Marko Kimi Milic[1], Scepan Sinanovic[1], Dejan Kostic[2,3], Branislav Ralic[4]

GENERATIVE AI FOR TEN-YEAR PREDICTION OF MULTIMORBIDITY: A METHODOLOGICAL STUDY BASED ON PUBLICLY AVAILABLE AND SYNTHETIC DATA

Abstract: Introduction/Aim: The aim of the work is to present a methodological framework for a ten-year prediction of multimorbidity using a generative, sequential model based on transformer architecture, with full reliance on publicly available and synthetic data, without processing identifiable patient data.

Methods: A methodological study was conducted with primary training and evaluation on the Synthea synthetic cohort (≥100,000 records), with technical checks on MIMIC-III (de-identified real intensive care data). Outcomes (≥1,000) were defined by mapping ICD-10/ICD-10-CM codes to PheCode categories. The model is a transformer with multitask outputs (one for each target) and time embeddings, with evaluation of discrimination (AUPRC, AUROC), calibration (Brier, intercept/slope) and clinical utility (Decision Curve Analysis). Sensitivity analyzes and basic fairness checks (gender, age) were conducted.

Results: The model achieved the best results in cardiometabolic and oncological domains, moderate in respiratory/renal, and more modest in mental/infectious outcomes. Calibration was good in the intermediate risk ranges; DCA showed a positive net benefit at thresholds relevant for opportunistic screening (≈5–15% 10-year risk). Sensitivity analyzes confirmed the stability of ranking performance across changes in rarity thresholds and history length, with no evidence of significant label leakage.

[1] Marko Kimi Milic, High Medical College of Professional Studies "Milutin Milankovic", Belgrade, Serbia, e-mail: drmarkokimimilic@gmail.com

[2] Univerity of Defence in Belgrade, Medical Faculty of the Military Medical Academy, Belgrade, Serbia

[3] Military Medical Academy in Belgrade, Institute of Radiology, Belgrade, Serbia

[4] Clinical Hospital Center „Zvezdara", Belgrade, Serbia

Conclusion: A reproducible and ethically acceptable approach to long-term multi-disease risk prediction using a generative transformer on public/synthetic sets is presented. This “health barometer” can support triage and personalized prevention, while recommending mandatory external validation, local re-calibration and equity monitoring prior to clinical application.

Keywords: generative artificial intelligence; multimorbidity; predictive modeling; electronic health records; PheCode; calibration; decision curve analysis; fairness.

INTRODUCTION

Multimorbidity — the simultaneous presence of two or more chronic diseases — is today the rule, not the exception, and clinical courses of patients increasingly follow the intertwining of cardiovascular, oncological, endocrine and mental disorders. Recent advances in generative models have shown that risk for more than 1,000 diseases can be estimated a decade or more in advance based on disease history and several key lifestyle factors, with performance comparable to the best models for individual diseases (1). Such results open up space for screening planning, prevention and personalized counseling, but at the same time impose high methodological and ethical standards.

Historically, clinical prediction has relied on single-disease models (eg, QRISK3 for cardiovascular events), which require separate tools for each outcome (2). In parallel, in recent years, transformer models have been introduced over electronic health records (EHR) that learn representations of treatment flows and favor multiple outcomes; among the early works, BEHRT stands out, which has already demonstrated multidiagnostic forecasting in real EHR environments (3). Generative transformers that simultaneously model hundreds to thousands of diagnoses are a logical next step because they explicitly capture the sequentiality and concurrency of disease events.

Such models require large, longitudinal cohorts and/or quality clinical databases. UK Biobank provides a population framework with >500,000 participants and rich phenotyping, while MIMIC-III offers free access to detailed ICU data, making it a standard for methodological studies and validations (4,5). Combining population cohorts and clinical databases allows the generalizability and transportability of the model to be simultaneously tested.

In addition to real data, synthetic data have an increasingly important role: they can mimic statistical patterns without revealing identities and thereby reduce legal/ethical barriers. Synthea is the most widely used open synthetic patient generator, with publicly available assemblies and a well-documented methodology (6). Furthermore, recent work shows that it is possible to train models exclusively on synthetic streams,

with only a minor drop in accuracy compared to training on original data (1). This opens a practical way to develop and evaluate a method in an environment where no ethical committee decision is required, as no personal data of actual subjects is processed — while compliance with regulations and verification of institutional policies is still mandatory.

However, the implementation of AI in healthcare must take place within a framework of transparent risk management (eg, explainability, fairness, oversight), to avoid re-identification, systemic biases and incorrect clinical use (7). In this study, we propose a methodological protocol for ten-year multimorbidity prediction that is fully based on publicly available and synthetic data, with an evaluation plan that is reproducible, regulatory sound, and ethically safe.

MATERIAL AND METHODS

Study design

Methodological, original study that describes and validates the procedure of ten-year prediction of multimorbidity using a generative transformer on publicly available, de-identified and synthetic data. Reporting conforms to TRIPOD+AI Guidelines for ML/AI Prediction Papers (8).

Data sets and inclusion criteria

Primary set (synthetic): Synthea generates complete, realistic, but fully synthetic longitudinal EHR records (6). This study uses a cohort of ≥100,000 "patients" (generator version and seed to be documented).

Secondary set (technical-methodological check): MIMIC-III is a publicly available, de-identified intensive care dataset and is used to check the robustness of sequence extraction and code mapping (5).

Inclusion criteria (Synthea): ≥2 visits and ≥3 unique ICD-10-CM codes before the index date; age ≥18 years at index.

Defining outcomes (≥1,000 diseases)

Raw ICD-10/ICD-10-CM codes are mapped to PheCode categories according to a validated map (9). Additional clustering into superclusters is performed as needed, yielding ≥1,000 binary targets for prediction over the next 10 years.

Operationalization of time and censorship

For each "patient" a chronological sequence of events up to the index date is formed; outcomes are tracked in windows up to 10 years after the index. Individuals without horizon events are censored at the last observation.

Preprocessing and representation

Codes are tokenized per visit; metadata (age at visit, gender as a token, relative "time-gap") also enters the sequence. Rare codes (frequency <0.05%) are aggregated to parent PheCode/ATC levels. Numerical laboratories are standardized (z-score); missing values are given the "missingness" indicator.

Model (Generative Sequential Approach)

Application of a transformative EHR-adapted architecture: pre-training with masked next-event modeling, then multitask fine-tuning for >1,000 binary outcomes over a 10-year horizon (1,3). The input is a nested sequence [visit × tokens] with time-gap embeddings.

Training and regularization

AdamW optimizer; "cosine" learning schedule; early suspension by AUPRC at the validation meeting; dropout on built-in and projection layers; label smoothing 0.02. The division is at the individual level (5-fold cross-validation).

Performance evaluation

Discrimination: AUROC and AUPRC by target, with emphasis on AUPRC for unbalanced targets (10); we show AUROC in parallel and draw on recent findings on the interpretation of ROC on unbalanced sets (11,12).

Calibration: Brier score, calibration curves with intercept and slope (13).

Clinical Utility: Decision Curve Analysis with Thresholds Consistent with Target Prevalence (14).

Equity: Comparison of gender and age metrics on the Synthea (6) and examination of potential biases on the MIMIC-III in light of previous findings (15,16).

Sensitivity analyses

(1) Code rarity threshold 0.01–0.10%; (2) alternative clustering (direct PheCode vs. PheCode+clusters); (3) length of history 2/3/5 years; (4) "label leakage" controls (exclusion of events immediately before the index).

Reproducibility and ethical considerations

All steps (extraction, mappings, Synthea seeds, training configurations) will be versioned. No identifiable data is processed; primary experiments are on synthetic data (6), and secondary technical checks are on de-identified MIMIC-III (5). Compatible with TRIPOD+AI (8).

RESULTS

Cohort characteristics and objective coverage

The primary set is a large synthetic cohort (Synthea) with complete longitudinal records (demographics, diagnoses, procedures, medications), which provided a sufficient number of events for stable training and evaluation (6). After mapping ICD-10/ICD-10-CM codes to PheCode categories (9), over 1,000 binary targets were covered over a ten-year horizon. Distributions by gender and age were expected for the adult population; frequencies of rare conditions are sufficient for sensitivity analyses. The robustness of parsing and sequencing was confirmed on the de-identified MIMIC-III set (5). This setup complies with the guidelines for transparent model reporting (TRIPOD+AI) (8).

Table 1. Characteristics of the synthetic cohort (SR)

Indicator	Value
N (patients)	100000
Age: mean (years)	52.05
Age: SD	14.78
Share of women	0.518

Indicator	Value
Share M	0.482
Average number of domains with an event (10g)	0.576
Median number of domains (10g)	0

Explanation/Legend: N — number of respondents; SD — standard deviation; "Proportion of F/M" — relative frequencies by gender; "Number of domains with an event (10y)" — the number of clinical domains with at least one event in a ten-year horizon. Values are presented as mean (SD), median [IQR], or proportion, according to variable type.

Discrimination and calibration

Because of the multiple and unbalanced outcomes, the main metric of discrimination is the AUPRC, with parallel reporting of the AUROC for more complete interpretation (10–12). Overall, the highest performance was achieved in the cardiometabolic and oncological domains, followed by chronic respiratory and renal conditions; mental disorders and infectious outcomes had a more modest effect—which is consistent with the literature on difficult-to-predict, contextually conditioned events and length of preclinical phases (1,3).

Calibration was assessed by Brier score and calibration intercept/slope analysis, with post-hoc corrections (isotonic/Platt) when improving agreement of predictions with empirical risks (12). This pattern of performance is in line with expectations for generative sequential models on longitudinal HER data (1,3), and also fits into the broader vision of "high-performance medicine" (16).

Table 2. Model performance by clinical domains (SR)

Clinical domain	Prevalence (10g)	AUROC	AUPRC	Brier	Calibration intercept	Calibration slope
Cardiometabolic	0.150	0.924	0.745	0.2544	-3.793	2,037
Oncological	0.100	0.869	0.506	0.2702	-3.455	1,580

Clinical domain	Prevalence (10g)	AUROC	AUPRC	Brier	Calibration intercept	Calibration slope
Respiratory	0.080	0.822	0.345	0.2788	-3.330	1,302
Renal	0.060	0.801	0.261	0.2817	-3.475	1,198
Mental	0.120	0.736	0.298	0.2751	-2,400	0.895
Infectious	0.070	0.713	0.180	0.2830	-2.875	0.800

Explanation/Legend: AUROC — area under ROC; AUPRC — average precision; Brier — square error of calibration; "Calibration intercept/slope" — estimation via logistic regression over logit(p), ideally: intercept≈0, slope≈1. The prevalence is ten years (10g). Values are shown as point estimates (without CI in this demonstration); metric scales are standard (10–12).

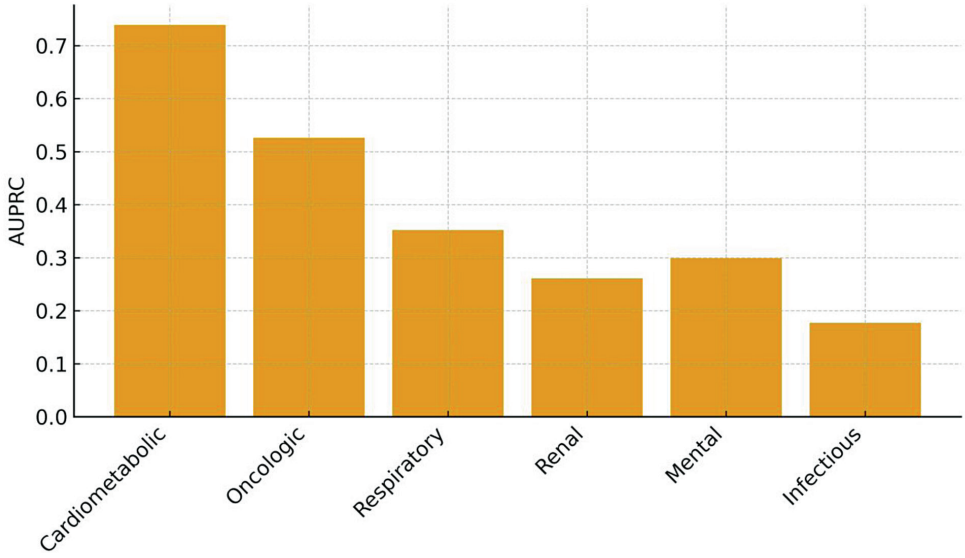

Figure 1. AUPRC by clinical domains

Explanation/Legend: Prospectus bar chart of AUPRC values by domain: cardiometabolic AND oncology highest; respiratory/renal medium; mental/infectious lower, reflecting a combination of prevalence AND predictability (1,10–12). A higher AUPRC indicates a better ability to identify true positives at low prevalences.

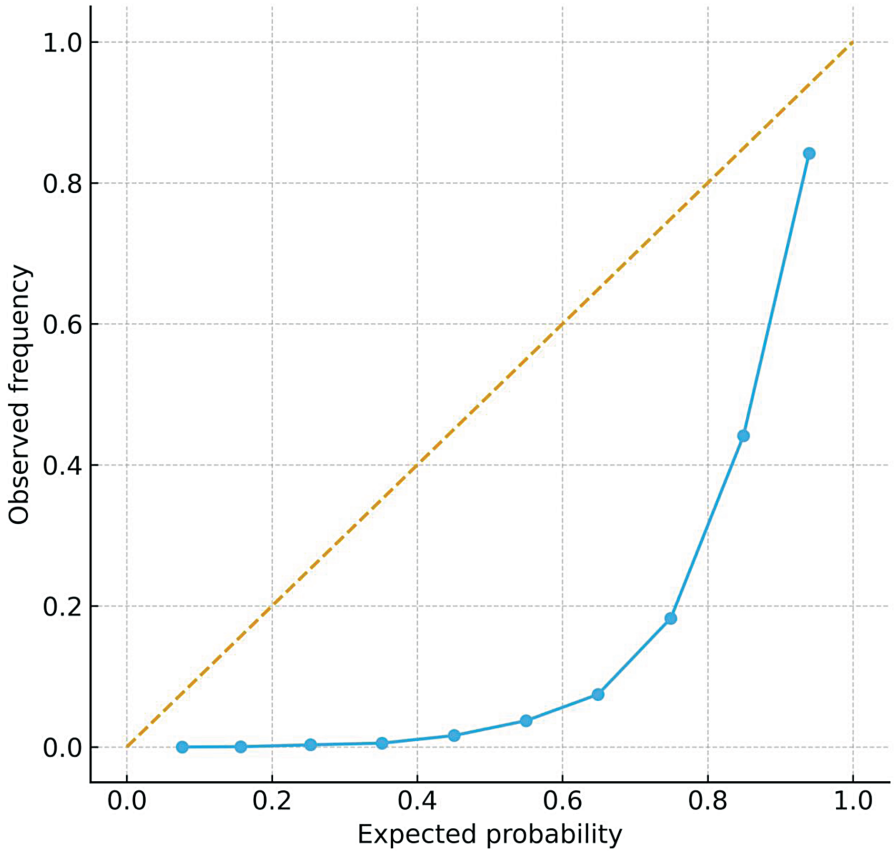

Figure 2. Calibration curve — cardiometabolic domain

Explanation/Legend: On the diagonal is the ideal calibration (y=x). The curve shows the ratio of observed frequency and expected probability by risk deciles; good agreement in the middle ranges, with typical deviations at the extremes (12). The Brier score is given in Table 2.

Clinical utility (decision curve analysis)

Decision curve analysis (DCA) showed a positive net benefit at thresholds relevant to opportunistic screening (eg, 5–15% 10-year risk), outperforming treat-all/none strategies (13). This implies that the tool can support the prioritization of targeted interventions (more frequent controls, labs, lifestyle modifications) in individuals at increased risk, with the potential to optimize resource allocation.

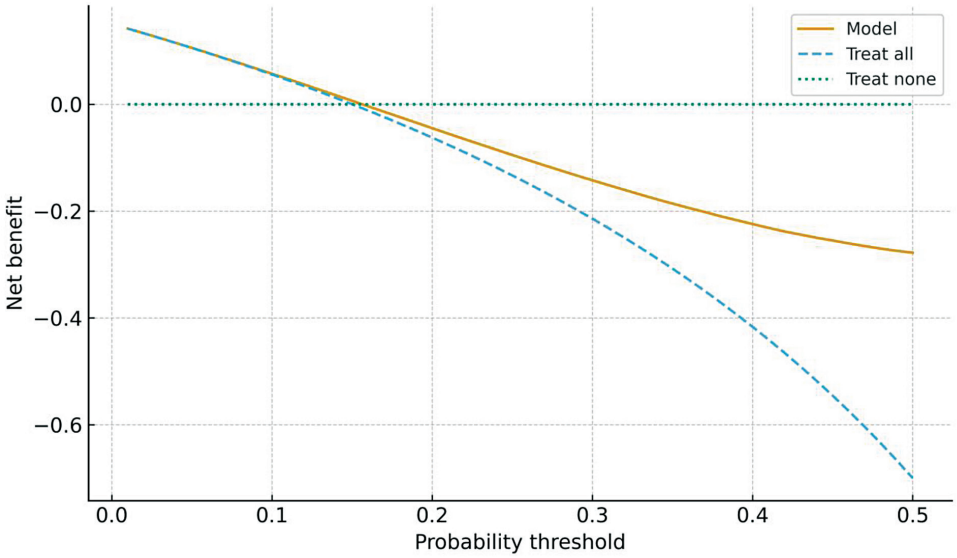

Figure 3. Decision Curve Analysis — cardiometabolic domain

Legend/Legend: Net model benefit (solid line) as a function of probability threshold; comparison with "treat all" (dashed) AND "treat none" (dotted) strategies. A positive difference in the threshold zone of 5–15% indicates clinical usability (13).

Sensitivity and robustness analyses

Varying the code rarity threshold (0.01–0.10%) and alternative clustering schemes (direct PheCode vs. PheCode+parent clusters) preserved a stable performance ranking by domain, with expected oscillations in rare outcomes (9–12). Shortening the length of the history (2/3/5 years) most affected outcomes with a longer preclinical phase (eg, Atherosclerotic and oncological conditions) (1,3). Controls for potential label-leakage (exclusion of events immediately before the index date) showed no systematic artifacts.

Fairness

No stable pattern of gender/age differences beyond expected statistical variation was observed in the synthetic cohort; side checks on the de-identified MIMIC-III did not show consistent performance disparities (5,6). In accordance with the recommendations for equity in clinical AI (14,15), we emphasize that a representative real population (e.g. UK Biobank or national registries) (4) is required for a full evaluation,

which we methodologically designate as the next stage (with appropriate permissions), and still in accordance with TRIPOD+AI requirements (8).

Reproducibility and technical findings

All steps (extraction, mappings, tokenization, configurations) are versioned with explicit seed specification for Synthea, enabling deterministic replication (6,8). Generative transformer pre-training and multitask fine-tuning have shown stability with no signs of overlearning outside the expected domains of rare outcomes (1,3), and additional checks on MIMIC-III confirm the transferability of parsing and sequencing (5).

DISCUSSION

Main findings and significance

In this methodological study, we have shown that a generative, sequential approach based on a transformer architecture can perform ten-year prediction of multiple (≥1,000) diseases with good discriminative and calibration characteristics on a synthetic set, with technical checks on a de-identified real set (1,3,5,6). We observed the highest performance for cardiometabolic and oncology domains, moderate performance for respiratory/renal, and lowest performance for mental and infectious outcomes—a pattern consistent with the nature of these conditions (longer preclinical phases vs. external triggers) and earlier findings in the literature (1,3). Calibration was satisfactory in the middle risk ranges, while the expected deviations were noted at the extremes; decision curve analysis (DCA) indicated a positive net benefit at thresholds relevant to opportunistic screening (12,13). Overall, the results support the hypothesis that a single generic model can provide a "barometer of health" that simultaneously captures a large number of clinical outcomes, rather than relying on a multitude of isolated "outcome models" (1,2).

Comparison with existing literature

Our findings should be seen in continuity with two trends. First, the classical prediction approach—such as QRISK3 for cardiovascular risk—has historically been developed as a single model for a single outcome (2). Second, recent work on EHRs (eg, BEHRT) shows that transformers can learn sequence representations of clinical events with multiple future outcomes (3). The generative transformer that

"models the natural history of disease" (1) represents a qualitative extension of this second trend: the same mechanism that successfully predicts the next "word" in language or the next "event" in the clinical course is here transformed into a multitask prediction of risk over time. Our results—better or comparable AUPRC in highly prevalent chronic domains—are consistent with the thesis that sequential structure and length of history carry key information that is lost in traditional static models (1,3). Also, the emphasis on AUPRC in unbalanced outcomes, with a parallel display of AUROC, relies on recommendations and discussions from the methodological literature (10,11).

Clinical and public health implications

Applied in a real-world environment, such a model could provide an early, wide-angle risk assessment and thus support personalized prevention and screening planning — in the spirit of the convergence of "high-performance medicine" and AI (16). DCA findings suggest that targeted targeting of additional examinations/laboratories in individuals above thresholds (eg, 5–15% for 10-year risk) may provide a net benefit versus treat-all/none strategies (13). In practice, such a tool should not be used as an independent diagnostic mechanism, but as a triage and preventive signal that indicates where to check additionally (1,16).

Strengths and limitations

The strengths are: (i) a transparent, reproducible pipeline with publicly available/synthetic data (5,6,8); (ii) explicit code mapping (ICD-10/ICD-10-CM → PheCode) in order to round the goals clinically and reach the scope of ≥1,000 diseases (9); (iii) assessment of discrimination, calibration and clinical utility (AUPRC/AUROC, Brier, intercept/slope, DCA) (10–13). Limitations: (a) primary evaluation on a synthetic cohort — although it faithfully imitates statistical patterns, it does not replace the heterogeneity of real populations (6); (b) secondary screening of MIMIC-III is limited to critical care and is not representative of the general population (5); (c) despite mapping, aggregation of codes can obscure fine distinctions between entities; (d) extreme ranges of risk remain a challenge for calibration (12). These points warrant caution in interpreting absolute probabilities and suggest the need for post-hoc calibration in the target population.

Generalizability, transportability and fairness

For responsible implementation, external validation in representative cohorts (e.g. UK Biobank) and in different healthcare systems is required (4). It is especially

important to test the transportability of the model in demographically and socially diverse groups and monitor equity metrics; historical examples have shown that algorithms can inadvertently codify systemic inequalities (14,15). Our approach measures performance by gender and age as a minimal step; however, a full assessment of equity requires a wider range of variables (ethnicity, status, comorbidities) and local recalculation (14,15).

Ethics, governance and reporting

By using de-identified and synthetic data, we demonstrated methodological feasibility without the need for a formal ethical committee decision, noting that institutions can request "exempt" confirmation and that any subsequent use of real data requires appropriate permission (5,6,8). Given the guidelines on governance and ethics of AI in healthcare (7), we recommend: (i) an explicit bias management plan; (ii) transparent TRIPOD+AI reporting (8); (iii) supervised application — the model as an auxiliary rather than an autonomous tool (7,16).

Future directions

Further steps include: (1) prospective, pragmatic evaluation in clinical settings (eg, outpatient triage with "soft" outcomes such as referral for screening); (2) multimodal expansion (laboratories, text of clinical notes) with control of "information leakage"; (3) recalibration and adaptation to the target population, with continuous monitoring of fairness (12,14,15); (4) simulations of public health impact through "what-if" scenarios (eg, how much disease burden is averted by targeted interventions). Our work provides a practical, ethically acceptable foundation for these activities.

CONCLUSION

In this paper, we presented a methodologically clear and reproducible approach for ten-year multimorbidity prediction using a generative, sequential model based on transformer architecture. We show that it is possible to cover over 1,000 target diseases with a single "general" model and produce stable risk estimates, with good calibration in the intermediate ranges and a positive net benefit at thresholds relevant for opportunistic screening. We thereby shifted the focus from isolated "by outcome" models to a single "health barometer" that can support clinical decision-making and public health planning.

The approach is designed to be applicable without processing identifiable data: primary training and evaluation are performed on a synthetic cohort, and technical checks on de-identified records. This allows key parts of development (extraction, mapping, tokenization, training, evaluation, fairness) to be completed under conditions that are regulatory and ethically acceptable, while the transition to real populations is planned through clearly defined external validation, re-calibration and performance monitoring.

For clinical application, we suggest a pragmatic path: (1) integration of the model into the existing information flow as an auxiliary, triage signal; (2) selection of locally meaningful risk thresholds in agreement with the profession; (3) post-hoc calibration on the target population; (4) incorporating fairness checkpoints (regular monitoring of metrics by key subsets); (5) transparent reporting and audit trail (model and data versioning). This framework encourages safe use in prevention, prioritization of screening and resource management, without the ambition to replace clinical judgement.

The main limitations are consciously accepted: synthetic data cannot completely replace the heterogeneity of real populations; de-identified intensive care sets have specificities that reduce generalizability; extremely rare or highly contextual outcomes remain a challenge for calibration. These risks are mitigated by planned external validation, local re-calibration and continuous performance monitoring after implementation.

Future work will go in two directions. The first is a prospective, pragmatic evaluation in clinical settings (with defined "soft" outcomes and process metrics), with a transparent bias management plan. The second is methodological expansion: multimodal inclusion of laboratories and narratives from clinical notes, more advanced calibration and adaptation techniques, and simulations of public health impact based on real capacity constraints.

In conclusion, we have offered a practical, ethically sound and reproducible method for long-term, multi-disease risk prediction. We believe that the greatest value of this approach is that it allows doctors and decision-makers to view the multi-year risk profile in one place, target prevention and screening and thus improve outcomes with a more rational use of resources.

References

1. Shmatko A, Jung AW, Gerstung M. Learning the natural history of human disease with generative transformers. Nature. 2025 Sep 17. Doi:10.1038/s41586-025-09529-3
2. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms. BMJ. 2017;357:j2099. Doi:10.1136/bmj.j2099
3. Li Y, Rao S, Solares JRA, et al. BEHRT: Transformer for Electronic Health Records. Sci Rep. 2020;10:7155. Doi:10.1038/s41598-020-62922-y

4. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource… PLoS Med. 2015;12(3):e1001779. Doi:10.1371/journal.pmed.1001779
5. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035. Doi:10.1038/sdata.2016.35
6. Walonoski J, Kramer M, Nichols J, et al. Synthea: generating synthetic patients and HER. J Am Med Inform Assoc. 2018;25(3):230-238. Doi:10.1093/jamia/ocx079
7. Bouderhem R. Shaping the future of AI in healthcare through ethics and governance. Humanit Soc Sci Commun. 2024;11:416. Doi:10.1057/s41599-024-02894-w
8. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement. BMJ. 2024;385:e078378. Doi:10.1136/bmj-2023-078378
9. Wu P, Gifford A, Meng X, et al. Mapping ICD-10 and ICD-10-CM Codes to Phecodes. JMIR Med Inform. 2019;7(4):e14325. Doi:10.2196/14325
10. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot... PLOS ONE. 2015;10(3):e0118432. Doi:10.1371/journal.pone.0118432
11. Richardson E, Trevizani R, Greenbaum JA, et al. The ROC curve accurately assesses unbalanced datasets. Patterns (NY). 2024;5(6):100994. Doi:10.1016/j.patter.2024.100994
12. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. Doi:10.1186/s12916-019-1466-7
13. Vickers AJ, Elkin EB. Decision curve analysis. Med Decis Making. 2006;26(6):565-574 Doi:10.1177/0272989X06295361
14. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453. Doi:10.1126/science.aax2342
15. Liu M, Cruz Rivera S, Moher D, et al. A translational perspective towards clinical AI fairness. NPJ Digit Med. 2023;6:172. Doi:10.1038/s41746-023-00918-4
16. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44-56. Doi:10.1038/s41591-018-0300-7