derived versus blood-measured HbA1c differed in separating those who would develop diabetes from those who remained diabetes-free.

### ROC AUC analysis for diabetes prediction at different time points

To evaluate the predictive performance of GluFormer-derived representations in forecasting the onset of diabetes, we conducted ROC curve analyses at three distinct temporal horizons: at the time of CGM

collection (baseline), and at 2 years and at 4 years after CGM monitoring. The study cohort comprised 337 individuals with prediabetes, defined by baseline HbA1c levels ranging from 5.7% to 6.4%, as per established clinical guidelines. Diabetes diagnosis was determined on the basis of blood HbA1c test or on blood fasting glucose during follow-up assessments conducted at the 2-year and 4-year marks. For each time point, we developed binary classification models using three different sets of predictors: (1) GluFormer-derived representations, generated by processing baseline CGM data through the pretrained GluFormer model to obtain 1,024-dimensional feature vectors; (2) GMI scores, calculated using the iglu package[24,25]; and (3) baseline HbA1c levels, serving as a traditional glycaemic marker.

Logistic regression was used to construct the classification models, with each set of predictors used independently to forecast diabetes onset. To assess model performance, ROC curves were plotted, and the AUC was computed for each predictor across all time points.

### Net benefit analysis for diabetes prediction at different time horizons

To evaluate the clinical utility of GluFormer-derived predictions in diabetes forecasting, we used decision curve analysis[28,29] to assess the net benefit of GluFormer compared with the GMI across varying threshold probabilities and temporal horizons. The analysis was conducted at three distinct time points: baseline (time of CGM collection), and at 2 years and 4 years after CGM monitoring. For each time horizon, we generated predicted probabilities of diabetes onset using both GluFormer and GMI models. Probabilities were calculated using logistic regression fitted on both the representations (as was done for the ROC AUC analysis, described in the section above), as well as GMI using the predict_proba method in sklearn. The true diabetes outcomes were determined on the basis of blood HbA1c tests or on blood fasting glucose during follow-up assessments conducted at the 2-year and 4-year marks. Using the dcurves Python package, we calculated the net benefit for each model across a range of threshold probabilities (0 to 1 in increments of 0.01), which represent the clinician's willingness to treat on the basis of predicted risk. This approach enables comparisons of models by quantifying the trade-off between true-positive and false-positive classifications, thereby providing insight into the clinical relevance of each predictive model.

For the decision curve analysis, net benefit curves were plotted alongside baseline strategies of 'treat all' and 'treat none' to contextualize the performance of GluFormer and GMI. The net benefit was calculated using the formula:

$$\text{Net benefit} = (\text{true positives}/n) - (\text{false positives}/n) \times (\text{threshold probability}/(1 - \text{threshold probability}))$$

where $n$ is the total number of participants.

### CGM-derived clinical metrics using iglu

To extract meaningful clinical insights from CGM data, we used the R package iglu[24,25]. Iglu is designed to derive a comprehensive set of metrics from CGM data, which are pivotal in assessing glucose control and variability across individuals.

### Functionality and application of iglu

Iglu integrates advanced computational algorithms to process CGM data and provides more than just basic reading and organizing functionalities. Unlike other tools that may offer limited analysis capabilities, iglu supports a full range of CGM-derived metrics. This tool facilitated an in-depth evaluation of glucose dynamics, which enabled the categorization of glucose management into several clinically relevant aspects[27].

For mean glucose measures, metrics such as mean glucose levels and estimated A1C were calculated to evaluate glycaemic control.

For postprandial glucose adaptation, measures such as s.d. and mean amplitude of glycaemic excursions assessed responses to meal intake.

For composite and range metrics, these included time in range and the *J*-index, which provide insights into overall glucose exposure and variability.

All iglu measurements used in this study are detailed Supplementary Table 2.

### Predictive modelling of medical measures on HPP participants using representations

To evaluate the predictive power of GluFormer representations for future health outcomes, we conducted a series of analyses targeting a broad range of clinical metrics. These metrics included blood tests, body measurements, anthropometric data and body composition parameters, such as muscle and fat mass. The primary goal was to assess the ability of GluFormer representations to forecast these health metrics over different time horizons, thereby validating their clinical applicability.

For each health metric, we trained a ridge regression model using iglu measures and GluFormer output representations as input features. The representations, representing the CGM data for each participant, were used to predict various clinical metrics recorded for those individuals. Specifically, we used a *k*-fold cross-validation approach, dividing the data into five folds and iteratively using each fold as the test set while aggregating results across all folds. We used this aggregated result to compute a Pearson's correlation between observed and predicted values for all samples in the set. This process was repeated ten times with different random seeds to ensure robustness, and the results were averaged, with s.d. values calculated. We evaluated the significance of our predictions using a permutation test across 100 different seeds, and we considered results as significant when random performance did not exceed the model performance more than five times ($P < 0.05$).

### GluFormer-derived score

The GluFormer-derived score is a measure created by processing CGM data through our GluFormer model to generate representations, which were then used to estimate HbA1c using a ridge regression model (with alpha = 80, as described in the 'Predictive modelling of medical measures' section). Note that this score is derived using only baseline CGM data.

### Methodology for out-of-cohort generalization

Here we specifically sought to determine how well GluFormer, trained only on the HPP training set, would generalize to these out-of-cohort datasets. First, each external dataset was fed into the pretrained GluFormer to obtain CGM representations. Next, using the same ridge regression approach described in the 'Predictive modelling of medical measures on HPP participants using representations' section, we evaluated how effectively these embeddings predict relevant clinical measures in the new populations. This procedure ensures consistency across all evaluations and provides a direct assessment of the robust transferability of GluFormer to diverse real-world settings.

### Out-of-cohort generalization

To evaluate the generalizability of our model to out-of-distribution data, we focused on assessing performance using all datasets that have CGM data, including the HPP, PNP1, PNP3, BREACPNT, PREDICT, T1DM, GDM, BREAD, IL Healthy, US Healthy, US Obese, Colas 2019, JDRF CGM RCT, Shanghai T2DM and Hypoglycemia in Older Adults (Supplementary Table 1). Using these datasets, we wanted to evaluate the performance of GluFormer under different clinical conditions (for example, type 1 and 2 diabetes, gestational diabetes, among others), different devices (for example, iPro2, Dexcom, Medtronic, among others) and different countries (Australia, United States, Spain and China).

### Data preparation and embedding for out-of-cohort

To be consistent with the preprocessing method used for the HPP dataset, glucose measurements from the external cohorts were discretized using the same bins that were established in the 'Tokenization' section based on the 80% training partition of the 10,812 participant dataset. If the data were from clinical trials, we further divided the data into two distinct phases: pre-intervention and post-intervention. We embedded the full sequence of glucose measurements directly into the 1,024-dimensional space, applying the maximum pooling method to generate a single representative vector for each phase of the study.

### Predicting outcomes of randomized clinical trials

For the outcome prediction analysis for randomized clinical trials (for full details of the studies and cohorts used, see Supplementary Table 1), we used the pre-intervention CGM GluFormer representations and added a binary variable to indicate the allocation of the participants. Each trial had two intervention arms, so each representation vector was increased to 1,025 in size, in which the last entry was either 0 or 1 depending on their allocation. We then used either iglu or the representations to predict the primary and secondary outcomes of the study and evaluated the results using Pearson's correlation. The trial cohorts in this category were PREDICT, BREACP and PNP3.

For the open-source datasets, we focused on primary outcomes and, using the same scheme, attempted to predict the primary outcome for each dataset.

For JDRF, we aimed to predict HbA1c at 26 weeks using CGM data from 1–6 months, evaluating the results using Pearson's correlation.

For Shanghai T2D, we aimed to predict HbA1c after 26 weeks using 2 weeks of baseline CGM data, evaluating the results using's Pearson correlation.

For Colas, we aimed to predict baseline HbA1c using 2 weeks of CGM data, evaluating the results using Pearson's correlation.

For Hypoglycaemia in Older Adults, we used 1 week of CGM data to predict whether there would be an event of hypoglycaemia in the subsequent weeks, evaluating the results using ROC AUC.

For Aleppo 2017, we aimed to predict HbA1c at baseline of study using 1 week of baseline CGM data, evaluating the results using Pearson's correlation.

For Chase 2005, we aimed to predict HbA1c at baseline of study using 1 week of baseline CGM data, evaluating the results using Pearson's correlation.

For Wadwa 2023, we aimed to predict HbA1c after 13 and 26 weeks using 1 week of baseline CGM data, evaluating the results using Pearson's correlation.

### Predictive modelling of randomized clinical trials

The predictive analysis was again conducted using ridge regression, consistent with our previous methodological approach to ensure comparability. We used ridge regression with *k*-fold optimized over ten different seeds. The folds were randomly assigned on the basis of participant identifiers. *P* values were obtained over a permutation test over 100 different seeds.

### Comparative analysis of SSL, plain transformer and convolutional neural network models

This section outlines the comparative performance of four distinct models: GluFormer; a transformer with the same architecture as GluFormer; frozen GluFormer representations; and a convolutional neural network (CNN) model. The models were tested in a downstream task of prediction HbA1c from CGM, as detailed in Supplementary Fig. 2. Each model was trained using the same dataset and evaluated on the same validation and test sets to ensure comparability. The split was the same 80% train, 10% validation and 10% test as was done for all other models.

### Fine-tuning the pretrained CGM transformer

To tailor the GluFormer architecture for the specific task of HbA1c prediction, we modified its output mechanism. Initially trained to predict the next token, the output layer of the model transformed representations into the token space. For fine-tuning, we replaced this output layer with a one-dimensional adaptive pooling layer that aggregated the 1,200 representations into a single vector. We then added a small, three-layer MLP with Gaussian error linear unit activation and a single neuron output to predict the measure from this vector. The fine-tuned model was optimized using mean squared error (MSE) for regression tasks and with the AdamW optimizer.

### Training CGM transformer from scratch (no SSL)

To evaluate the intrinsic benefit of SSL, a CGM transformer model with the same architecture (including the transformer backbone, the one-dimensional pooling and the three-layer MLP) was trained from scratch with randomly initialized weights. This approach aimed to isolate the effect of SSL from the architectural benefits. The model was optimized using MSE with AdamW.

### Fine-tuning frozen transformer representations

In this variant, the weights of the transformer were frozen, and only the added one-dimensional adaptive pooling layer and the three-layer MLP were trained. This model setup tested the efficacy of the transformer representations when not allowed to adapt during the training of the downstream task. Optimization and hyperparameter tuning followed the same protocol as the other models, focusing on MSE and AdamW optimization.

### CNN model

A CNN was also tested for comparison. The CNN model was extensively tuned across a range of hyperparameters, including different kernel sizes, numbers of layers and MLP configurations. The best-performing setup in the validation set featured a five-layer CNN with a kernel size of three, followed by a three-layer MLP. This model was similarly optimized using MSE with AdamW.

### Adjustments for long CGM sequences

Some external datasets featured very long CGM recordings that exceeded the memory constraints of our system. Specifically, for Wadwa 2023, we split the CGM sequence into two intermediate chunks and embedded each chunk separately. We then computed the mean of these two embeddings to obtain one final representation per individual. For Aleppo 2017, the maximum sequence length was around 26,000 readings (after removing the final day and re-sampling at 15-min intervals). To accommodate the capacity of the model, we took the last 15,000 measurements of each sequence, embedded them and used those embeddings for downstream analyses.

These adjustments ensured that our embedding process remained tractable for longer CGM data while preserving as much relevant information as possible and adjusting for the different tasks for each dataset.

### Temporal modelling

To capture the temporal impacts of glucose fluctuations over time, we incorporated date and time information into our model. We explored two methods to directly integrate temporal information into the representations used by our transformer model.

**Temporal positional encoding.** Building on the positional encoding framework presented in 'attention is all you need'[45], we introduced additional sine and cosine functions to represent temporal dimensions. This method, which we named temporal positional encoding, extends traditional positional encoding by incorporating time-specific waveforms corresponding to minute, hour, day of the week and month.

To tailor these encodings to the physiological context of glucose measurements, we adjusted the phase and wavelength of the sine and cosine functions to align with real-world time cycles.

**Learned temporal representations.** Our second approach involved the use of learned representations for temporal values: minute, hour, day of the week and month. These representations were added directly to the corresponding data representations in the transformer architecture. Unlike the fixed mathematical formulations of sinusoidal encodings, these learned representations were optimized during training, which enabled the model to dynamically adjust and refine its understanding of time as it related to physiological changes and glucose measurements.

In practice, the learned temporal representations outperformed the sinusoidal temporal positional encoding in the generation task over the test set (Supplementary Fig. 5).

### GluFormer with diet tokens

**Preprocessing of CGM and diet data.** To integrate CGM data with dietary logs, we established a preprocessing pipeline that combined these datasets, ensuring accurate alignment and comprehensive representation of nutritional intake alongside glucose measurements. Glucose measurements were recorded every 15 min, and dietary intake was aligned to the nearest 15-min glucose measurement. The diet was also further broken into its macronutrient values.

**Data collection and initial processing.** We collected CGM and dietary data from 10,844 participants. The CGM device recorded glucose levels subcutaneously at regular intervals, and participants logged dietary intake through a mobile application, detailing their consumption of calories, carbohydrates, proteins, lipids and water.

**Temporal alignment and data cleaning.** Dietary entries were aligned with CGM time stamps to ensure temporal correspondence between nutrient intake and glucose readings. We excluded any dietary data not temporally coinciding with CGM records. Glucose values were also clipped to a physiological range of 40 mg $dl^{-1}$ to 500 mg $dl^{-1}$, and nutrient values exceeding the 99th percentile were trimmed to reduce the impact of outliers.

**Caloric and meal filters.** Entries from participants who logged fewer than 1,000 calories throughout the monitoring period were removed. Days with total caloric intake outside the range of 500 to 7,000 calories, or with fewer than 3 meal logs, were also excluded to maintain consistency in dietary patterns.

**Data imputation.** In rare instances of device recording lapses, missing CGM measurements were linearly imputed to maintain a continuous glucose profile.

**Meal consolidation and adjustment.** Multiple dietary logs recorded in the same hour were consolidated into a single meal entry to simplify the dataset. Optionally, we adjusted meal timings to correlate better with observed glucose spikes to enhance the ability of the model to associate dietary intake with glycaemic responses. This was done by observing that high sugar meals that did not have a subsequent spike in blood glucose were probably mis-logged; therefore, we looked in a window of 1 h before and after the meal to see whether we could locate a glucose spike (an increase of more than 30 mg $dl^{-1}$ over that period). If we found one, we moved the logging to 15 min before the spike.

**Data binning and tokenization.** Nutrients were categorized into quantile-based bins to facilitate uniformity in representation. Similarly, minutes were binned into 15-min intervals to synchronize with the frequency of CGM recordings. Glucose measurements and nutrient intake were then discretized and tokenized into integers to provide a standardized scale for model input.

**Statistics.** The preprocessing steps reduced the initial dataset from 10,844 participants to 5,875 participants, with the number of analysable days decreasing from 76,862 to 57,137. This refinement was crucial in ensuring that the dataset comprised only the most reliable and representative entries for model training.

**Usage.** The models that used this preprocessing step were the models using temporal information and the diet variant. Previous results used the previous preprocessing that does not take into account temporal or nutrient-based information.

### Tokenization and temporal encoding

To effectively process and analyse the CGM and dietary data in our GluFormer + diet model, a comprehensive tokenization strategy was used as detailed above. This approach not only standardizes the data but also integrates crucial temporal information to facilitate the ability of the model to capture temporal dynamics associated with glucose fluctuations and dietary intakes.

**One-dimensional sequence formation.** Both CGM and dietary data were tokenized into a unified one-dimensional sequence. This linear sequence format is crucial for the model to enable it to process time-series data as a continuous stream of events, which is essential for capturing the dynamics of glucose responses and nutrient effects over time.

**Incorporation of temporal tokens.** Alongside each glucose and diet token, we included four additional temporal tokens representing the minute, hour, day of the week and month. These tokens were vital for providing the model with context about the timing of glucose readings and dietary logs, which enriched the input data with layers of temporal granularity. For diet-related tokens, time did not progress between entries on macronutrients of the same log.

**Tokenization details.** All glucose measurements and nutrient values were discretized into predefined bins before tokenization, which ensured that each entry conformed to a standardized integer value. Glucose values were tokenized as described above into integer units, and diet nutrients were binned on the basis of quantiles. The following nutrients were used for diet: (1) calories (kcal); (2) carbohydrates (g); (3) protein (g); (4) caffeine (mg); (5) water (g); (6) total lipids (g); (7) alcohol (g); (8) total sugars (g).

The following bin values were used: energy (kcal): 0.02, 31.68, 84.52, 146.04, 195.00, 262.12, 340.58, 423.10, 518.50, 623.53, 745.56, 891.08, 1069.25, 1317.02, 1726.57 and 5832.24; carbohydrate (g): 0.00, 3.47, 9.05, 15.00, 21.60, 28.87, 36.13, 43.08, 51.81, 61.97, 73.86, 88.02, 106.90, 132.38, 177.13 and 608.09; protein (g): 0.00, 0.66, 1.63, 2.81, 4.42, 6.80, 9.61, 13.64, 18.23, 23.76, 30.46, 38.61, 49.34, 65.22, 93.21 and 354.53; caffeine (mg): 0.30, 188.52, 377.04, 754.08 and 2639.28; water (g): 0.00, 4.56, 36.15, 78.24, 129.46, 182.26, 247.63, 311.44, 389.16, 468.20, 498.60, 575.58, 687.39, 855.37, 1109.35 and 4507.10; total lipid (g): 0.00, 0.41, 0.89, 2.59, 5.68, 8.74, 11.97, 15.84, 19.97, 24.95, 30.71, 37.91, 47.38, 60.38, 83.04 and 306.73; alcohol (g): 0.02, 0.88, 2.60, 6.60, 9.90, 13.20, 18.00, 20.00, 26.40, 28.80, 33.00, 49.50, 66.00 and 172.84; total sugars (g): 0.00, 1.25, 2.82, 4.67, 6.88, 9.27, 12.12, 14.70, 18.02, 22.04, 26.04, 31.28, 37.90, 48.22, 65.61 and 266.78.

This method simplified processing of the model by reducing the complexity of input data and ensuring consistency across the dataset.

**Parallel temporal information.** To ensure that the model accurately recognizes the sequence of events, temporal tokens accompanied each glucose or diet token without advancing time unnecessarily during diet entries. This approach kept the temporal data consistent and precise and reflected the actual timing of meals and glucose measurements without artificial shifts.

### Exploring temporal encoding in CGM data generation

To enhance the capacity of GluFormer to capture temporal information, we incorporated date and time features into the architecture of GluFormer through learned representations for minute, hour, day of the week and month (Temporal learned tokens). We then trained two versions of the model using a preprocessed HPP dataset (Data

consistency across models) and evaluated their performance in generating CGM data for test participants (Supplementary Fig. 5). The temporal-informed model achieved a Pearson's correlation of 0.22 ($P < 0.001$) with participants' observed CGM data, outperforming the original GluFormer, which exhibited a correlation of 0.15 ($P < 0.001$).

### Transformer configuration

The GluFormer + diet model used a transformer architecture that was specifically tailored to effectively handle both CGM and dietary data. This setup enabled the model to learn from the complex relationships between nutrient intake and glucose levels while incorporating temporal dynamics. The following key components constituted the architecture of the model: (1) embedding size, whereby each input token was represented in a 1,024-dimensional space; (2) heads and layers, whereby the model featured 8 attention heads and 10 transformer layers; (3) feed-forward dimension, whereby a dimension of 2,048 was used for the feed-forward networks in each transformer block.

### Additional training parameters

The model was trained on sequences that included both glucose and diet tokens. However, the diet tokens served exclusively as contextual information. The predictive task was focused on forecasting glucose tokens only. This selective prediction approach helped the model specialize in glucose dynamics while still understanding the impact of dietary inputs as contextual modifiers.

### Positional and modality encodings

**Positional encoding.** Traditional positional encodings were used to preserve the order of tokens in the sequence, which was crucial for the model to understand sequence-dependent changes in glucose levels.

**Modality tokens.** Each token type (glucose, calories, sugars, carbohydrates, proteins, lipids, alcohol and water) was associated with a learned modality embedding. These representations were added to the input representations and provided the model with information about the nature of each token, which enhanced its ability to effectively process mixed data types. This approach is similar to segment representations in BERT[50], in which they add learned representations to indicate sentence structure.

**Temporal positional encoding.** To integrate temporal information more directly, the model included an additional positional encoding for time.

**Non-learned temporal encoding.** The last eight dimensions of the embedding space were reserved for sinusoidal positional encodings that represent minute, hour, week and month. Each unit of time was encoded using both sine and cosine functions, which resulted in eight dimensions that helped the model grasp the cyclic nature of time.

This encoding ensured that at any point in the sequence, the model had immediate access to precise temporal information, which enhanced its ability to make time-sensitive predictions.

**Temporal learned tokens.** Alongside the static temporal encodings, the model also incorporated learned temporal tokens for minute, hour, day and month. These tokens were preprocessed and added to the embedding of each token before entering the transformer blocks. This setup enabled the model to adjust its response on the basis of learned patterns associated with specific times, thereby providing a dynamic and responsive predictive mechanism.

### Autoregressive generation and inference procedure

A key capability of GluFormer is to generate synthetic CGM time series through an autoregressive inference process. In an autoregressive model, each newly generated token (that is, a discretized glucose measurement) depends on all previously observed or generated tokens. Below, we describe this inference procedure in detail.

**Context and initialization.** To begin generation, we selected an initial context of observed CGM data (for example, the first day of glucose readings for a participant) or, optionally, a shorter warm-up window (for example, a few hours).

These observed glucose tokens were fed into the trained GluFormer model to provide the model with individual-level or population-level glycaemic patterns as a starting point.

**Next-token distribution.** Given the context tokens, the model produced, for each future time step, a distribution over all possible next glucose tokens. This distribution was effectively a probability vector across the entire glucose vocabulary (for example, 460 bins spanning 40–500 mg $dl^{-1}$). Because GluFormer is trained with a causal (autoregressive) mask, it uses only historical (context) tokens, not any future or current masked tokens, to compute the next-token probabilities.

**Sampling the next glucose value.** Rather than taking the single most likely token (that is, argmax decoding), we sampled from the predicted distribution of the model. Sampling introduces variability that can produce more realistic, person-specific glucose trajectories.

For instance, if the model assigns 10% probability to a mild hyperglycaemic bin and 90% probability to a euglycaemic bin, the sampling procedure will usually pick the euglyacemic bin but, occasionally, will sample the hyperglycaemic bin. This stochasticity helped capture the inherent uncertainty and biological variability of daily glucose fluctuations.

**Iterative (autoregressive) generation.** The newly sampled glucose token was then appended to the sequence to become part of the context for generating the subsequent token.

This process was repeated, step by step, until the desired generation length was reached, typically a full day (for example, 24 h at 15-min intervals, which produces 96 tokens). Each newly generated glucose token was conditioned on all preceding observed or generated tokens, thereby creating a coherent time series.

Through this procedure, GluFormer can produce realistic, subject-specific CGM trajectories that reflect known (or proposed) glucose dynamics. When contextual information (for example, previous glucose patterns or dietary intake) is provided, the autoregressive approach leverages the model's learned representation of glycaemic physiology to generate sequences that mirror individual variability and daily fluctuations in glucose levels.

### Analysis details

In Fig. 2a, we illustrate GluFormer's day-by-day generation on three representative participants from the HPP test set. For each participant, we took two consecutive days of CGM measurements: the first day (black curve) served as context input to the model, whereas the second day was held out for comparison. We then generated three independent synthetic day-2 trajectories (each with a different random seed) to demonstrate how small sampling variations can lead to slightly different, yet realistic, glucose curves (coloured lines). To gauge clinical similarity, we computed a set of CGM-based composite metrics (mean, c.v., GMI, time in range and time above 180 mg $dl^{-1}$) on both real and generated data for the second day, plotting them side-by-side in the radar charts. These three examples highlight how GluFormer preserved day-to-day glucose dynamics and produced synthetic CGM patterns that closely matched actual measurements.

In Fig. 2b, moving beyond individual examples, we systematically generated an unseen final day of CGM for each participant in the HPP test set, using all other available days as context. To quantify similarity, each scatter plot compares observed versus generated composite

scores (below 70 (%), in range 70–180 (%), mean, s.d., c.v. and GMI) for each participant's final day. Each point therefore represents one real-versus-synthetic comparison. We repeated generation three times per participant to account for sampling variability, then averaged the synthetic metrics to form a single data point. The strong Pearson's correlations (≥0.75, $P < 0.001$) confirmed that GluFormer reproduces key glycaemic indicators with high fidelity, and illustrated its capacity to generate clinically consistent CGM time series.

Last, in Fig. 2c, we assessed the out-of-distribution robustness of the model by applying the same generative strategy to ten additional cohorts, each with different populations (for example, type 2 diabetes or gestational diabetes) or device types. As in Fig. 2b, we provided all but the last day of CGM as context, generated the final day, then calculated each iglu-based metric on real versus generated data. For each cohort, we display Pearson's correlations across six key metrics (below 70 (%), in range 70–180 (%), s.d., mean, c.v. and GMI). The consistently high correlations, even in cohorts with markedly different glycaemic profiles, underscored the broad generalizability of the model and confirmed that GluFormer can accurately simulate diverse CGM dynamics across varying clinical conditions.

### Statistical analysis of generation

To evaluate the quality of generated samples from GluFormer, we used an autoregressive strategy to produce full days of CGM data on unseen test participants and then compared these generated time series to the true CGM data collected for those same days.

**Glucose-only GluFormer variant.** For the glucose-only version of GluFormer (that is, the model without diet tokens), we provided one full day of an individual's CGM readings ($4 \times 24 = 96$ measurements, sampled every 15 min) as context. The model then inferred the subsequent day's glucose values in an autoregressive manner, predicting each next glucose token based only on previously generated or observed tokens (while masking future tokens).

To assess generation quality, we computed a variety of standard CGM metrics (for example, mean glucose, percentage time in range and c.v.) using the iglu package in R on both the generated and actual CGM days. We then compared these metrics by calculating Pearson's correlations between the observed and generated composite scores.

We also computed the Pearson's correlation and MAE between the observed and generated glucose time series (that is, data points at each 15-min interval).

We report two-tailed $P$ values for each Pearson's correlation using a $t$-test to determine significance.

**Multimodal GluFormer variant (with diet tokens).** For the multimodal GluFormer model, which integrates both glucose and diet tokens, we provided one full day of CGM plus corresponding diet tokens as context. The model then autoregressively generated the subsequent day's CGM values. Whenever the participant ate, we inserted the relevant diet tokens (that is, macronutrient bins) into the generated sequence at the appropriate time.

After generation, we removed the diet tokens in the output to focus on the generated glucose sequence. The same set of evaluation metrics was used: Pearson's correlation and MAE between the observed and generated glucose time series.

**Context-length experiment.** To investigate how the number of CGM days provided as context affects the quality of generated data, we conducted an experiment (Supplementary Figs. 3 and 4) whereby we varied the amount of CGM context (ranging from a single token up to multiple days). For each context length, the model generated a new day of CGM data, and we again calculated Pearson's correlations between observed and generated CGM scores or the raw time series. We observed that more context about an individual's previous CGM data improved the generation fidelity (that is, produced higher correlations and lower error between observed versus generated curves).

All $P$ values for correlations were determined using two-tailed $t$-tests, and the MAE was similarly evaluated as an additional measure of discrepancy between observed and generated glucose signals.

### Optimization and training details

For optimizing the GluFormer + diet model, we used a carefully configured setup designed to promote stable learning while effectively minimizing the loss over time. A detailed overview of the optimization parameters and training conditions is provided below.

AdamW, a variant of the Adam optimizer that incorporates weight decay, was selected for its robustness and effectiveness in managing sparse gradients and complex dependencies in high-dimensional data.

Set at 0.0001, the relatively low learning rate helped prevent the model from overshooting minima, thereby allowing finer adjustments in weight updates.

For beta coefficients (b1 and b2), values of 0.9 and 0.99 for the first and second moment estimates, respectively, balanced the trade-off between momentum and stability.

For weight decay, a value of 0.01 was used to regularize and prevent overfitting, which was particularly useful in complex models with a high capacity, such as this transformer.

For learning rate scheduler, a gamma of 0.0001 was used, and the the learning rate decayed by this factor, gradually reducing the step size to fine-tune the model's weights as training progressed.

### Training dynamics

The model was trained for 20 epochs, which provided sufficient time for the complex patterns in the dataset to be learned without leading to excessive training times. Set at 0.2, dropout was used as a regularization method to prevent overfitting by randomly dropping units (along with their connections) during the training process. A total of 16 samples per batch were used.

### Hardware and computational details

For GPU, training used a single NVIDIA A100, known for its powerful computational capabilities and efficiency in handling large datasets and complex models.

Training was conducted using 16-bit floating point precision (FP16), which reduces memory consumption and can speed up training without significantly affecting the accuracy of the results.

A maximum gradient norm (max_grad_norm) of 1 was used to prevent the exploding gradient problem, which can lead to destabilized training dynamics.

A consistent numpy, torch and random math seed of 42 was set to ensure reproducibility of the results, thereby enabling consistent initialization and stochastic processes across different runs.

The model used 100 warm-up steps at the start of training. This gradual ramp-up in learning rate helped stabilize the parameters of the model before entering the full training regime.

### Comparative analysis setup using GluFormer with and without diet

To rigorously evaluate the impact of integrating dietary data into the GluFormer model, we conducted a comparative analysis between two versions of the model: one incorporating diet data (GluFormer + diet) and the other excluding it (GluFormer). This comparison aimed to assess the added value of dietary information in enhancing the predictive accuracy and understanding of glucose dynamics of the model.

### Data consistency across models

**Unified preprocessing pipeline.** Both versions of the GluFormer model were trained using the same preprocessing pipeline detailed above for the GluFormer and GluFormer + diet models. This ensured

that any observed differences in model performance can be directly attributed to the inclusion of dietary data rather than discrepancies in data handling or preprocessing.

**Identical datasets.** To ensure a fair comparison, the training, validation and test sets were identical across both model variants. This approach eliminated variability in data distribution as a confounding factor, which enabled a clear assessment of the impact of diet data integration.

Hyperparameter searches for these models used the following values: beta coefficients (b1 and b2): b1, {0.85, 0.88, 0.9} and b2, {0.9, 0.95, 0.99}; diet modelling (diet_modeling) values, {true, false} (determines whether dietary data are incorporated into the model); dimension of feedforward network (dim_feedforward): {512, 2,048}; dropout rate (dropout): {0, 0.1, 0.2}; number of training epochs (epochs): {10, 15, 20, 30}; learning rate decay: {0.1, 0.01, 0.001} (specifies the ratio between starting learning rate (after warm up) and ending learning rate; learning rate: {1e-06, 1e-05, 0.0001, 0.0005, 0.001}; maximum gradient (norm): {1, 10, 100}; embedding size (n_embd): {256, 512, 1,024} (size of each token embedding); number of attention heads (n_heads): {8, 16}; number of transformer layers (n_layers): {8, 10, 12, 16}; warm-up steps: {10, 100, 1,000} (number of steps to increase the learning rate at the beginning of training; and weight decay: {0, 0.001, 0.01, 0.1}.

**Code and packages.** Python (v.3.10) was used with the following packages: torch (v.2.3.1), torchaudio (v.2.3.1), torchelastic (v.0.2.2), torchvision (v.0.18.1), tqdm (v.4.66.4), numpy (v.1.26.4), pandas (v.2.2.2), scikit-learn (v.1.5.0), scipy (v.1.14.0), seaborn (v.0.13.2), matplotlib-inline (v.0.1.6), matplotlib (v.3.9.0), wandb (v.0.17.3), umap-learn (v.0.5.5) and dcurves (v.1.1.1). Iglu (v.4.2.1) was calculated in R (v.4.5.1).

### Ethics approval

All participants signed an informed consent form after arrival to the research site. All identifying details of the participants were removed before the computational analysis. The 10K cohort study was conducted according to the principles of the Declaration of Helsinki and was approved by the Institutional Review Board of the Weizmann Institute of Science. ClinicalTrials.gov registration identifier: NCT05817734.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The data used in this paper are part of the HPP and are accessible to researchers from universities and other research institutions (https://humanphenotypeproject.org/data-access). Interested bona fide researchers should contact info@pheno.ai to obtain instructions for accessing the data. Deidentified participant data from the AEGIS study will be made available upon publication through the Runa Digital Repository (runa.sergas.gal). Access will require a signed data access agreement, and proposals should be directed to F.G.

## Code availability

Implementation of GluFormer is available at GitHub (https://github.com/Guylu/GluFormer).


45. Vaswani, A. et al. Attention is all you need. In *Adv. Neural Information Processing Systems* (eds Guyon, I. et al.) **30**, 5998–6008 (2017).
46. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *Proc. 7th International Conference on Learning Representations* https://openreview.net/forum?id=Bkg6RiCqY7 (2019).
47. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *Proc. 37th International Conference on Machine Learning* (eds Daumé, H. D. III & Singh, A.) **119**, 1597–1607 (2020).
48. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9729–9738 (2020).
49. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning* (eds. Meila, M. & Zhang, T.) **139**, 8748–8763 (2021).
50. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds. Burstein, J. et al.) Vol. 1 (Long and Short Papers), 4171–4186 (2019).
51. Yuan, H. et al. Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality. *npj Digit. Med.* **7**, 86 (2024).



**Acknowledgements** We thank all members of the Segal Laboratory, the Pheno.AI data science and NVIDIA Tel Aviv Research groups, and E. Barkan and A. Shocher for discussions. J.M. was supported by Novo Nordisk Foundation grant NNF23SA0084103, an EFSD/Novo Nordisk Foundation Future Leaders Award (no. 0094134) and the European Union (HORIZON-EIC-2023-PATHFINDERCHALLENGES-01-101161509). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.


**Author contributions** G.L. conceived the project, designed and conducted all analyses, interpreted the results and wrote the manuscript. G.S. developed protocols, interpreted the results and wrote the manuscript. A.G. designed pipelines and created preprocessing scripts. S.S. interpreted the results and wrote the manuscript. J.R.G. and D.S.-B. acquired the PREDICT cohort data. R.D. interpreted the results and wrote the manuscript. F.G. wrote the manuscript. J.M. interpreted the results and wrote the manuscript. S.M. guided computational analyses. E.M. guided computational analyses and managed code-running infrastructure. E.P.X. interpreted the results and wrote the manuscript. G.C. interpreted the results and wrote the manuscript, and directed the project. H.R. conceived and directed the project and analyses, designed the analyses, interpreted the results and wrote the manuscript. E.S. conceived and supervised the project and analyses, designed the analyses, interpreted the results and wrote the manuscript.

**Competing interests** G.S. and H.R. are employees in Pheno.AI, a biomedical data science company from Tel-Aviv, Israel. E.S. is a paid consultant of Pheno.AI. G.L.'s work was done during an internship at NVIDIA Research. The other authors declare no competing interests.

**Additional information**


**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-09925-9.



**Correspondence and requests for materials** should be addressed to Hagai Rossman or Eran Segal.



**Peer review information** *Nature* thanks Jessilyn Dunn and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Eran Segal

Last updated by author(s): 21/05/2025

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed |
|---|---|
| ☐ | ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ A description of all covariates tested |
| ☐ | ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

Data collection: No data collection software was used

| | |
|---|---|
| Data analysis | Python 3.10, with packages:<br>torch==2.3.1<br>torchaudio==2.3.1<br>torchelastic==0.2.2<br>torchvision==0.18.1<br>tqdm==4.66.4<br>numpy==1.26.4<br>pandas==2.2.2<br>scikit-learn==1.5.0<br>scipy==1.14.0<br>seaborn==0.13.2<br>matplotlib-inline==0.1.6<br>matplotlib==3.9.0<br>wandb==0.17.3<br>umap-learn==0.5.5<br>dcurves==1.1.1<br>R version 4.5.1<br>Iglu version 4.2.1 |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data used in this paper is part of the Human Phenotype Project (HPP) and is accessible to researchers from universities and other research institutions at: https://humanphenotypeproject.org/data-access. Interested bona fide researchers should contact info@pheno.ai to obtain instructions for accessing the data.

## Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

| | |
|---|---|
| Reporting on sex and gender | Participants' sex was determined based on self-reporting. Participants were not asked about their gender. Study population included 5837 women and 4941 men. 34 did not answer when asked about their sex. |
| Reporting on race, ethnicity, or other socially relevant groupings | Participants included in this study are of all the participants who met the inclusion criteria detailed in Shilo et al. (2021). No analysis based on race, etchnicity, etc.. was conducted in this study. |
| Population characteristics | Samples analyzed in this study were collected as part of the Human Phenotype Project (HPP), described in details in: Shilo, Smadar, et al. "10K: a large scale prospective longitudinal study in Israel" European journal of epidemiology 36.11 (2021): 1187-1194. The 13,018 participants with detailed data have a mean age of 51.4±8.5 years (males: 50.8±8.8; females: 51.0 ±9.1) and a sex distribution of 53.6% male to 46.4% female. The cohort is ethnically diverse, including 6,270 Ashkenazi, 943 North African, 780 Middle Eastern, 264 Sephardi, 249 Yemenite, and 2,837 participants of mixed ancestry. |
| Recruitment | The recruitment process relied on self-assignment of volunteers who register to the 10K trial website (https://www.project10k.org.il/en) |
| Ethics oversight | The 10K cohort is conducted according to the principles of the Declaration of Helsinki and was approved by the Institutional Review Board (IRB) of the Weizmann Institute of Science. |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | This studies goal was to build a deep learning model using the HPP dataset, thus no prior knowledge on the expected effect sizes exists, or sample size calculation was done. Samples included in this study are of all the participants who met the inclusion criteria detailed in Shilo et al. (2021). The results reported in the manuscript are those that passed the statistical significance threshold. |
| Data exclusions | Participants included in this study are of all the participants who met the inclusion criteria detailed in Shilo et al. (2021). |
| Replication | A lot of the associations found in this study were also observed in previous works as referenced and detailed in the Discussion section of the manuscript. In addition, results were replicated on many external cohorts, as mentioned in the manuscript. |
| Randomization | There was no group allocation in this study. |
| Blinding | There was no group allocation in this study. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☐ | ☒ Clinical data |
| ☒ | ☐ Dual use research of concern |
| ☒ | ☐ Plants |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

| | |
|---|---|
| Clinical trial registration | NCT05817734. All other NCTS of cohorts used in this study can be found in page 47 of the paper in Table 1. |
| Study protocol | The full description of the study can be found in: Shilo, Smadar, et al. "10K: a large scale prospective longitudinal study in Israel" European journal of epidemiology 36.11 (2021): 1187-1194. |
| Data collection | The trial is being held in Rehovot, Israel. The study started on 2018-10-01 and is expected to be completed on 2045-12-31. |
| Outcomes | Development of medical conditions based on participant's self-reporting coded to ICD11. Spesific outcomes for each study are mentioned in NCTs in Table 1 of the paper. |

## Plants

| | |
|---|---|
| Seed stocks | No seed stocks or other plant material was used in this study. |
| Novel plant genotypes | No seed stocks or other plant material was used in this study. |
| Authentication | No seed stocks or other plant material was used in this study. |