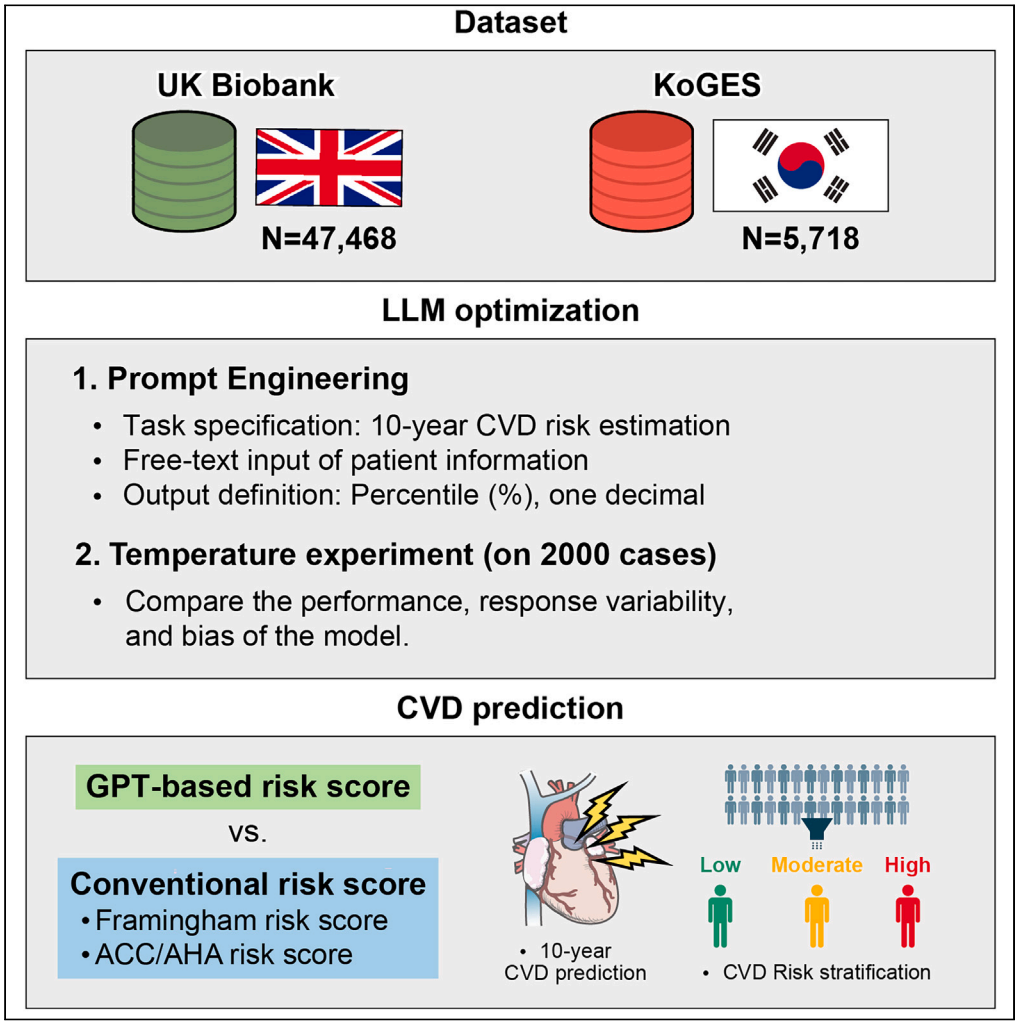
iScience

CellPress
OPEN ACCESS

Article

Evaluation of GPT-4 for 10-year cardiovascular risk prediction: Insights from the UK Biobank and KoGES data



Changho Han, Dong Won Kim, Songsoo Kim, Seng Chan You, Jin Young Park, SungA Bae, Dukyong Yoon

cardiobsa@yuhs.ac (S.B.)
dukyong.yoon@yonsei.ac.kr (D.Y.)

Highlights

Quantitative evaluation of GPT-4 in CVD risk scoring

GPT-4 shows robust performance regardless of data omission

GPT-4 consistent in multi-ethnic dataset

Study underscores GPT-4's potential in AI-driven healthcare

Han et al., iScience 27, 109022
February 16, 2024
https://doi.org/10.1016/j.isci.2024.109022

iScience

CellPress
OPEN ACCESS

Article

Evaluation of GPT-4 for 10-year cardiovascular risk prediction: Insights from the UK Biobank and KoGES data

Changho Han,[1,7] Dong Won Kim,[1,7] Songsoo Kim,[1,7] Seng Chan You,[1,2] Jin Young Park,[3,4,6] SungA Bae,[3,5,*] and Dukyong Yoon[1,2,3,8,*]

SUMMARY

Cardiovascular disease (CVD) remains a pressing global health concern. While traditional risk prediction methods such as the Framingham and American College of Cardiology/American Heart Association (ACC/AHA) risk scores have been widely used in the practice, artificial intelligence (AI), especially GPT-4, offers new opportunities. Utilizing large scale of multi-center data from 47,468 UK Biobank participants and 5,718 KoGES participants, this study quantitatively evaluated the predictive capabilities of GPT-4 in comparison with traditional models. Our results suggest that the GPT-based score showed commendably comparable performance in CVD prediction when compared to traditional models (AUROC on UKB: 0.725 for GPT-4, 0.733 for ACC/AHA, 0.728 for Framingham; KoGES: 0.664 for GPT-4, 0.674 for ACC/AHA, 0.675 for Framingham). Even with omission of certain variables, GPT-4's performance was robust, demonstrating its adaptability to data-scarce situations. In conclusion, this study emphasizes the promising role of GPT-4 in predicting CVD risks across varied ethnic datasets, pointing toward its expansive future applications in the medical practice.

INTRODUCTION

Cardiovascular disease (CVD) is the leading cause of morbidity and mortality worldwide, accounting for a considerable proportion of healthcare costs and posing a substantial public health risk.[1] The accurate and timely prediction of an individual's risk of developing CVD can facilitate early intervention and prevention strategies, which reduce the incidence and devastating impact of CVD.[2] Conventional CVD risk prediction models, such as the Framingham risk score[3] and the American College of Cardiology/American Heart Association (ACC/AHA) risk score,[4] are widely used in clinical settings. These models were derived from cohorts in the United Kingdom (UK) and the United States, respectively, and provide valuable insights into CVD risk prediction and aid in patient management.

Recently, artificial intelligence (AI) has been widely adopted across various fields of medicine.[5] Large language models (LLMs), particularly the generative pretrained transformer 4 (GPT-4) model developed by OpenAI, exhibit remarkable proficiency in producing human-like languages and have potential for application in various industries.[6,7] In the medical field, reports suggest that ChatGPT not only possesses knowledge sufficient to pass the United States Medical Licensing Examination (USMLE), but also has the potential to assist in various aspects of the medical workflow, such as making medical diagnosis and aiding in clinical decision making.[8,9] Particularly in the field of cardiology, ChatGPT has shown promise in verifying the appropriateness of recommendations for the prevention of cardiovascular diseases.[10]

Despite growing expectations for GPT's potential in medicine, many aspects of its practical applications remain unexplored. The efficacy of GPT in estimating the risk of CVD has not been studied. Additionally, due to its inherent language model nature, GPT may offer greater flexibility in terms of input compared to conventional CVD risk prediction models, and this flexibility warrants further evaluation. Furthermore, GPT is not without concerns. A significant current limitation of GPT impeding its application in the medical domain is the occurrence of inconsistent and potentially incorrect answers, because it is based on probabilistic algorithms.[11–13] When presented with identical prompts, the model's responses often vary, and while this variation can sometimes lead to more insightful responses, it can also result in less accurate or even incorrect answers.[12] There is a need to assess the variability of GPT's responses, quantitatively if possible, in order to further understand its reliability in clinical settings. Another concern with GPT is that the composition of its training corpus, as well as its training processes,

[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Yongin, Republic of Korea
[2]Institute for Innovation in Digital Healthcare, Severance Hospital, Seoul, Republic of Korea
[3]Center for Digital Health, Yongin Severance Hospital, Yonsei University Health System, Yongin, Republic of Korea
[4]Department of Psychiatry, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin, Republic of Korea
[5]Department of Cardiology, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin, Republic of Korea
[6]Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Yonsei University Health System, Seoul, Republic of Korea
[7]These authors contributed equally
[8]Lead contact
*Correspondence: cardiobsa@yuhs.ac (S.B.), dukyong.yoon@yonsei.ac.kr (D.Y.)
https://doi.org/10.1016/j.isci.2024.109022

are not fully transparent.[12] This lack of transparency raises questions about whether GPT will operate similarly across various groups, cohorts, or ethnicities.

Thus, in this study, we aimed to evaluate the efficacy and reliability of the GPT models in predicting 10-year CVD risk. Specifically, our objectives were to: (1) quantitatively evaluate the efficacy and reliability of the GPT models in evaluating 10-year CVD risk through comparative analysis with established benchmarks such as the Framingham risk score using real-world, longitudinal data from different ethnic groups including the UK Biobank and the Korean Genome and Epidemiology Study (KoGES) data[14,15]; (2) quantitatively evaluate the variability of the GPT-4 risk score by conducting multiple iterations of experiments for each subject at different GPT-4 temperature settings; and (3) investigate the adaptability and flexibility of GPT-4 in scenarios of incomplete data, a common challenge in clinical settings. This way, we aimed to gain insights into the capabilities of GPT in estimating CVD risk within the complex landscape that entails both promise and pitfalls in applying LLMs to the medical domain.

RESULTS

Cohort selection and baseline characteristics

The UK Biobank study included 502,396 participants aged 40–69 years at the time of assessment, recruited between 2006 and 2010 (Figure S1). Data pertaining to age, sex, diabetes diagnosed by a doctor, blood pressure medication, smoking status, total cholesterol, high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, triglycerides, systolic blood pressure (SBP), diastolic blood pressure (DBP), standing height, weight, date of attending the assessment center, and date of death were extracted, and the outcome was 10-year CVD risk defined by major adverse cardiovascular events (MACEs). A total of 103,817 participants with missing data were excluded. Of the remaining participants, after randomly selecting 50,000 participants, 2,532 patients who had previously experienced MACEs were further excluded, leaving 47,468 subjects for the analysis.

For the KoGES study population, variables analogous to those used in the UK Biobank were extracted. Out of an initial 10,030 participants, 875 with missing data were excluded (Figure S2). An additional 3,437 participants, who were not followed up at the 10-year mark, were also omitted, resulting in a final cohort of 5,718 subjects for subsequent analysis.

Baseline characteristics of the UK Biobank and KoGES participants and the CVD risk derived from GPT-4 are shown in Table 1. Among a total of 47,468 individuals from the UK Biobank for analysis, the participants had an overall median age of 57 years (IQR 50–63), with 21,224 (44.7%) men and 3,136 (6.6%) experiencing MACE within 10 years. When grouped by the GPT-4 risk group category (details on deriving and categorizing 10-year CVD risk from GPT-4 will be described in the next section), 15,190 individuals were classified as low-risk, 10,290 as moderate-risk, and 21,268 as high-risk. Among the 5,718 patients from KoGES included in the study, the participants had an overall median age of 49 years (IQR 44–59), with 2,663 (46.6%) men and 176 (3.1%) experiencing MACE within 10 years. In both the UK Biobank and KoGES cohorts, the higher-risk groups had older individuals, a higher proportion of males, a higher prevalence of diabetes mellitus, more antihypertensive treatment, a higher proportion of smokers, more unfavorable lipid profiles, higher blood pressure, a higher body mass index (BMI), and a higher incidence of 10-year MACE ($p < 0.001$).

Performances of the GPT models in 10-year CVD risk prediction and comparison with traditional models

To predict the incidence of CVDs using GPT, we transformed the variables into a sentence structure, as exemplified in Figure 1. Predefined information on each participant was provided to the GPT and we prompted the GPT to answer only the risk percentage rather than extensive text narratives. Based on the 10-year CVD risk percentage, <10% was classified as low-risk, 10% and <20% as moderate-risk, and >20% as high-risk.

Performances of each risk scoring method (GPT-4, GPT-3.5-turbo, Framingham risk score and ACC/AHA risk score) on 10-year CVD risk prediction were evaluated and compared in both the UK Biobank and KoGES cohorts. In cases of GPT-4 and GPT-3.5-turbo, we utilized the temperature of 0.4 as the optimal setting (the details of the process by which the optimal temperature was predetermined to be 0.4 are described in the method details section). In the UK Biobank cohort, the highest area under the receiver operating characteristics curve (AUROC) was found for the ACC/AHA risk score with 0.733, followed by the Framingham risk score at 0.728, GPT-4 at 0.725, and GPT-3.5-turbo at 0.706 (Figure 2A). The DeLong test revealed statistically significant differences between GPT-3.5-turbo and both the Framingham risk score and the ACC/AHA risk score ($p < 0.001$). While GPT-4 and the ACC/AHA risk score also showed statistically significant differences in the DeLong test ($p < 0.001$), the difference between GPT-4 and the Framingham risk score was not statistically significant ($p = 0.120$). In the KoGES dataset, the highest AUROC scores were observed for the ACC/AHA risk score at 0.674, the Framingham risk score at 0.675, GPT-4 at 0.671, and GPT-3.5-turbo at 0.664 (Figure 2B). The DeLong test indicated no statistically significant differences between GPT-3.5-turbo and both the Framingham risk score and the ACC/AHA risk score ($p = 0.715$, $p = 0.805$). Also, no statistically significant differences were found between GPT-4 and both the Framingham risk score and the ACC/AHA risk score ($p = 0.145$, $p = 0.166$). Detailed metrics are provided in Table 2. In Table 2, the risk threshold for calculating accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) was set at 20%, which was used to differentiate the high-risk group from the rest in all risk scoring methods.

In the UK Biobank cohort, the Pearson correlation coefficient (Pearson's r) revealed a substantial correlation between the GPT-based score and the ACC/AHA risk score, with a Pearson's r value of 0.882 (Figure 2C). Similarly, in the KoGES cohort, there was a notable correlation between the GPT-based score and the ACC/AHA risk score, evidenced by a Pearson's r value of 0.867 (Figure 2D). As delineated in Figure S3, correlations between the GPT-based score and the Framingham risk score were also discerned in both the UK Biobank (Pearson's r = 0.890)

Table 1. Baseline characteristics (grouped by GPT-4 based risk category) of UK Biobank and KoGES patients

	Low risk	Moderate risk	High risk	Overall	p-value
UK Biobank					
N	15910	10290	21268	47468	
10-year CVD incidence, n	255 (1.6)	508 (4.9)	2373 (11.2)	3136 (6.6)	<0.001
Age	49 [44,54]	57 [51,61]	63 [59,66]	57 [50,63]	<0.001
Sex (male)	2932 (18.4)	4536 (44.1)	13756 (64.7)	21224 (44.7)	<0.001
Total cholesterol, mmol/L	5.4 [4.8,6.1]	5.9 [5.3,6.6]	5.9 [5.0,6.7]	5.7 [5.0,6.5]	<0.001
HDL, mmol/L	1.5 [1.3,1.8]	1.4 [1.2,1.7]	1.3 [1.1,1.6]	1.4 [1.2,1.7]	<0.001
LDL, mmol/L	3.3 [2.9,3.8]	3.7 [3.2,4.3]	3.7 [3.1,4.3]	3.6 [3.0,4.1]	<0.001
Triglyceride, mmol/L	1.1 [0.8,1.6]	1.5 [1.1,2.2]	1.8 [1.3,2.5]	1.5 [1.0,2.1]	<0.001
SBP, mm/hg	124 [116,134]	135 [126,143]	147 [135,158]	136 [124,149]	<0.001
DBP, mm/hg	78 [72,84]	82 [76,88]	85 [79,92]	82 [75,89]	<0.001
BMI, mg/kg^2	25.1 [22.8,28.0]	26.8 [24.4,29.7]	27.7 [25.2,30.8]	26.6 [24.1,29.8]	<0.001
Smoking (current)	904 (5.7)	900 (8.7)	3070 (14.4)	4874 (10.3)	<0.001
Blood pressure medication	645 (4.1)	1333 (13.0)	6715 (31.6)	8693 (18.3)	<0.001
Diabetes	15 (0.1)	119 (1.2)	2044 (9.6)	2178 (4.6)	<0.001
Framingham risk score	4.9 [3.4,6.6]	10.3 [8.5,12.7]	20.1 [15.2,27.3]	11.4 [6.3,19.1]	<0.001
ACC/AHA risk score	1.6 [0.9,2.5]	5.0 [3.8,6.5]	12.1 [8.5,17.2]	5.6 [2.4,11.3]	<0.001
GPT-based score	2.9 [2.0,5.3]	15.6 [13.2,17.9]	27.3 [22.2,34.7]	18.0 [5.2,26.0]	<0.001
KoGES					
N	3070	1190	1458	5718	
10-year CVD incidence, n	55 (1.8)	37 (3.1)	84 (5.8)	176 (3.1)	<0.001
Age	46 [43,52]	53 [45,61]	60 [49,65]	49 [44,59]	<0.001
Sex (male)	796 (25.9)	762 (64.0)	1105 (75.8)	2663 (46.6)	<0.001
Total cholesterol, mmol/L	192.0 [170.0.214.0]	198.0 [176.0.221.8]	209.0 [183.0.235.0]	196.0 [174.0.221.0]	<0.001
HDL, mmol/L	49.0 [43.0,57.0]	47.0 [40.0,55.0]	45.0 [39.0,53.0]	48.0 [41.0,56.0]	<0.001
LDL, mmol/L	115.0 [96.2,135.8]	119.5 [98.6,142.3]	124.2 [99.5,148.8]	118.0 [97.2,140.6]	<0.001
Triglyceride, mmol/L	108.0 [77.0.153.0]	132.0 [94.0.181.8]	164.0 [113.0.236.8]	124.0 [87.0.181.0]	<0.001
SBP, mm/hg	114 [105,124]	122 [110,133]	129 [117,144]	118 [108,131]	<0.001
DBP, mm/hg	77 [70,84]	81 [74,89]	85 [77,92]	80 [72,88]	<0.001
BMI, mg/kg^2	24.3 [22.5,26.3]	24.5 [22.6,26.5]	24.9 [22.9,26.9]	24.5 [22.6,26.5]	<0.001
Smoking (current)	298 (9.7)	120 (10.1)	165 (11.3)	583 (10.2)	0.244
Blood pressure medication	140 (4.6)	128 (10.8)	369 (25.3)	637 (11.1)	<0.001
Diabetes	9 (0.3)	44 (3.7)	273 (18.7)	326 (5.7)	<0.001
Framingham risk score	3.7 [2.2,5.8]	9.6 [7.7,12.0]	18.5 [13.6,25.7]	7.0 [3.4,12.8]	<0.001
ACC/AHA risk score	1.2 [0.6,2.3]	5.0 [3.7,6.6]	10.7 [7.8,15.2]	3.1 [1.1,7.0]	<0.001
GPT-based score	2.2 [1.2,3.6]	15.8 [13.6,18.1]	25.1 [21.6,33.6]	7.7 [2.1,20.1]	<0.001

GPT: generative pretrained transformer, UK: United Kingdom, KoGES: Korean Genome and Epidemiology Study, BMI: body mass index, SBP: systolic blood pressure, DBP: diastolic blood pressure, HDL: high-density lipoprotein, LDL: low-density lipoprotein, MACE: major adverse cardiovascular event. Data are median (IQR) or n (%), ACC/AHA: American College of Cardiology/American Heart Association.

and the KoGES (Pearson's r = 0.896) cohorts. Furthermore, the correlation between the Framingham risk score and the ACC/AHA risk score was 0.956 in the UK Biobank and 0.954 in the KoGES.

The Kaplan-Meier method was applied to plot survival curves for the low-, moderate-, and high-risk groups based on the risk scoring methods (Figure 3). Distinct segregation was observed between the survival curves of the three risk categories across all risk scoring methods, with all pairwise comparisons yielding statistically significant differences according to the log rank test with a post-hoc Bonferroni correction. This confirmed the risk stratification capability of the GPT-based score. However, the degree of separation was not more pronounced than what was achieved using the Framingham risk score or the ACC/AHA risk score.

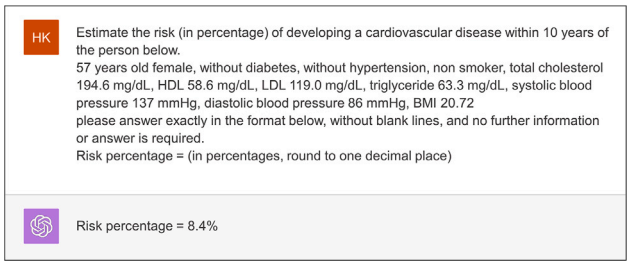


Figure 1. Example of a GPT prompt and response

Tabular data extracted from UK Biobank and KoGES were organized and queried into a sentence format. The 10-year cardiovascular disease risk percentage was extracted using regular expressions from the corresponding answers. GPT: generative pretrained transformer, UK: United Kingdom, KoGES: Korean Genome and Epidemiology Study, HDL: high-density lipoprotein, LDL: low-density lipoprotein, BMI: body mass index.

Performance of GPT-4 with omission of variables in the UK Biobank cohort

To evaluate GPT-4's adaptability when certain information cannot be obtained, we conducted further experiments deliberately omitting specific variables from the input prompt within the UK Biobank cohort. Initially, we excluded laboratory data—total cholesterol, HDL cholesterol, LDL cholesterol, and triglycerides—resulting in an AUROC of 0.722 and an area under the precision recall curve (AUPRC) of 0.141 (Table S1). These results are nearly consistent with GPT-4's original performance metrics (AUROC 0.725, AUPRC 0.145). Subsequently, we omitted physical examination data—SBP, DBP, and BMI—and observed an AUROC of 0.715 and an AUPRC of 0.134, which are also comparable to the baseline performances (Table S1). Detailed metrics can be found in Table S1.

DISCUSSION

In this study, we quantitatively evaluated the efficacy and reliability of LLMs in predicting 10-year CVD risk using real world longitudinal data, and compared their performances with that of conventional risk prediction models. Our findings indicate that GPT-4's performance is comparable to conventional risk prediction models such as the Framingham risk score (GPT-4 AUROC 0.725 vs. Framingham risk score AUROC 0.728). The Kaplan-Meier analyses confirmed the risk stratification capability of the GPT-based CVD risk score when stratified into low, moderate, and high risk groups. GPT-4's performance remained robust even with the omission of certain variables, highlighting its adaptability in circumstances where certain information might not be acquirable.

The recent integration of AI into the medical sphere signifies a pivotal evolution in healthcare practices, particularly with recent evidence demonstrating its proficiency in numerous prediction tasks, including the risk assessment of diseases.[16,17] Of late, the rapid advancements in LLMs, particularly the GPT series, have significantly heightened expectations regarding the potential application of LLMs in the field of medicine.[18–21] GPT is a state-of-the-art language model employing deep learning to generate responses that closely mimic human conversation in reaction to natural language prompts.[22,23] As one of the largest language models available to the public, GPT was trained utilizing a vast corpus of text data to understand and replicate the subtleties of human language, thus producing relevant and contextually aware responses to a diverse range of prompts.[22,23]

Table 2. 10-year cardiovascular disease risk prediction performances of the risk scoring methods

	AUROC	AUPRC	Sensitivity	Specificity	PPV	NPV	F1 score
UK Biobank							
GPT-4	0.725	0.145	0.757	0.574	0.112	0.971	0.194
GPT-3.5-turbo	0.706	0.135	0.308	0.879	0.152	0.947	0.204
ACC/AHA risk score	0.733	0.151	0.211	0.935	0.188	0.944	0.199
Framingham risk score	0.728	0.149	0.496	0.790	0.143	0.957	0.222
KoGES							
GPT-4	0.664	0.054	0.477	0.752	0.058	0.978	0.103
GPT-3.5-turbo	0.671	0.059	0.108	0.957	0.073	0.971	0.087
ACC/AHA risk score	0.674	0.059	0.662	0.974	0.071	0.970	0.066
Framingham risk score	0.675	0.061	0.278	0.893	0.077	0.975	0.120

AUROC: area under the receiver operating characteristics curve, AUPRC: area under the precision recall curve, PPV: positive predictive value, NPV: negative predictive value, UK: United Kingdom, GPT: generative pretrained transformer, ACC/AHA: American College of Cardiology/American Heart Association, KoGES: Korean Genome and Epidemiology Study.

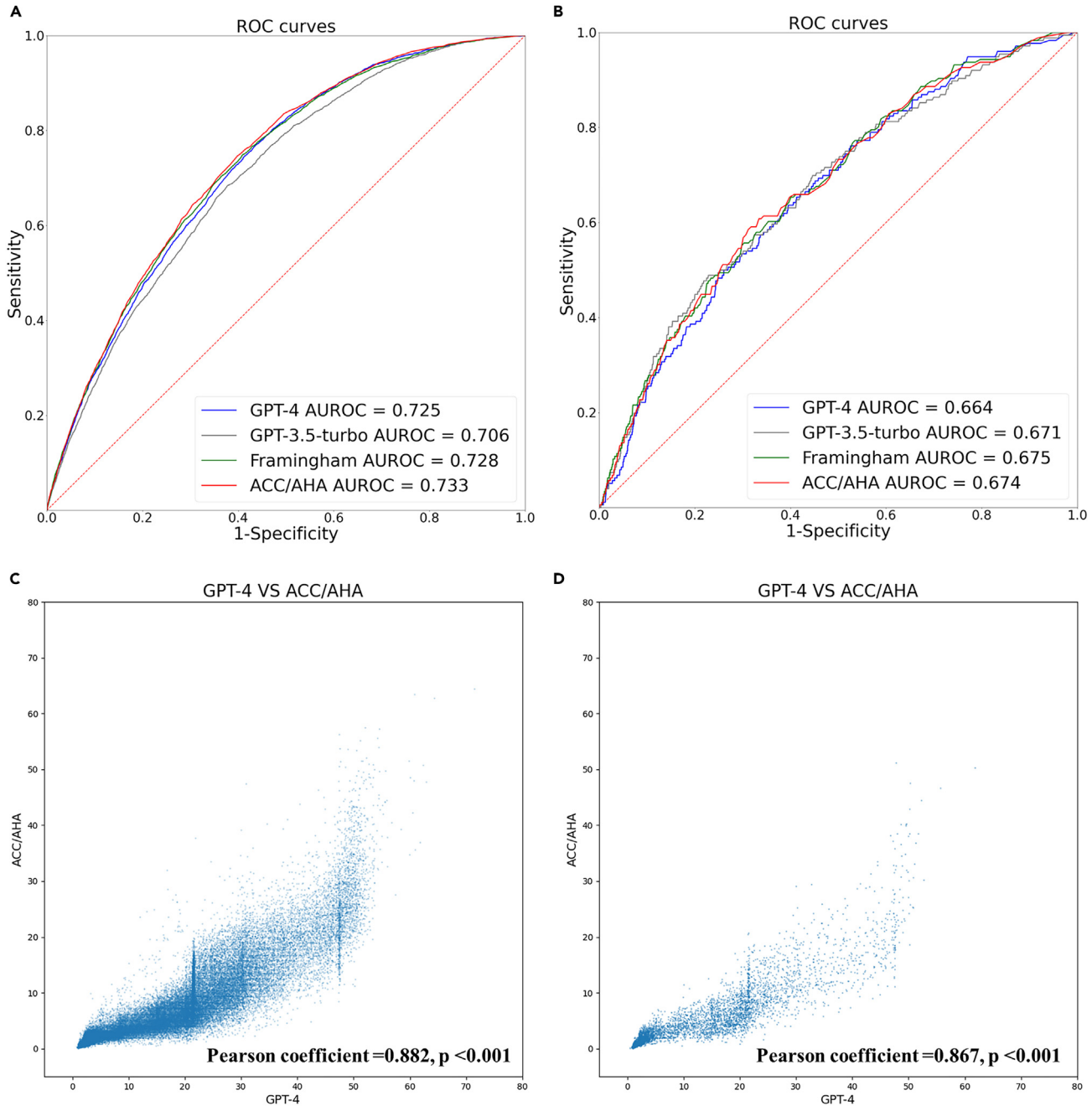


Figure 2. Performance evaluation and comparison of the risk scoring methods in the UK Biobank and KoGES cohorts

GPT-4's performance was comparable to conventional risk prediction models. Substantial correlation was found between the GPT-based risk score, ACC/AHA risk score and Framingham risk score.

(A) AUROC curves (UK Biobank).

(B) AUROC curves (KoGES).

(C) Scatterplot (UK Biobank, GPT-4 vs. ACC/AHA risk score).

(D) Scatterplot (KoGES, GPT-4 vs. ACC/AHA risk score). UK: United Kingdom, KoGES: Korean Genome and Epidemiology Study, GPT: generative pretrained transformer, ACC/AHA: American College of Cardiology/American Heart Association, AUROC: area under the receiver operating characteristics curve.

GPT's proficiency in cognitive tasks and interactive communication, nearly paralleling human capability, is increasingly recognized for its prospective transformative impact on medical practices, prompting numerous studies and investigations in recent months to corroborate its utility[18–21]: GPT has demonstrated the ability to achieve passing scores on the United States Medical Licensing Examinations, with GPT-4

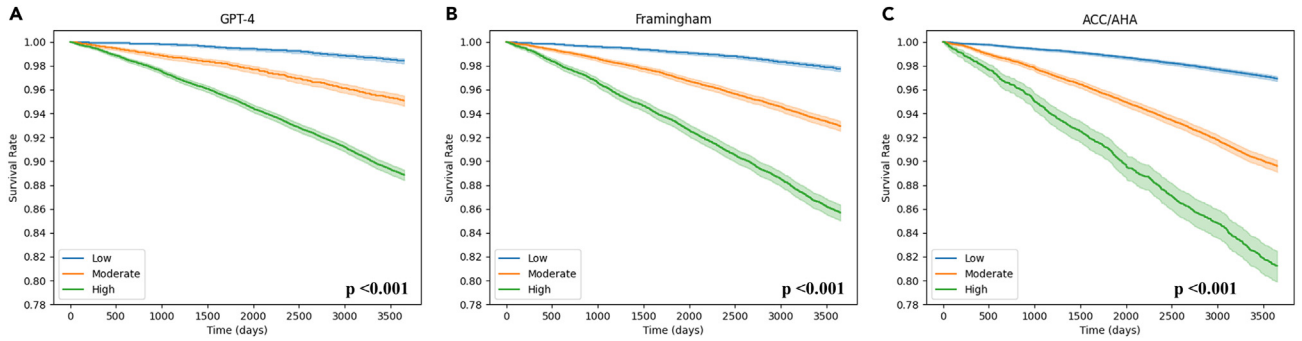


Figure 3. Kaplan-Meier curves stratified by risk categories in the UK Biobank cohort

All pairwise comparisons between curves with the log rank test with post-hoc Bonferroni correction were statistically significant.

(A) Kaplan-Meier curve stratified by GPT-4 based risk category.

(B) Kaplan-Meier curve stratified by Framingham risk score category.

(C) Kaplan-Meier curve stratified by ACC/AHA risk score category. GPT: generative pretrained transformer, ACC/AHA: American College of Cardiology/American Heart Association, UK: United Kingdom.

exhibiting a significant improvement over its predecessor, GPT-3.5-turbo[8,24]; GPT provided largely accurate responses to 284 medical questions posed by physicians across 17 specialties, as evaluated by academic physician specialists, and also provided largely appropriate responses to CVD prevention questions as evaluated by cardiology clinicians[10,25]; when comparing GPT's responses to patient questions with those given by doctors on a social network, the language model's outputs were favored for their quality and empathetic tone, as assessed qualitatively by medical professionals[26]; GPT has proven effective at converting various free-text radiology reports into structured formats with minimal effort, suggesting potential applications in standardization and data mining for research purposes.[27]

Despite the recent upsurge in interest regarding GPT, numerous aspects of its utility in medical domains are yet to be explored. Literature review revealed that the application of GPT in medical predictive tasks, including its ability to predict future cardiovascular events, was largely unexplored. Moreover, it is also important to acknowledge and address the concerns and limitations currently associated with the use of GPT in medicine. Most importantly, due to its reliance on probabilistic algorithms, GPT models can yield variable results to the same prompts, and their lack of transparency in training data and processes leads to uncertainties about their accuracy, calling for persistent human monitoring to ensure reliability.[11–13]

In this context, our study has numerous implications. To the best of our knowledge, this study is the first to quantitatively evaluate GPT's performance in predicting 10-year CVD risk. The similar performance of GPT-4 in generating CVD risk scores with that of the traditional model is notable (GPT-4 AUROC 0.725 vs. Framingham risk score AUROC 0.728 vs. ACC/AHA risk score AUROC 0.733). Traditional models such as the Framingham or ACC/AHA risk scores rely on mathematical calculations based on various variables to derive their results.[3,4,28] On the other hand, GPT-4 employs a fundamentally different mechanism by vectorizing words to learn patterns and predicting the most probabilistically appropriate next word or sentence.[29] Despite these divergent methodologies, both GPT-4 and traditional regression-based models have shown comparable effectiveness in actual CVD prediction and patient risk stratification. The capability of GPT-4 for risk stratification was further confirmed through Kaplan-Meier analysis. To assess the model's efficacy in reflecting actual patient outcomes, we conducted this analysis using the same cut-off points commonly applied to Framingham risk scores, namely 10 and 20.[30,31] The resulting Kaplan-Meier curves for each risk group, as categorized by GPT-based scores, were statistically distinct, underlining the model's usefulness in this regard. However, before utilizing GPT-4 for actual risk stratification, it is imperative to establish carefully considered cut-off points and interpretative guidelines to ensure the model's effective application.

This study is also the first to quantitatively assess the variability in GPT's predictions of 10-year CVD risk. As aforementioned, a notable concern when comparing traditional prediction models to LLMs like GPT-4 is the variability or inconsistency in outputs from GPT-4. To accurately understand this variability, thorough testing is essential. The variability in 10-year CVD risk predictions of GPT-4 is quantifiable because the risk is expressed in numerical terms, which is what our study has rigorously analyzed. Our experiments focused on how different temperature settings affected the consistency and reliability of GPT-4's outputs. We found that lower temperature settings led to reduced variability, as indicated by lower standard deviations and coefficients of variation. The AUROC was maximized at a temperature setting of 0.4, although the difference was not substantial across various settings. However, lower temperature settings resulted in an undesirable clustering of GPT-based scores around specific values, a phenomenon we refer to as "streaking," which limited the model's capacity for fine-grained risk stratification. Thus, a temperature setting of 0.4 was deemed optimal, as it mitigated the "streaking" issue while maintaining strong overall performance.

This study highlights GPT-4's adaptability in managing incomplete clinical data, a common challenge in healthcare settings. Unlike traditional models that require discarding or imputing missing data, potentially leading to biases or inaccuracies, our study showed that GPT-4 maintains robust predictive performance even with missing information. This is demonstrated by the consistent performance in predicting CVD risk even with the omission of certain key variables as shown in Table S1. GPT-4's sophisticated algorithmic design enables flexible administration of input prompt, which is invaluable in clinical decision-making where full datasets may not be readily available. Although

traditional models like the PCE remain relevant where computational resources are sparse, GPT-4 might offer a novel solution to data irregularities, marking an advancement in AI-driven healthcare applications.

We further validated the adaptability of GPT-4 by confirming consistent tendencies in two datasets from two different ethnic groups, namely the UK Biobank and KoGES datasets. Traditional models are typically derived from specific cohorts, necessitating verification for generalizability across diverse populations with varying demographic, clinical, and genetic characteristics before application.[30] Similar challenges exist for GPT-4 and other LLMs, primarily because the composition of their training corpus is not fully transparent.[12] Consequently, there is a risk that an LLM model might not yield consistent results across different cohorts. Despite these potential limitations, our study demonstrated that GPT-4 yielded consistent tendencies across datasets with varying demographic and clinical characteristics. This performance lends support to the model's adaptability and potential for broader applications.

The rate of advancement in LLMs has accelerated significantly in a brief period. As recently as December 2020, the GPT-Neo model set a new precedent in the MedQA dataset, which is a medical benchmark dataset comprising questions in the style of the USMLE, with an accuracy of 33.3%.[32,33] This milestone was rapidly surpassed many times, and by December 2022, the Flan-PaLM model achieved an accuracy of 67.6%.[33] By May 2023, GPT-4 marked a significant leap forward, attaining an accuracy of 86.1%.[34] This study also corroborates this trajectory of progress, revealing that GPT-4 outperforms GPT-3.5-turbo in 10-year CVD risk prediction within the UK Biobank cohort (GPT-3.5-turbo AUROC 0.73 vs. GPT-4 AUROC 0.75). Moreover, as various studies unfold, there is accumulating evidence that LLMs can integrate extensive medical data and patient information, potentially contributing to clinical decision-making and education in healthcare.[35–37] These developments, together with the capability of GPT in medical predictive tasks that we have shown in our study, underscore the critical need for additional research to assess the unexplored potentials and broad applicability of LLMs in diverse medical contexts.

Conclusions

Our study was the first to quantitatively evaluate the efficacy and reliability of GPT-4 in predicting 10-year CVD risk, and our findings indicate that GPT-4's performance is comparable to conventional risk prediction models. The Kaplan-Meier analyses verified the risk stratification capability of the GPT-based CVD risk score. The robustness of GPT-4 was maintained even after some key variables were excluded, emphasizing its flexibility in situations where certain information may be unavailable. Furthermore, we validated the adaptability of GPT-4 by confirming consistent tendencies in two datasets from two different ethnic groups. Considering the rapid pace of development in LLMs, the future holds even greater promise, and there is a need for additional research to assess the untapped possibilities and wide usability of LLMs in various medical fields.

Limitations of the study

This study has a few limitations. First, our experiments on GPT-4 were not conducted on the entire cohort population but rather on selected subsets of 50,000 and 2,000 instances, introducing the possibility of selection bias. The cost of using the GPT-4 model for this research is based on a per-token pricing model for both input and output. Running multiple iterations on tens of thousands of cases would incur a substantial expense. Given these constraints, we aimed to select an optimal sample size that would both be cost-effective and yield the strongest statistical power for our analyses. Second, we only tested a single prompt for the task. We utilized a zero-shot prompt to predict CVD risk. However, other prompt engineering techniques, such as few-shot learning and Chain of Thought, have been reported to potentially enhance the model's performance.[6,38] This underscores the growing expectations surrounding the capabilities of LLMs, and future studies should focus on this prompt engineering technique. Additionally, an unavoidable limitation arises from the non-disclosure of specific details regarding GPT-4's training data. Given that the inner workings and training corpus of GPT-4 are not publicly available, it is unclear whether the model's predictions are influenced by direct references to ACC/AHA guidelines or other specific medical literature. This lack of detailed transparency inherently limits our ability to fully understand the underpinnings of the model's risk predictions, while not diminishing the overall value and insights provided by our research. Finally, we focused solely on well-known variables when interacting with the LLMs. However, one of the key advantages of LLMs lies in their ability to handle diverse formats of input. This aspect was not experimentally explored in our study, limiting the scope of our findings. Unlike traditional models that depend exclusively on structured data, LLMs are capable of assimilating various types of data, including unstructured medical records.[39,40] Exploring this capacity could lead to the discovery of new risk factors and advancements in risk prediction models. Future research may delve deeper into these facets, thereby unveiling new possibilities for LLM-based risk prediction models.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Participant data information

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109022.

ACKNOWLEDGMENTS

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI22C0452). This study was conducted using data from UK Biobank (application number: 85037). Data in this study were obtained from the Korean Genome and Epidemiology Study (KOGES; 6635-302), National Research Institute of Health, Korea Disease Control and Prevention Agency, Republic of Korea. We appreciate the Medical Illustration & Design (MID) team, a member of Medical Research Support Services of Yonsei University College of Medicine, for their excellent support with medical illustration.

AUTHOR CONTRIBUTIONS

Conceptualization: S.A.B.; methodology: D.Y., C.H., D.W.K., S.K., and JY.P.; validation: S.C.Y.; investigation: C.H., D.W.K., and S.K.; writing – original draft preparation: C.H., D.W.K., and S.K.; writing – review and editing: D.Y., S.A.B., and S.C.Y.; supervision: D.Y. and S.A.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 11, 2023
Revised: November 28, 2023
Accepted: January 22, 2024
Published: January 24, 2024

REFERENCES

1. Timmis, A., Vardas, P., Townsend, N., Torbica, A., Katus, H., De Smedt, D., Gale, C.P., Maggioni, A.P., Petersen, S.E., Huculeci, R., et al. (2022). European Society of Cardiology: cardiovascular disease statistics 2021. Eur. Heart J. *43*, 716–799. https://doi.org/10.1093/eurheartj/ehab892.
2. Piepoli, M.F., Hoes, A.W., Agewall, S., Albus, C., Brotons, C., Catapano, A.L., Cooney, M.-T., Corrà, U., Cosyns, B., Deaton, C., et al. (2016). 2016 European Guidelines on cardiovascular disease prevention in clinical practice. Rev. Esp. Cardiol. *69*, 939. https://doi.org/10.1016/j.rec.2016.09.009.
3. D'Agostino, R.B., Vasan, R.S., Pencina, M.J., Wolf, P.A., Cobain, M., Massaro, J.M., and Kannel, W.B. (2008). General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation *117*, 743–753. https://doi.org/10.1161/CIRCULATIONAHA.107.699579.
4. Goff, D.C., Jr., Lloyd-Jones, D.M., Bennett, G., Coady, S., D'Agostino, R.B., Gibbons, R., Greenland, P., Lackland, D.T., Levy, D., O'Donnell, C.J., et al. (2014). 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation *129*, S49–S73. https://doi.org/10.1161/01.cir.0000437741.48606.98.
5. Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine Learning in Medicine. N. Engl. J. Med. *380*, 1347–1358. https://doi.org/10.1056/NEJMra1814259.
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Adv. Neural Inf. Process. Syst. *33*, 1877–1901.
7. GPT-4. https://openai.com/research/gpt-4.
8. Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., and Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit. Health *2*, e0000198. https://doi.org/10.1371/journal.pdig.0000198.
9. Rao, A., Pang, M., Kim, J., Kamineni, M., Lie, W., Prasad, A.K., Landman, A., Dreyer, K., and Succi, M.D. (2023). Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. J. Med. Internet Res. *25*, e48659. https://doi.org/10.2196/48659.
10. Sarraju, A., Bruemmer, D., Van Iterson, E., Cho, L., Rodriguez, F., and Laffin, L. (2023). Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. JAMA *329*, 842–844. https://doi.org/10.1001/jama.2023.1044.
11. Ye, W., Ou, M., Li, T., Chen, Y., Ma, X., Yanggong, Y., Wu, S., Fu, J., Chen, G., Wang, H., et al. (2023). Assessing Hidden Risks of LLMs: An Empirical Study on Robustness, Consistency, and Credibility. Preprint at arXiv. https://doi.org/10.48550/ARXIV.2305.10235.
12. Harrer, S. (2023). Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine *90*, 104512. https://doi.org/10.1016/j.ebiom.2023.104512.
13. Thirunavukarasu, A.J. (2023). Large language models will not replace healthcare professionals: curbing popular fears and hype. J. R. Soc. Med. *116*, 181–182. https://doi.org/10.1177/01410768231173123.
14. Kim, Y., and Han, B.-G.; KoGES group (2017). Cohort Profile: The Korean Genome and Epidemiology Study (KoGES) Consortium. Int. J. Epidemiol. *46*, 1350. https://doi.org/10.1093/ije/dyx105.
15. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779. https://doi.org/10.1371/journal.pmed.1001779.

16. Briganti, G., and Le Moine, O. (2020). Artificial Intelligence in Medicine: Today and Tomorrow. Front. Med. *7*, 27. https://doi.org/10.3389/fmed.2020.00027.
17. Kaul, V., Enslin, S., and Gross, S.A. (2020). History of artificial intelligence in medicine. Gastrointest. Endosc. *92*, 807–812. https://doi.org/10.1016/j.gie.2020.06.040.
18. Lee, P., Bubeck, S., and Petro, J. (2023). Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N. Engl. J. Med. *388*, 1233–1239. https://doi.org/10.1056/NEJMsr2214184.
19. Dave, T., Athaluri, S.A., and Singh, S. (2023). ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front. Artif. Intell. *6*, 1169595. https://doi.org/10.3389/frai.2023.1169595.
20. Haupt, C.E., and Marks, M. (2023). AI-Generated Medical Advice-GPT and Beyond. JAMA *329*, 1349–1350. https://doi.org/10.1001/jama.2023.5321.
21. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., and Ting, D.S.W. (2023). Large language models in medicine. Nat. Med. *29*, 1930–1940. https://doi.org/10.1038/s41591-023-02448-8.
22. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language Models are Few-Shot Learners. Preprint at arXiv. https://doi.org/10.48550/ARXIV.2005.14165.
23. OpenAI. (2023). GPT-4 Technical Report. Preprint at arXiv. https://doi.org/10.48550/ARXIV.2303.08774.
24. Nori, H., King, N., McKinney, S.M., Carignan, D., and Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. Preprint at arXiv. https://doi.org/10.48550/ARXIV.2303.13375.
25. Goodman, R.S., Patrinely, J.R., Stone, C.A., Jr., Zimmerman, E., Donald, R.R., Chang, S.S., Berkowitz, S.T., Finn, A.P., Jahangir, E., Scoville, E.A., et al. (2023). Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw. Open *6*, e2336483. https://doi.org/10.1001/jamanetworkopen.2023.36483.
26. Ayers, J.W., Poliak, A., Dredze, M., Leas, E.C., Zhu, Z., Kelley, J.B., Faix, D.J., Goodman, A.M., Longhurst, C.A., Hogarth, M., and Smith, D.M. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern. Med. *183*, 589–596. https://doi.org/10.1001/jamainternmed.2023.1838.
27. Adams, L.C., Truhn, D., Busch, F., Kader, A., Niehues, S.M., Makowski, M.R., and Bressem, K.K. (2023). Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. Radiology *307*, e230725. https://doi.org/10.1148/radiol.230725.
28. Arnett, D.K., Blumenthal, R.S., Albert, M.A., Buroker, A.B., Goldberger, Z.D., Hahn, E.J., Himmelfarb, C.D., Khera, A., Lloyd-Jones, D., McEvoy, J.W., et al. (2019). 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation *140*, e596–e646. https://doi.org/10.1161/CIR.0000000000000678.
29. Kıcıman, E., Ness, R., Sharma, A., and Tan, C. (2023). Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.00050.
30. Kavousi, M., Leening, M.J.G., Nanchen, D., Greenland, P., Graham, I.M., Steyerberg, E.W., Ikram, M.A., Stricker, B.H., Hofman, A., and Franco, O.H. (2014). Comparison of application of the ACC/AHA guidelines, Adult Treatment Panel III guidelines, and European Society of Cardiology guidelines for cardiovascular disease prevention in a European cohort. JAMA *311*, 1416–1423. https://doi.org/10.1001/jama.2014.2632.
31. Anderson, T.J., Grégoire, J., Pearson, G.J., Barry, A.R., Couture, P., Dawes, M., Francis, G.A., Genest, J., Jr., Grover, S., Gupta, M., et al. (2016). 2016 Canadian Cardiovascular Society Guidelines for the Management of Dyslipidemia for the Prevention of Cardiovascular Disease in the Adult. Can. J. Cardiol. *32*, 1263–1282. https://doi.org/10.1016/j.cjca.2016.07.510.
32. Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. (2020). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Preprint at arXiv. https://doi.org/10.48550/ARXIV.2009.13081.
33. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. Nature *620*, 172–180. https://doi.org/10.1038/s41586-023-06291-2.
34. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al. (2023). Towards expert-level medical question answering with large language models. Preprint at arXiv. https://doi.org/10.48550/ARXIV.2305.09617.
35. Grupac, M., Zauskova, A., and Nica, E. (2023). Generative artificial intelligence-based treatment planning in clinical decision-making, in precision medicine, and in personalized healthcare. Contemp. Read. Law Soc. Justice *15*, 45.
36. Peters, M.A., Jackson, L., Papastephanou, M., Jandrić, P., Lazaroiu, G., Evers, C.W., Cope, B., Kalantzis, M., Araya, D., Tesar, M., et al. (2023). AI and the future of humanity: ChatGPT-4, philosophy and education – Critical responses. Educ. Philos. Theor. 1–35. https://doi.org/10.1080/00131857.2023.2213437.
37. Kovacova, M., Kevicky, F., and Popescu, G.H. (2023). Generative artificial intelligence-driven healthcare systems in patient record analysis, in disease diagnosis and monitoring, and in customized treatment plans. Contemp. Read. Law Soc. Justice *15*, 152.
38. Zhang, Z., Zhang, A., Li, M., and Smola, A. (2022). Automatic Chain of Thought Prompting in Large Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2210.03493.
39. Yang, X., Chen, A., PourNejatian, N., Shin, H.C., Smith, K.E., Parisien, C., Compas, C., Martin, C., Costa, A.B., Flores, M.G., et al. (2022). A large language model for electronic health records. NPJ Digit. Med. *5*, 194. https://doi.org/10.1038/s41746-022-00742-2.
40. Jiang, L.Y., Liu, X.C., Nejatian, N.P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H.A., Laufer, I., Punjabi, P., et al. (2023). Health system-scale language models are all-purpose prediction engines. Nature *619*, 357–362. https://doi.org/10.1038/s41586-023-06160-y.
41. Steinfeldt, J., Buergel, T., Loock, L., Kittner, P., Ruyoga, G., Zu Belzen, J.U., Sasse, S., Strangalies, H., Christmann, L., Hollmann, N., et al. (2022). Neural network-based integration of polygenic and clinical information: development and validation of a prediction model for 10-year risk of major adverse cardiac events in the UK Biobank cohort. Lancet. Digit. Health *4*, e84–e94. https://doi.org/10.1016/S2589-7500(21)00249-1.
42. OpenAI Platform. https://platform.openai.com/docs/api-reference/chat/create.
43. Rademaker, D.T., Xue, L.C., 't Hoen, P.A.C., and Vriend, G. (2022). Entropy and Variability: A Second Opinion by Deep Learning. Biomolecules *12*, 1740. https://doi.org/10.3390/biom12121740.
44. DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics *44*, 837–845.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
UK Biobank dataset	https://www.ukbiobank.ac.uk/	The UK Biobank data access has been approved. The application number is 85037.
KoGES dataset	https://nih.go.kr/eng/	The KoGES data access has been approved. The application number is 6635-302.
Software and algorithms		
Python	https://www.python.org/	Version 3.10.6
scikit-learn	https://scikit-learn.org/	Version 1.2.2
Scipy	https://scipy.org/	Version 1.10.1
lifelines	https://pypi.org/project/lifelines/	Version 0.27.7
matplotlib	https://matplotlib.org/	Version 3.7.1
openai	https://openai.com/blog/openai-api/	Version 0.27.8
gpt-4	https://platform.openai.com/docs/models/gpt-4/	Version gpt-4-0613
gpt-3.5	https://platform.openai.com/docs/models/gpt-3-5/	Version gpt-3.5-turbo-0613
R	https://www.r-project.org/	Version 4.2.0
pROC	https://github.com/cran/pROC/	Version 1.18.4

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Dukyong Yoon (dukyong.yoon@yonsei.ac.kr).

Materials availability

This study did not generate new materials.

Data and code availability

The UK Biobank data access has been approved. The application number for UK Biobank is 85037. And The KoGES data access has been approved. The application number for KoGES is 6635-302.

All original code has been deposited at https://github.com/CMI-Laboratory/GPTCVD/ and is publicly available as of the date of publication. DOIs are listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Participant data information

The UK Biobank is a large-scale prospective cohort study initiated in 2006 to investigate the genetic and non-genetic determinants of diseases common in middle-aged and older populations. The study compiled health data from around 500,000 participants, ranging in ages from 40 to 70 years, from diverse socioeconomic and ethnic backgrounds across the United Kingdom. The collection process spanned across England, Scotland, and Wales at 22 assessment centers. The assessment involved both self-completed touchscreen questionnaires and face-to-face interviews to garner detailed information. Additionally, physical and functional measures were taken, and a variety of biological samples were collected.[15]

The KoGES is a community-based cohort study that has collected health-related information from over 10,000 Korean participants aged 40 and over since 2001. This study conducts recurrent health screenings and surveys every 2 to 4 years and employs passive follow-up through linkage with other national databases.[14]

In this study, we used data from 47,468 participants from the UK Biobank and 5,718 participants from KoGES, focusing on completeness of records for cardiovascular disease diagnosis, age, sex, cholesterol levels, systolic and diastolic blood pressure, body mass index, smoking

status, hypertension medication usage, and diabetes status. More detailed information related to participants is provided in the "method details" section of this paper.

METHOD DETAILS

Data sources and outcome

We used data from the UK Biobank cohort, a large-scale biomedical database of the UK general population. Established in 2006, the UK Biobank cohort is one of the major international health resources that has collected extensive data and biological samples from approximately 500,000 participants aged between 40 and 69 years at the time of assessment (Figure S1). We used the UK Biobank database to extract data pertaining to age, sex, diabetes diagnosed by a doctor, blood pressure medication, smoking status, total cholesterol, HDL cholesterol, LDL cholesterol, triglycerides, SBP, DBP, standing height, weight, date of attending the assessment center, and date of death.

Utilizing the UK Biobank, the outcome was 10-year CVD risk defined by MACE, which represents the most fatal and predominant occurrence of CVD.[41] A MACE is defined as a composite outcome comprising myocardial infarction or ischemic stroke. For extracting a MACE, we employed an outcome variable known as first occurrences, which consolidates data from various sources within the UK Biobank, including primary care and hospital inpatient records, the death register, and self-reported medical conditions. This outcome is organized based on the earliest recorded instance for each condition, as classified by the International Classification of Diseases, 10th Revision (ICD-10) codes. For our analysis, we focused on the ICD-10 codes I21, I22, I23, I24, and I25 for fatal or non-fatal myocardial infarction, and I63 and I64 for ischemic stroke. A total of 103,817 participants with missing data were excluded (Figure S1). Of the remaining participants, after randomly selecting 50,000 participants, 2,532 patients who had previously experienced MACEs were further excluded, leaving 47,468 subjects for the analysis.

In addition, we used KoGES data as an additional validation cohort. The KoGES is a large-scale prospective study designed to investigate the genetic and environmental factors contributing to chronic diseases in the Korean population.[14] We used baseline data from the KoGES cohort collected between 2001 and 2002 to extract variables analogous to those used in the UK Biobank. These variables included age, sex, diagnosis of diabetes by a physician, blood pressure medication use, smoking status, total cholesterol, HDL cholesterol, triglycerides, SBP, DBP, height, and weight. LDL cholesterol levels were calculated based on total cholesterol, HDL cholesterol, and triglyceride levels. In alignment with the criteria established by the KoGES, the onset of MACE was defined as the occurrence of either myocardial infarction or ischemic stroke subsequent to the investigation date. Out of an initial 10,030 participants, 875 with missing data were excluded (Figure S2). An additional 3,437 participants, who were not followed up at the 10-year mark, were also omitted, resulting in a final cohort of 5,718 subjects for subsequent analysis.

Cardiovascular risk calculation with conventional risk prediction models

The Framingham and the ACC/AHA risk scores are widely accepted algorithms for estimating an individual's 10-year risk of developing CVD. For the calculation of these risk scores, we referred to the guidelines outlined by Anderson et al. (2016) for Framingham risk score and Arnett et al. (2019) for ACC/AHA risk score.[28,31] These risk assessment tools incorporate multiple variables such as age, sex, blood pressure, cholesterol levels, smoking status, and the presence of diabetes to generate a risk percentage. For the Framingham risk score, individuals were categorized into low, moderate, or high-risk groups based on calculated risk percentages, utilizing thresholds of 10% and 20%. On the other hand, the ACC/AHA risk score used thresholds of 7.5% and 20%. It should be noted that the original ACC/AHA risk score guidelines categorize risk as lowest for scores below 5% and borderline for scores between 5% and 7.5%. For the purpose of this study, we have simplified the categories to low, moderate, and high risk and have considered scores below 7.5% as low risk.

Cardiovascular risk prediction leveraging GPT-3.5-turbo and GPT-4

To predict the incidence of CVDs using GPT, we transformed the variables into a sentence structure, as exemplified in Figure 1. The decision to use this conversion was based on the inherent language model nature of LLMs. We prompted the GPT to answer only the risk percentage rather than extensive text narratives. Information on each participant (age, sex, diabetes, hypertension, smoking status, total cholesterol, LDL cholesterol, HDL cholesterol, triglycerides, systolic blood pressure, diastolic blood pressure, and BMI (calculated from height and weight) was provided to the LLMs, and the 10-year CVD risk percentage was extracted using regular expressions from the corresponding answers. Based on the 10-year CVD risk percentage, <10% was classified as low risk, 10% and <20% as moderate risk, and >20% as high risk.

Determining the optimal temperature settings for GPT-4

For using GPT, we used the OpenAI application programming interface (API) (GPT-3.5-turbo and GPT-4) in a Python environment to streamline the extraction of results. In the GPT model, various hyperparameters are available to control the variability of responses, notably the 'temperature' and 'top-k' settings. The temperature parameter in GPT-4 adjusts the model's output diversity, acting as a measure of 'creative freedom', while top-k limits the word choices to enhance prediction accuracy.[42] Given that the significance of the top-k setting tends to be greater when the temperature is high, we chose to keep the top-k setting at its default value for this experiment. We aimed to find the optimal temperature settings for GPT-4 as follows:

In the UK Biobank cohort, prior to our main experiments involving 47,468 subjects as depicted in Figure S1, we initially selected a random sample of 2,000 individuals from this 47,468 subjects to determine the optimal temperature settings for GPT-4 and to assess GPT's response

variability. We adjusted the temperature in increments of 0.2, ranging from 0 to 1, and for each setting, we performed five iterations of the 10-year CVD risk prediction using GPT on the same sample. Firstly, we computed the average risk score from the five iterations for each subject and then calculated the AUROC for this average risk score in predicting 10-year CVD incidence. Secondly, we determined the coefficient of variation (CV) for the iterations per subject and calculated the average CV across all subjects to quantify the variability of the GPT-based risk score. Third, we calculated the entropy of the average risk score from the five iterations for each subject to statistically assess the spread and distribution of the GPT-based predictions.[43] The calculated AUROC, CV and entropy values are shown in Figure S4A. The CV and entropy increased gradually with higher temperature settings: from 0.087 to 8.233 at a temperature of 0.0, to 0.179 and 10.021 at a temperature of 0.4, and 0.290 and 10.196 at a temperature of 1.0. The highest AUROC was observed at a temperature of 0.4 (0.735), while the lowest was at a temperature of 1.0 (0.717).

To determine the optimal temperature setting, we juxtaposed CV, entropy, and AUROC within a single graph (Figure S4B). To align the trends of CV, in which an increase in value indicates increased variability of the GPT-based risk score, with that of entropy, in which an increase in value indicates a more evenly distributed spread of the predicted scores (more fine-grained), and AUROC, in which an increase in value indicates an increase in GPT's 10-year CVD prediction performance, we inverted the CV values, representing them as -CV in Figure S4B for a coherent comparison. A tradeoff between entropy and -CV values can be observed as temperature increases. We identified the temperature setting of 0.4 as the confluence point where entropy and -CV intersect, also with the highest AUROC values, signifying it as the optimal juncture for the model's performance, as elucidated in Figure 2B. Examining the distribution of answers via scatterplots at different temperature settings, we observed that at lower temperatures such as 0.0 or 0.2, the answers were not uniformly distributed. Instead, they were more likely to cluster around specific values (e.g., 10, 20), which we call the 'streaking' phenomenon, as illustrated in Figures S4A and S5. GPT-4 risk score derived from an experiment with one iteration from the samples resulted in a more accentuated 'streaking' than GPT-4 risk score derived from an experiment with five iterations from the samples (Figure S6).

For the evaluation of GPT-4's 10-year CVD risk prediction capabilities using the optimal temperature setting determined by the above procedure, we conducted five iterations for both GPT-3.5-turbo and GPT-4 models on the selected UK Biobank and KoGES study populations (Figures S1 and S2), comparing their predictive performance for MACEs against established models such as the Framingham and ACC/AHA risk scores. The average risk score from the five iterations for each subject was determined to be the final risk score.

Performance evaluation

To assess the predictive accuracy of each scoring method for MACE, receiver operating characteristic (ROC) curve analysis was conducted. The AUROC and AUPRC was calculated to quantify the overall discriminative ability of each scoring system. The risk threshold for calculating accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) was set at 20%, which was used to differentiate the high-risk group from the rest in all risk scoring methods.

QUANTIFICATION AND STATISTICAL ANALYSIS

To rigorously assess the statistical significance of differences in baseline characteristics among the risk groups, we conducted tests tailored to the characteristics of each variable. As all continuous variables failed to meet the criteria for normality as assessed by the Shapiro–Wilk method, we employed the Kruskal–Wallis test for comparisons across different risk groups. For categorical variables, the chi-squared test was used to evaluate the statistical significance of differences among the groups. To compare the AUROCs between different scoring systems, the DeLong test was conducted.[44] The relationships between the scoring systems (GPT-4, GPT-3.5-turbo, Framingham, and ACC/AHA risk scores) were evaluated by plotting scatterplots and calculating Pearson's correlation coefficient. The Kaplan–Meier method was applied to plot survival curves for the low-, moderate-, and high-risk groups based on the risk scoring methods. To statistically compare the survival functions across these risk groups, the pairwise log rank test with post-hoc Bonferroni correction was utilized. A p-value <0.05 was considered significant in all tests.