Integrating Polygenic Risk Improves Generative Forecasting of Disease Trajectories

Chris German[1], Suyash Shringarpure[1], Payam Dibaeinia[1], James Ashenhurst[1], Bertram L. Koelsch[1], Adam Auton[1], and Aly A. Khan[1,2]

[1] *23andMe, Inc., Palo Alto, CA, USA*
[2] *Departments of Family Medicine, and Pathology, and Institute for Population and Precision Health, University of Chicago, Chicago, IL, USA*
E-mail: chrisg@23andme.com; aakhan@uchicago.edu

Predicting the longitudinal sequence of diseases an individual will develop over their lifetime is a central challenge in medicine. While recent AI models can process health histories, they have been limited by cohort size and the omission of genetic data. Here we introduce the Next Health Event (NHE) model, a generative transformer trained on the health trajectories of 7.1 million research participants. By using a transformer architecture to integrate demographic data, longitudinal BMI, and polygenic risk scores (PRS) for 297 traits with sequential health history, NHE significantly outperforms baseline models, including XGBoost with the same inputs, in predicting the next diagnosis across 129 conditions (Top-1 accuracy 25.5% vs. 22.3%). Systematic ablation studies reveal that both PRS and longitudinal BMI provide substantial, non-redundant predictive power, whereas self-reported lifestyle information offers limited additional value. The model's predictive accuracy is the same when forecasting prospectively reported incident outcomes vs. combined prospectively and retrospectively reported outcomes (AUROC 0.917), demonstrating its utility for real-world risk assessment. By uniting large-scale health histories with genetics, our work establishes a new framework for predictive health and demonstrates that generative models can effectively forecast individual disease pathways.

Keywords: Disease Prediction, Polygenic Risk Score, Longitudinal Data, BMI.

1. Introduction

Predicting an individual's lifetime health trajectory is a central goal of personalized medicine. These trajectories can be represented as ordered sequences of clinical events, characterized by complex temporal dependencies and patterns of comorbidity. While traditional machine learning can predict risk for single diseases, it often fails to capture this rich longitudinal structure. Recent advances in natural language processing, particularly transformer-based large language models (LLMs), provide a powerful framework for overcoming this limitation by treating health histories as sequences of clinical events, and integrating multimodal context such as genetic information.

The application of LLMs to health data is rapidly maturing, with distinct methodological

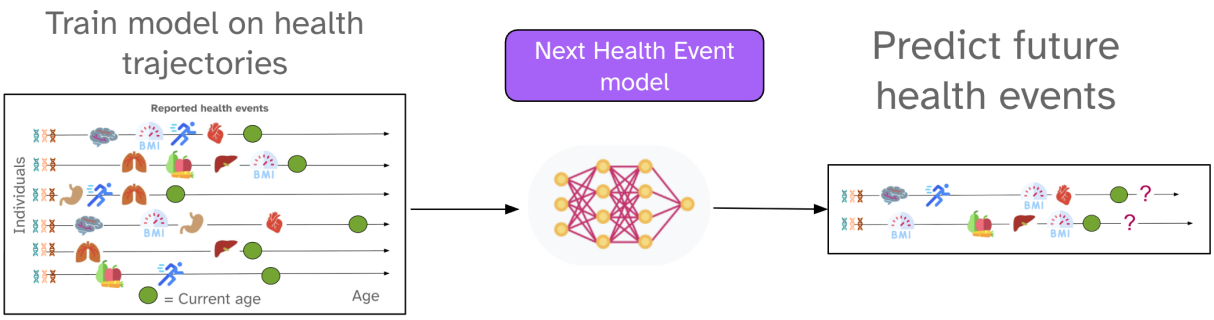

Fig. 1. The Next Health Event (NHE) model framework for forecasting disease trajectories. The model trains on multi-modal health histories (diagnoses, genetics, BMI) from over 7 million individuals (left) using a generative transformer (center) to predict an individual's future health events based on their current history (right).

approaches demonstrating success. Some models focus on multimodal data integration; for instance, HeLM combines structured clinical data with high-dimensional inputs like spirograms, while PH-LLM leverages time-series data from wearables to generate personalized health insights.[1,2] Another class of models learns deep representations from medical terminologies themselves. For example, GRASP utilizes a masked language modeling approach on structured Electronic Health Record (EHR) data to create robust embeddings of medical codes, and life2vec models an individual's entire life course, including health, social, and economic events, to predict outcomes like mortality.[3,4]

Most recently, generative pre-trained transformer (GPT) architectures have been adapted to directly forecast future health events. Foresight models clinical histories from structured data and clinical notes to predict subsequent medical outcomes, while Delphi-2M, a GPT-inspired model trained on the UK Biobank cohort, has shown strong performance in predicting over 1,000 distinct diagnoses.[5,6] Notably, the superior accuracy of a domain-specific model like Delphi-2M compared to zero-shot predictions from a general-purpose LLM underscores the necessity of specialized training on health records.

Despite this rapid progress, two critical gaps limit the current state-of-the-art. First, and most critically, existing models have not integrated an individual's innate genetic risk, a fundamental determinant of health, as a core input. Prior work has shown accounting for genetic risk improves prediction beyond clinical factors alone.[7–9] Second, these generative models have mostly been trained on datasets of fewer than one million individuals, which restricts their power to model rare diseases and to accurately dissect the complex interplay between genetic predispositions and environmental factors.

To bridge these gaps, we introduce the Next Health Event (NHE) model, a generative transformer trained and evaluated on a large cohort of over 7 million research participants from the 23andMe Research Cohort (Figure 1).[10] 23andMe, Inc. is a consumer genetics and research company where customers are offered the opportunity to consent to participate in research studies via online surveys. The key innovation of NHE is its ability to create a unified

representation of an individual by integrating their unique sequence of diagnosed health conditions with demographics, polygenic risk scores (PRS) for 297 traits, longitudinal body mass index (BMI) measurements, and lifestyle survey data. This scale and multimodal approach provide a critical testbed directly relevant to major population-scale health initiatives, such as Our Future Health and All of Us, which are building cohorts of over one million people.

We demonstrate through comprehensive ablation studies and analyses of prospectively collected data that the NHE model establishes a new state-of-the-art in health forecasting. Our analyses show that: (1) the generative sequence model significantly outperforms strong non-sequential baselines (e.g., XGBoost); (2) integrating genetic risk and longitudinal BMI provides substantial and non-redundant contributions to predictive accuracy across a wide range of diseases; and (3) the model generalizes effectively, maintaining high performance when predicting prospectively collected, incident health outcomes. By successfully fusing deep genetic and longitudinal phenotypic data at population scale, NHE provides a powerful and validated blueprint for the next generation of personalized health models.

2. Methods

2.1. *Study cohort*

Research participants from 23andMe, Inc., a consumer genetics and research company, were used to train and evaluate the Next Health Event model. These participants were genotyped as part of the 23andMe Personal Genome Service. All participants included in the analysis provided informed consent and volunteered to participate in online research under a protocol approved by Ethical & Independent (E&I) Review Services, an external AAHRPP-accredited IRB. As of 2022, E&I Review Services is part of Salus IRB (https://www.versiticlinicaltrials.org/salusirb).

2.2. *Data details*

Data were collected via self-reported online surveys. A baseline "Health Profile" survey captured each participant's health history, including diagnosed conditions and age of onset. Age of onset information was collected on 129 conditions. The 129 conditions were selected based on a combination of factors, including their prevalence within the research cohort, their clinical significance, and the availability of reliable age-of-onset data

Since 2016, a longitudinal"Health Update" survey was administered every 1-2 years to capture incident diagnoses. Information on incident diseases from follow-up surveys was available for 65 of the 129 conditions. Individuals were included in the analysis if they reported at least one diagnosed condition. The model integrated the following data modalities:

- **Health History:** Sequences of diagnosed conditions and their corresponding ages of onset were constructed for 129 conditions. If multiple conditions were reported during the same age, the ordering was randomly assigned so the model did not learn any specific ordering.
- **Demographics:** Self-reported birth year, sex, and genetically determined ancestry.
- **Longitudinal BMI:** Body mass index (BMI) was measured at multiple time points, discretized into deciles, and integrated into the health sequence.

- **Genetic Risk:** 297 polygenic risk scores (PRS) were calculated for traits related to the modeled conditions. Individuals were classified as high-risk (top 5%) or low-risk (bottom 5%) relative to a genetic ancestry-matched distribution. Individuals that fell within the middle 90% of the distribution for any single PRS had no classification represented in the model for that specific PRS. 36/297 PRSs were trained using data from the 23andMe research cohort as detailed below. The remaining PRSs were taken from the PGSCatalog.
- **Lifestyle:** Responses from a lifestyle survey (e.g., diet, sleep, activity) were encoded as tokens and assigned to the participant's age at the time of survey completion. The lifestyle survey is only surfaced to participants after the Health Profile survey is completed.

BMI was discretized into deciles of the age-and-sex-matched distribution to provide a non-parametric, normalized representation of weight status that is robust to population-level shifts over time. This approach was chosen for its simplicity and effectiveness in preliminary experiments. Simialrly, PRS were encoded as binary high-risk (top 5%) and low-risk (bottom 5%) indicators to create distinct, high-impact tokens that the model could easily differentiate. While this approach simplifies the continuous nature of genetic risk, it effectively captures individuals at the extremes of the genetic liability spectrum where risk is most pronounced, and reduces model complexity.

The 23andMe PRSs were trained excluding the individuals in the NHE validation and test sets. We used a cross-validation approach to train these PRSs where we split all European individuals in the training cohort into 5 folds. PRSs were estimated for each fold by performing GWAS on the other 4 folds and running LDPred2[11] to estimate the PRS on the left out fold. This approach was done to reduce overfitting to the training set. After PRSs were estimated, we standardized each PRS within each fold with mean 0 and variance 1 to transform them the same scale. The remaining 261 PRSs were taken from the PGSCatalog.[12,13] In total, data from 7,184,380 individuals were used, split into a training set (5,747,786), a validation set (718,041), and a held-out test set (718,553).

2.3. *Model details and training*

We adapted the generative pretrained transformer (GPT-2) architecture for the NHE model.[14] We selected a GPT-2 architecture due to its autoregressive nature, which enables the model to learn conditional dependencies over time and predict future health events by sequentially conditioning on an individual's prior history. Each age, condition, PRS, BMI decile, unique value of demographic variables, and lifestyle response included in the model prompt was encoded into the model vocabulary as distinct tokens. It has an embedding size of 288, 12 attention heads, 12 layers, and a maximum sequence length of 1024. The total number of parameters in the model including the language modeling head is 13,071,744. It was trained from scratch on 8 NVIDIA A100 GPUs for up to 50 epochs, using early stopping based on validation set performance. With this setup the model training time is under 3 days. This relatively compact architecture was chosen to ensure computational efficiency and facilitate reproducibility, demonstrating that powerful predictive performance can be achieved without requiring massive-scale computational resources.

2.4. *Model evaluations*

2.4.1. *Comparison with baselines*

We compared the NHE model to several baselines in a held-out test set to investigate whether we gain additional prediction accuracy by (1) including personalized health information as inputs (i.e., PRS and health history) and (2) using a language modeling framework. To assess (1), we compared the NHE model to a naive prevalence based estimator that predicts the top condition(s) in the training data as well as an age and sex-specific prevalence estimator that uses sex-matched condition rates in the training data ±2 years from the target prediction age. To assess (2) we compared the NHE model trained without BMI and lifestyle data (trained on demographic, PRS, and health history data) to an XGBoost model that has the same features. XGBoost was selected as a baseline due to its strong empirical performance in a variety of structured prediction tasks and its ability to capture non-linear interactions and handle missing data, making it a strong benchmark for assessing the value of a generative, sequence-based approach. Training of all baseline models took under an hour. The primary evaluation metric was the top-1 and top-5 accuracy in predicting the most recent health condition in an individual's history. Top-k accuracy measures how often the correct answer is among the top k predictions a model makes.

For the NHE model, we prompted the model with each person's health history in the test set omitting their last condition (and any following BMI values/lifestyle responses) to generate top candidates for the held-out condition. We also compared macro-/micro-AUROC estimates, per-condition AUROC, precision, and recall between the NHE and XGBoost models. We quantify discrimination with the micro-averaged area under the receiver-operating characteristic curve (micro-AUROC). Unless noted otherwise, 'AUROC' refers to this micro-averaged value, computed by pooling true-positive, false-positive, and false-negative counts across all conditions before drawing the ROC curve.

2.4.2. *Ablation studies*

To assess the contribution of different data modalities to the predictive performance of our model, we conducted ablation studies. In these experiments, we systematically removed one or more types of data from the model's training input and evaluated the resulting impact on its accuracy and other relevant metrics. The data modalities we investigated included polygenic risk scores (PRS), body mass index (BMI), lifestyle data (such as smoking status, physical activity levels, and dietary habits), and demographic information (including birth year, sex, and genetic ancestry). By comparing the model performance trained on the complete set of modalities to its performance after the exclusion of each individual modality, as well as combinations, we aimed to quantify the relative importance and potential redundancy of these different data sources in predicting the health events. This systematic removal and evaluation process allowed us to gain a deeper understanding of which types of information were most critical for achieving optimal model performance and to identify potential areas for future model refinement and data collection efforts.

Since BMI and lifestyle may be indirectly captured by other signals via health history (e.g.,

BMI is correlated with many metabolic conditions), we conducted a subanalysis on individuals without prior health conditions, examining model performance when specific data types were available. By examining model performance within this specific subset of individuals, we can better understand each modality's unique contributions.

We further investigated the impact that PRSs have on the model by comparing the full model to the model trained without PRS information, restricting to individuals whose held-out trait is one that had a direct match with a PRS in the model and who were either at elevated or decreased genetic risk for that trait (they were in the top or bottom 5% of the PRS distribution). No individuals in the NHE validation or test sets were included in the GWAS or PRS training datasets used to develop the 23andMe PRSs incorporated into the model. Because they are directly related to a trait of interest, the PRSs capture relevant risk information for the phenotypes, and these comparisons can highlight the importance of including genetic risk information into the model.

2.4.3. *Incident data prediction*

The majority of the data being trained on is retrospective in nature - research participants fill out a health profile survey that asks them questions about prior conditions they have been diagnosed with since birth. Some participants have completed follow-up longitudinal health update surveys that collect incident data, asking about recent diagnoses. The model is trained on the combination of these data types. We tested the model's ability to specifically predict incident data, which is available on 65 out of 129 phenotypes.

In one evaluation, we used the previous task of holding out the last condition and having the model predict it. We stratified by if the condition was coming from a retrospective survey or a prospective survey, and compared top-1 and top-5 accuracies.

In the other evaluation we used the individual's prompt prior to any incident data as input into the model, and estimated approximate probabilities of each condition occurring over 5 years since completion of the health profile survey. The approximate probability is defined as:

$$\hat{P}_c^{(k)} := \sum_{i=1}^{k} P(a = \alpha + i|\pi) \cdot P(c|\pi, a = \alpha + i)$$

where:

- $\hat{P}_c^{(k)}$ is the estimated probability of condition c occurring within k years
- α is the age of the individual and π is the prompt (e.g., demographic and PRS information, BMI, lifestyle, and health history prior to the follow-up survey)
- $P(a = \alpha + i|\pi)$ is the model's predicted probability that the next age is $\alpha + i$ given prompt π
- $P(c|\pi, a = \alpha + i)$ is the model's predicted probability of condition c given prompt π and the next age is $\alpha + i$

We used these probabilities to calculate AUROC across the 65 incident phenotypes, and compared it to the AUROC coming from the held-out last token task from the combined data (retrospective and prospective) restricted to the 65 phenotypes. Individuals who reported no new conditions in the health update surveys were kept in this analysis, as it more accurately

reflects the target population.

3. Results

3.1. *NHE Model Outperforms Baselines in Disease Prediction*

The NHE model significantly outperformed all baselines, including population-based prevalence estimators and a powerful XGBoost classifier. In a held-out test set, the NHE model achieved a top-1 accuracy of 25.5% for predicting the next diagnosis, compared to 22.3% for the strongest baseline, XGBoost (Figure 2a). This performance advantage extended to top-5 accuracy (56.1% vs. 52.2%). The generative model also demonstrated superior discriminative ability, yielding a higher micro-averaged AUROC of 0.938 compared to 0.926 for XGBoost (Figure 2b). Macro-averaged AUROC, and per-condition AUROC, precision, and recall were also consistently higher for the NHE model (Figures 2c–f). Overall, across the 129 conditions in the model, the NHE model had significantly higher accuracy compared to all baselines.

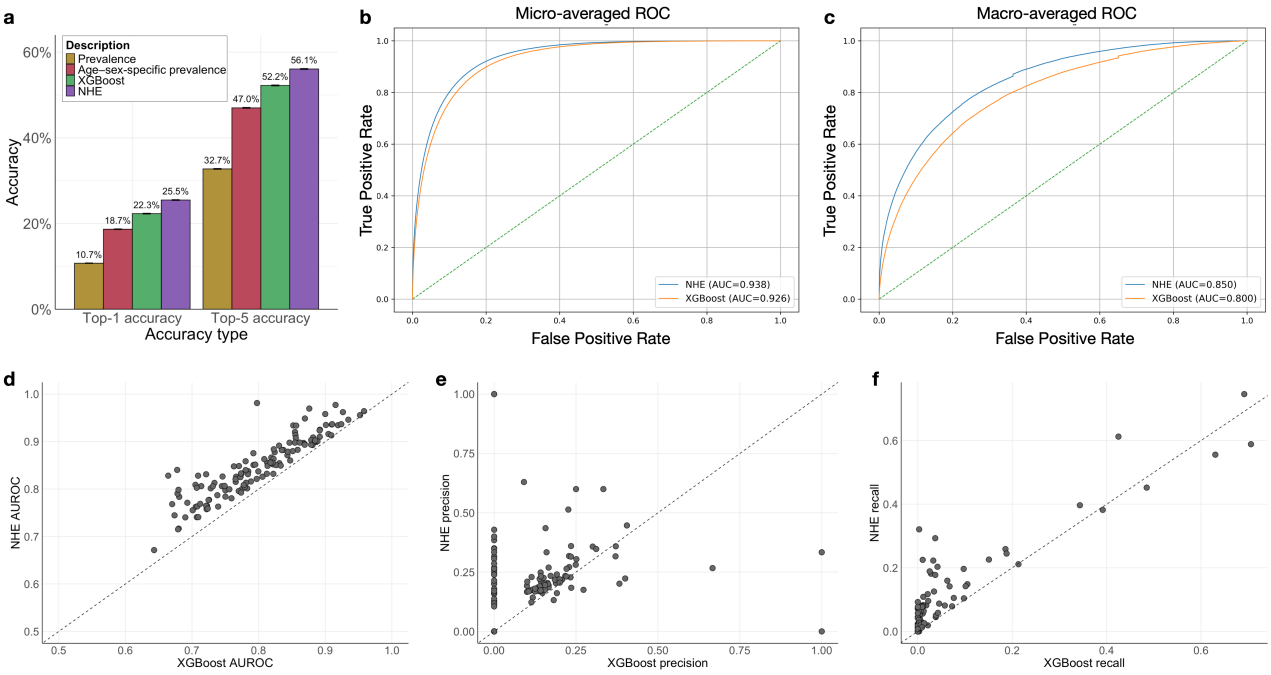

Fig. 2. The NHE model significantly outperforms baselines in next-disease prediction. Performance on a held-out test set (N=718,553) comparing NHE to XGBoost and prevalence estimators. (a) NHE achieves the highest Top-1 and Top-5 accuracy. (b-f) The model also demonstrates superior discriminative ability, with consistently higher AUROC, precision, and recall over XGBoost, highlighting the value of its sequence-aware architecture. Error bars indicate 95% confidence intervals.

3.2. *Genetic Risk and Longitudinal BMI Provide Non-Redundant Predictive Power*

To evaluate the contribution of each data modality, we performed ablation studies in which specific input types were systematically removed from the model. The NHE model incorpo-

rating all information and the model with lifestyle data removed achieved the highest top-1 and top-5 prediction accuracies. There was no significant difference between the model with all information and the one with lifestyle data removed, suggesting lifestyle data does not offer additional predictive accuracy when accounting for BMI, PRS, and demographic information. Accuracy decreased upon the removal of any of the other single data modalities or combinations thereof, indicating that each of those data types provide non-redundant contributions to the model's predictive performance (Figure 3a). The model using only health history information, "No demographics, PRS, BMI, lifestyle" had the lowest performance.

In our stratified analyses of individuals with only a single reported condition (i.e., no prior health conditions in the prompt), we found that excluding BMI and lifestyle data from training significantly reduced performance compared to the full model (top-1 accuracy: 21.1% vs 19.4%, $P < 0.01$; Figure 3b). However, when comparing the full model to a version trained only without lifestyle data (while retaining BMI), there was no significant difference in performance among individuals who had lifestyle data available ($P = 0.62$).

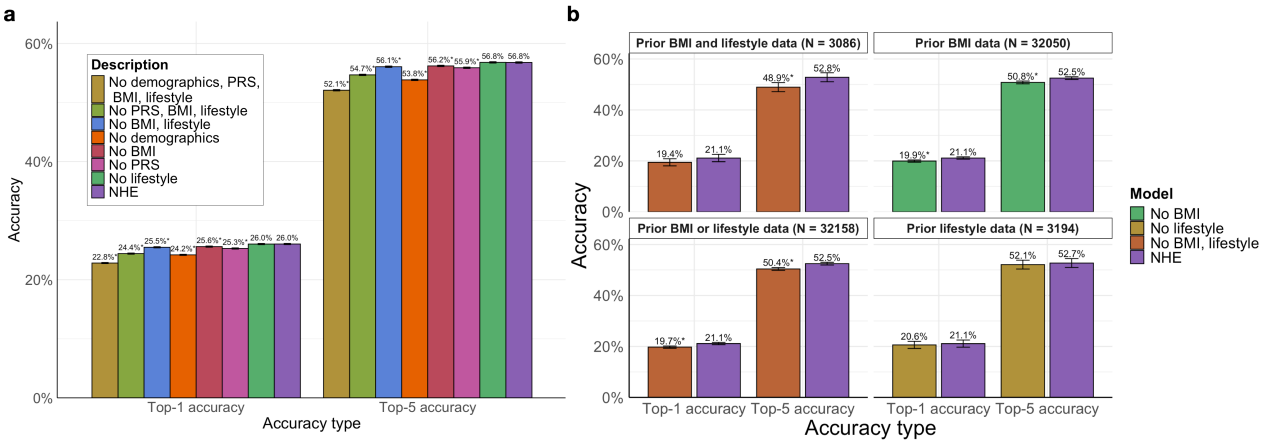

Fig. 3. Genetic risk (PRS) and longitudinal BMI provide non-redundant predictive power. Ablation studies on the test set (N=718,553) quantify the contribution of each data modality. (a) Removing PRS or BMI significantly degrades model accuracy, while removing lifestyle data does not. (b) In a stratified analysis of individuals with no prior health conditions, BMI and PRS are shown to be critical for an initial diagnosis. Asterisks (*) denote a significant ($P < 0.05$) accuracy drop compared to the full model.

For individuals with a held-out trait that matched one of the included PRSs and who were at either elevated or reduced genetic risk for that same trait, the full NHE model outperformed the version trained without PRS with higher top-1 accuracy (34.2% vs 28.2%, $P <$ 2e-16) and top-5 accuracy (65.5% vs 59.0%, $P <$ 2e-16), showing that genetic risk information meaningfully improves the model's predictive ability (Figure 4).

The ablation studies demonstrate that demographics, PRS, and BMI contribute to the NHE model's overall performance. The importance of each modality varies depending on an individual's health history and the specific condition being predicted. Incorporating PRS data improves prediction accuracy for traits with available PRS data. These findings underscore

the value of leveraging diverse data modalities for developing comprehensive and accurate health prediction models.

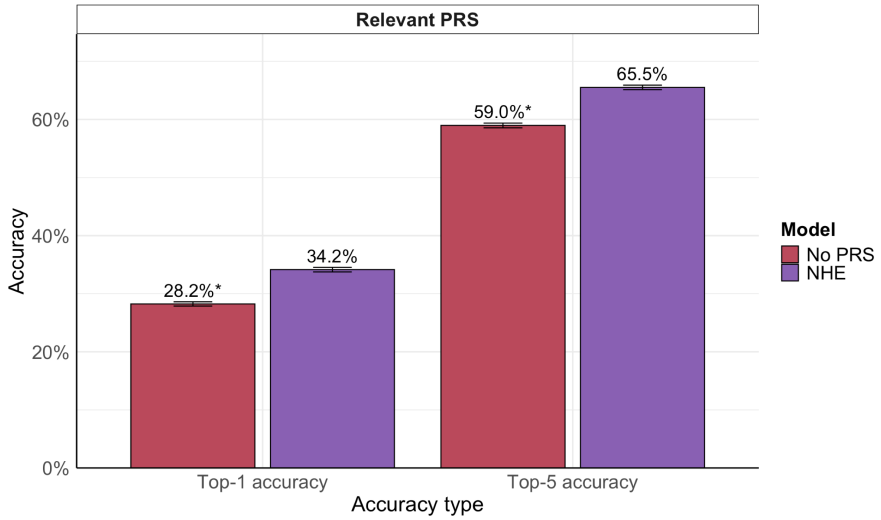

Fig. 4. Integrating Polygenic Risk Scores provides a dramatic boost in predictive accuracy for relevant traits. Analysis of individuals (N=56,519) with a condition matching their high/low genetic risk profile. The full NHE model (with PRS) substantially outperforms a model without PRS in both Top-1 (34.2% vs 28.2%) and Top-5 (65.5% vs 59.0%) accuracy, demonstrating the critical value of genetic data. Error bars indicate 95% confidence intervals.

3.3. *Model Generalizes to Prospectively Reported Incident Outcomes*

To assess real-world utility, we tested the model's ability to predict prospectively reported diagnoses. We evaluated its performance on incident outcomes – conditions newly reported by participants in follow-up health update surveys, surfaced every 1-2 years after the initial health profile survey. The NHE model had strong generalization to prospective outcomes. In predicting the held-out last condition task, the accuracy was similar regardless of whether the condition was coming from the retrospective health profile survey or the prospective health update survey, with only slightly decreased accuracy (top-1 accuracy 24.1% vs 26.1%, top-5 accuracy 57.3% vs 58.9%).

In the separate evaluation using the 5-year probability estimates derived from model outputs prior to any incident data, performance, measured by AUROC, was comparable when restricted to the 65 phenotypes with available incident data and evaluated on either prospective outcomes alone or on the combined dataset (0.917 for both). These findings indicate that the model is able to learn patterns from a mix of retrospective and prospective data and effectively generalize to future risk prediction.

4. Discussion

By integrating genetics with longitudinal health records at population scale, the Next Health Event (NHE) model establishes a new framework for predicting individual disease trajecto-

ries. Trained on a dataset of over 7 million individuals, our generative model learns complex patterns of comorbidity and age of onset to outperform strong baselines. More importantly, our work quantifies the distinct contributions of multiple data modalities, providing critical insights for the development of future predictive health tools.

Our findings confirm the power of transformer architectures for modeling sequential health data, aligning with previous work such as Delphi-2M and Foresight. However, NHE addresses a critical gap in the field by being the first, to our knowledge, to integrate an individual's innate genetic risk via polygenic risk scores (PRS) as a core input. The resulting enhancement in predictive accuracy, particularly for genetically-linked conditions, underscores the necessity of unifying genetics and health histories to create truly personalized predictive models.

A key finding from our work is the quantification of the non-redundant predictive power of both genetic risk and longitudinal BMI. Through systematic ablation, we demonstrated that each modality provides unique, valuable information. Conversely, self-reported lifestyle data offered minimal additional benefit, suggesting its predictive signal is likely captured by more proximal biological measures like BMI and the presence of related diagnoses in the health history. Our stratified analyses further refined these insights: BMI proved critical for risk assessment in individuals with sparse health histories, while PRS substantially improved predictions for those with known genetic predispositions. This reaffirms the value of multimodal data for creating robust and personalized clinical and preventive strategies.

A significant strength of our study is the demonstration of the model's prospective validity. The NHE model's high performance on incident outcomes, reported years after the initial data collection, shows that it learns generalizable patterns rather than simply memorizing historical sequences. This robust generalization to future events is a crucial step toward establishing real-world clinical utility for short-term risk forecasting.

Broader Implications and Potential Applications

Beyond its technical advancements, the NHE framework has significant implications for both clinical practice and biomedical research. For instance, the model could be used to inform personalized screening schedules, suggesting earlier or more frequent monitoring for individuals forecast to be at high risk for specific conditions like breast cancer or heart disease. It could also accelerate therapeutic development by enriching clinical trial cohorts, identifying individuals with a high prospective risk for a disease of interest more efficiently than traditional criteria. Perhaps most importantly, NHE serves as a validated blueprint for large-scale population health initiatives like the All of Us Research Program and the UK's Our Future Health, demonstrating a clear path to translate these massive data collection efforts into tangible predictive health tools.

Limitations and Future Directions

Despite its strong performance, the current NHE architecture has several limitations that represent clear avenues for future work. First, and most critically, our binary encoding of PRS as high/low-risk indicators, while effective, discards the majority of the continuous risk signal for 90% of individuals. Future work should prioritize the development of modality-specific

encoders to integrate the full, continuous PRS values, which could unlock a more nuanced understanding of genetic liability. Second, the 23andMe PRSs included were trained only on individuals of European genetic ancestry, reflecting the limited sample sizes available for running the cross-validation GWAS/PRS construction pipeline in other groups. To partially address this, we applied ancestry-matched thresholds (top and bottom 5%) to define high and low genetic risk within each group. Developing more robust PRSs that integrate information across ancestries will be essential to capture both shared and population-specific genetic architecture. Third, the model's current temporal encoding does not explicitly handle co-occurring diagnoses at the same age. Adopting a more flexible positional encoding strategy could better represent the complex structure of multimorbidity.

The model also faces limitations related to its data source. The reliance on self-reported data introduces the possibility of recall bias and healthy volunteer bias. While our prospective validation mitigates some of these concerns, a critical next step is to validate the model's predictions against structured clinical records (e.g., EHRs) to quantify any potential bias and its impact. Additionally, while the model incorporated PRS developed for multiple ancestries where available, many were derived from European-ancestry GWAS. Expanding the inclusion of well-powered, diverse-ancestry PRS is essential for improving predictive equity and generalizability across all populations. Finally, it is crucial to emphasize that NHE is a predictive, not a causal, model. While it can forecast what might happen next, it does not explain the underlying biological mechanisms.

In conclusion, the NHE model provides a powerful and validated blueprint for the next generation of predictive health tools. By successfully fusing deep genetic and longitudinal phenotypic data at population scale, our work highlights a clear path toward more accurate and comprehensive health forecasting. Continued refinement and validation of these methods will be essential for realizing a future where medicine can shift from a reactive to a proactive and deeply personalized paradigm.

Acknowledgements

We would like to thank the research participants and employees of 23andMe for making this work possible. We also thank David Hinds, Steve Pitts, Bertram Michael Holmes, Stella Aslibekyan, and Nick Eriksson for their comments on the manuscript. The authors gratefully acknowledge support from AWS for GPU computing and credits.

References

1. A. Belyaeva, A. Belyaeva, A. Shmatko, I. Ushinin, A. Avramov, N. Egorov, K. Sargsyan, M. Kartashev, R. Shalaby, T. Rocktäschel *et al.*, Multimodal llms for health grounded in individual-specific data, *arXiv preprint arXiv:2307.09018* (2023).
2. J. Cosentino, O. Shaik, A. Nag, T.-C. Liu, H. Adam, A. Adithyan, J. Agnes, B. Al-Banna, M. Al-Hajj, V. Anand *et al.*, Towards a personal health large language model, *arXiv preprint arXiv:2406.06474* (2024).
3. M. Kirchler, M. Ferro, V. Lorenzini, FinnGen, C. Lippert and A. Ganna, Large language models improve transferability of electronic health record-based predictions across countries and coding systems, *Health Informatics* (2025).

4. G. Savcisens, T. Eliassi-Rad, L. K. Hansen, L. H. Mortensen, L. Lilleholt, A. Rogers, I. Zettler and S. Lehmann, Using sequences of life-events to predict human lives, *Nat. Comput. Sci.* **4**, 43 (2023).
5. Z. Kraljevic, C. Tomlinson, R. Bendayan, T. Searle, L. Roguski, K. Noor, D. Bean, A. Mascio, J. T. Teo, R. J. Dobson *et al.*, Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study, *Lancet Digit. Health* **6**, e281 (2024).
6. A. Shmatko, A. Belyaeva, I. Ushinin, A. Avramov, N. Egorov, K. Sargsyan, M. Kartashev, R. Shalaby, T. Rocktäschel, S. Bach *et al.*, Learning the natural history of human disease with generative transformers, *Epidemiology* (2024).
7. A. King, L. Wu, H.-W. Deng, H. Shen and C. Wu, Polygenic risk score improves the accuracy of a clinical risk score for coronary artery disease, *BMC medicine* **20**, p. 385 (2022).
8. L. Li, S. Pang, F. Starnecker, B. Mueller-Myhsok and H. Schunkert, Integration of a polygenic score into guideline-recommended prediction of cardiovascular disease, *European heart journal* **45**, 1843 (2024).
9. G. R. Zirpoli, R. M. Pfeiffer, K. A. Bertrand, D. Huo, K. L. Lunetta and J. R. Palmer, Addition of polygenic risk score to a risk calculator for prediction of breast cancer in US black women, *Breast Cancer Research* **26**, p. 2 (2024).
10. J. Y. Tung, C. B. Do, D. A. Hinds, A. K. Kiefer and N. Eriksson, Efficient replication of over 180 genetic associations with self-reported medical data, *PLoS ONE* **6**, p. e23473 (2011).
11. F. Privé, J. Arbel and B. J. Vilhjálmsson, Ldpred2: better, faster, stronger, *Bioinformatics* **36**, 5424 (2021).
12. S. A. Lambert, L. Gil, S. Jupp, S. C. Ritchie, Y. Xu, A. Buniello, A. McMahon, G. Abraham, M. Akiyama, A. Al-Chalabi *et al.*, The polygenic score catalog as an open database for reproducibility and systematic evaluation, *Nat. Genet.* **53**, 420 (2021).
13. S. A. Lambert, L. Gil, S. C. Ritchie, M. Inouye, P. Würtz, S. Jupp, A. Buniello and J. A. MacArthur, Enhancing the polygenic score catalog with tools for score calculation and ancestry normalization, *Nat. Genet.* **56**, 1989 (2024).
14. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever, Language models are unsupervised multitask learners, *OpenAI Blog* **1**, p. 9 (2019).