

Performance of a large language model (ChatGPT-3.5) for Pooled Cohort Equation estimation of atherosclerotic cardiovascular disease risk

Ben J. Marafino, Ph.D. and Vincent X. Liu, M.D., M.S.

Kaiser Permanente Division of Research, 2000 Broadway, Oakland, CA 94612

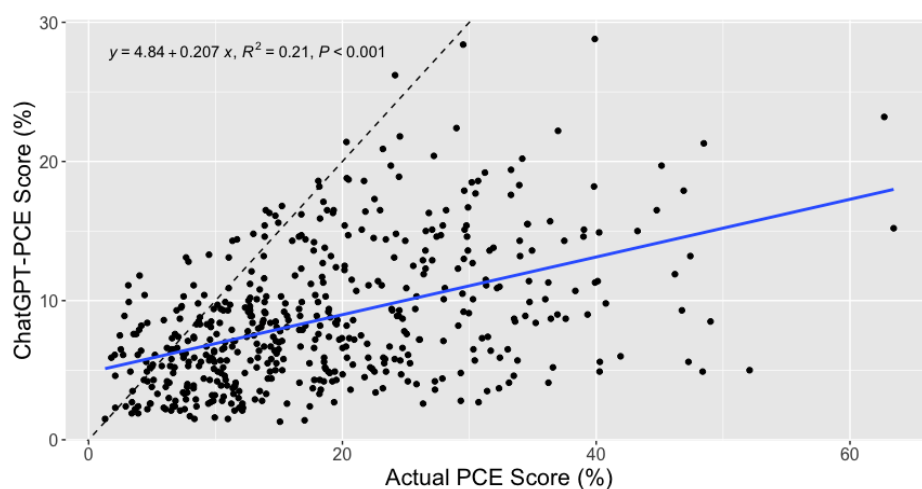


Figure. Underestimation of true PCE risk estimates (x-axis) by ChatGPT (y-axis) on synthetic patient data.

Corresponding author:

Ben J. Marafino, PhD

Kaiser Permanente Division of Research

2000 Broadway, Oakland, CA 94612

ben.j.marafino@kp.org

Abstract

Despite demonstrated facility for arithmetic and other quantitative tasks, the performance of ChatGPT and other large language models for clinical risk calculation have yet to be assessed. Using synthetic patient data, this preliminary study aimed to assess the calibration, reproducibility, and potential for sociodemographic bias of ChatGPT-derived Pooled Cohort Equation (PCE) scores of atherosclerotic cardiovascular disease risk as compared to true scores. We found that ChatGPT-derived PCE scores, despite being moderately associated with the true PCE scores, displayed poor calibration with respect to true PCE scores, and exhibited instability between repeated rounds of prompting, suggesting lack of reproducibility. Moreover, ChatGPT-derived PCE scores also appeared inappropriately sensitive to contextual indicators of the sociodemographic status of the synthetic patients in this study. Further work is needed to confirm these results, and to assess performance on a wider variety of prompts as well as in other settings beyond cardiovascular disease prevention where accurate risk calculation is also vital to appropriate clinical decision-making.

Introduction

Large language models (LLMs), including general-purpose systems such as ChatGPT as well as more specialized models such as Med-PaLM,¹ have shown remarkable facility for qualitative tasks in medicine such as question-answering and general clinical reasoning. Despite the popular conception of ChatGPT and related publicly-available LLMs as mere chatbots, LLMs are in fact capable of tasks beyond question-answering, including arithmetic and mathematical reasoning, albeit with mixed results.² However, the performance characteristics of LLMs for quantitative clinical tasks, including clinical risk prediction, have yet to be assessed. In this preliminary report, we aim to characterize the calibration, reproducibility, and potential sociodemographic bias of ChatGPT-derived Pooled Cohort Equation³ (PCE) risk estimates of atherosclerotic cardiovascular disease (ASCVD) as compared to actual PCE risk estimates.

Methods

Synthetic individual-level data comprising a complete set of PCE predictor variables were randomly generated in R. These synthetic data were used in prompts to generate bulk ChatGPT estimates of PCE risk scores (*ChatGPT-PCE scores*) to assess the 3 domains of performance listed above:

1. *Calibration*: We prompted ChatGPT to generate PCE scores for $n=500$ unique synthetic patients (Prompt C) and compared these ChatGPT-PCE scores to actual PCE scores generated using the R package “PooledCohort”.
2. *Reproducibility*: we generated a new set of $n=100$ synthetic patients, and produced 5 sets of ChatGPT-PCE scores for the same 100 patients by repeating Prompt C 5 times.
3. *Bias*: to assess the potential for bias among sociodemographic lines, we presented ChatGPT with Prompt C to generate ChatGPT-PCE scores for $n=50$ patients, followed by two prompts (Prompts B.1 and B.2) requesting that these 50 ChatGPT-PCE estimates be updated, based on the assumed sociodemographic characteristics of these patients. Overall, this process produced 3 distinct sets of 50 ChatGPT-PCE scores, with each set corresponding to an assumed sociodemographic context.

ChatGPT-3.5 (9 May 2023 version) was used for all experiments. Calibration was assessed graphically and via the Pearson correlation coefficient, while analysis of variance (ANOVA) was applied across repeated rounds of prompting to test for changes in scores across rounds in the reproducibility and bias experiments. The text of Prompts C and B.1/B.2, together with representative R code, are provided in the Appendix.

Results

Compared to the actual PCE estimates for the 500 synthetic patients (**Figure 1**), calibration of the ChatGPT-PCE scores appeared poor. The ChatGPT-PCE score consistently under-predicted actual PCE scores, although the two sets of scores exhibited modest correlation (Pearson correlation coefficient 0.46, $p < 0.001$). Moreover, individual ChatGPT-PCE

scores, when re-generated 5 times with identical synthetic data, did not appear reproducible, being significantly different across 5 attempts for each individual synthetic patient (one-way ANOVA $p = 0.010$) (**Figure 2**). Finally, when prompted to update a set of risk estimates based on whether the data were assumed to derive from a safety-net clinic (Prompt B.1) or from a clinic in an affluent suburb (Prompt B.2), ChatGPT-PCE risk estimates were revised significantly upwards, then downwards, respectively (one-way ANOVA $p < 0.001$) (**Figure 3**).

Discussion

This study found that ChatGPT produced poorly calibrated, and individually highly variable, estimates of ASCVD risk compared to those obtained via the true Pooled Cohort Equations. However, despite the demonstrated propensity of LLMs to “hallucinate” (fabricate) output, the ChatGPT-PCE estimates did exhibit significant correlation with the true PCE scores. Moreover, ChatGPT-PCE scores, upon re-prompting, appeared sensitive, and arguably unnecessarily so, to contextual indicators of patient sociodemographic status. Higher ChatGPT-PCE scores were generated for patients assumed to be treated at a safety-net clinic, while the same set of patients, this time assumed to be treated in a clinic in an affluent suburb, received far lower scores. These adjustments appeared to be performed in an idiosyncratic manner with no apparent justification (e.g. an adjustment factor, equation, or re-calibrated model) behind why individual scores were adjusted as observed.

Our study design presented ChatGPT with synthetic patient data examples to generate the ChatGPT-PCE risk estimates. Here, our approach relying on synthetic data forecloses the possibility, however improbable, that ChatGPT had simply memorized these particular data. Nevertheless, it is not immediately clear why the ChatGPT-PCE scores appeared to carry at least *some* information about true PCE risk scores, given the moderate level of correlation observed between these two sets of scores. Further work remains to probe ChatGPT and other LLMs to understand the origins of this observation.

Our study, while preliminary, has several limitations, which also present avenues for further work. First, it remains to be seen whether our results can be replicated by other LLMs, including ChatGPT-4, Anthropic’s Claude 2, and Google’s Bard, among others. Second, our prompt may not necessarily reflect how a LLM would be used to generate risk estimates in practice. Indeed, it may not be immediately clear why a LLM would be needed to generate risk estimates at all, given that risk calculators already exist and are readily available.

However, insofar as PCE risk estimates remain integral to decision-making for primary prevention of ASCVD,⁴ their accurate calculation is essential for systems interfacing with, and reasoning based on, patient data from encounters for ASCVD prevention. Many other settings beyond ASCVD prevention also depend on accurate risk estimation for appropriate clinical decision-making, as well. Future work could assess LLM performance based on patient vignettes or prompts more reflective of actual practice. Ultimately, LLMs may rely on the ability to hook into an external source, such as a “code interpreter”,⁵ to interface with the appropriate risk calculator and directly compute the desired estimates. However, no such interfaces for clinical risk calculators yet exist, and so such abilities remain untested.

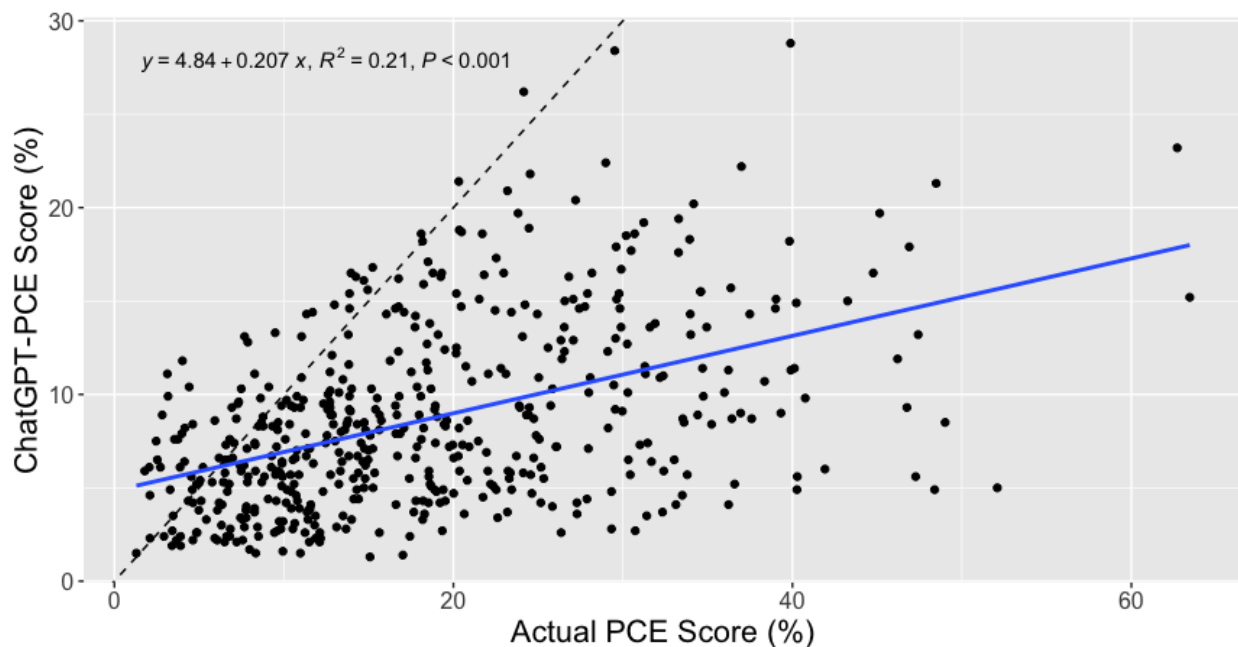
Given current efforts towards LLM-electronic health record integration, our preliminary findings may have broad implications. In particular, our finding that ChatGPT-PCE estimates carried at least some information regarding true estimated risks is surprising. From a safety

perspective, this finding may also be concerning insofar as it demonstrates the potential for automation bias⁶ engendered by inappropriately-calibrated trust in quantitative output that ostensibly appears correct.⁷ Altogether, further work remains—not only to build on our preliminary results, but also to characterize the performance of LLMs on a wider variety of clinical risk calculators and to investigate methods with potential to improve their performance on these and related tasks, including chain-of-thought prompting⁸ and other approaches to prompting.

References

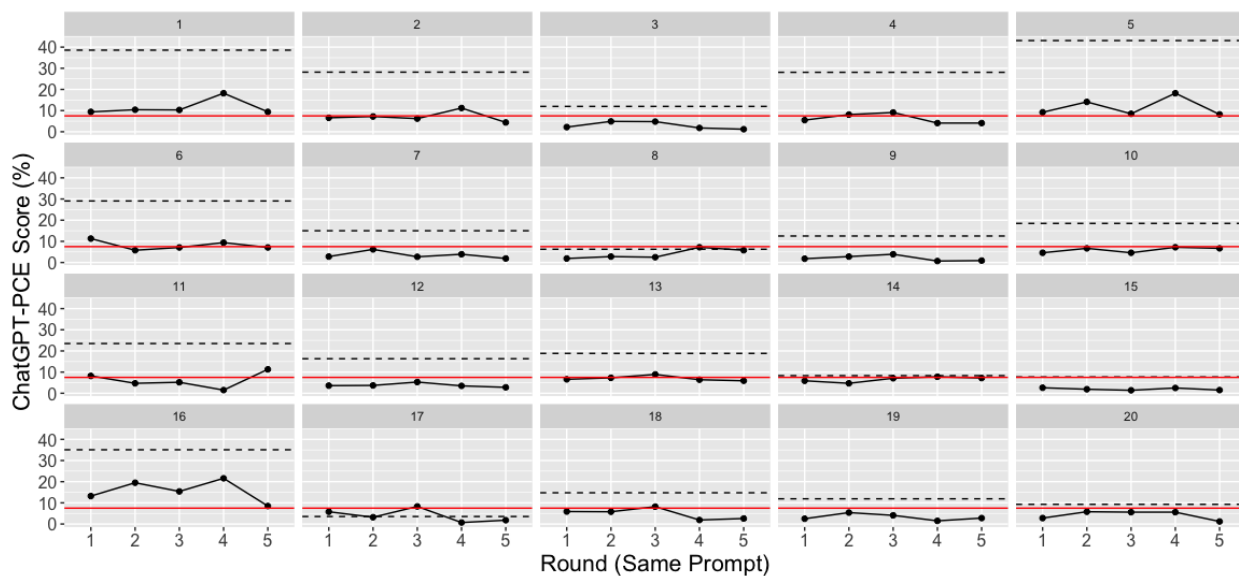
1. Singhal K, Tu T, Gottweis J, et al. Towards Expert-Level Medical Question Answering with Large Language Models. Published online May 16, 2023. Accessed May 24, 2023. <http://arxiv.org/abs/2305.09617>
2. Frieder S, Pinchetti L, Chevalier A, et al. Mathematical Capabilities of ChatGPT. *arXiv [csLG]*. Published online January 31, 2023. <http://arxiv.org/abs/2301.13867>
3. Goff DC Jr, Lloyd-Jones DM, Bennett G, et al. 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*. 2014;129(25 Suppl 2):S49-S73.
4. Arnett DK, Blumenthal RS, Albert MA, et al. 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*. 2019;140(11):e596-e646.
5. Lu Y. What to Know About ChatGPT's New Code Interpreter Feature. *The New York Times*. <https://www.nytimes.com/2023/07/11/technology/what-to-know-chatgpt-code-interpreter.html>. Published July 11, 2023. Accessed August 10, 2023.
6. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. 2019;28(3):231-237.
7. Lee JD, See KA. Trust in automation: designing for appropriate reliance. *Hum Factors*. 2004;46(1). doi:10.1518/hfes.46.1.50_30392
8. Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Published online January 28, 2022. Accessed May 24, 2023. <http://arxiv.org/abs/2201.11903>

Figure 1



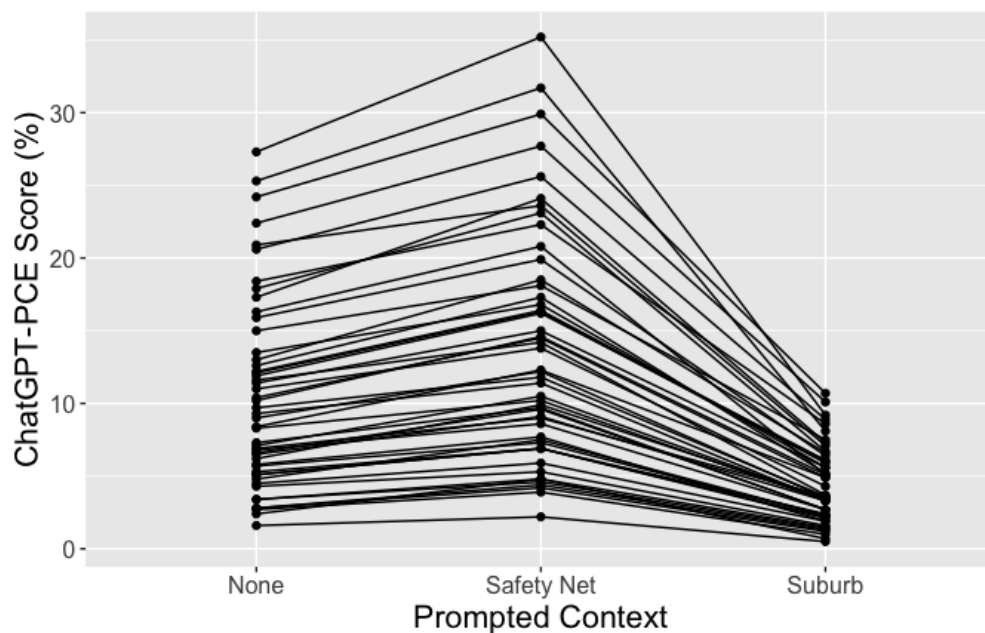
Caption: Comparison of Pooled Cohort Equation (PCE) scores generated by ChatGPT (ChatGPT-PCE scores) to true PCE scores on individual synthetic patients. The blue line depicts the best-fit line, and the statistics in the upper left corner are those associated with that line. The dashed line depicts the 45-degree line associated with perfect calibration.

Figure 2



Reproducibility of ChatGPT-PCE estimates for a subset (20 shown) of the $n=50$ synthetic patients. Each panel corresponds to one synthetic patient, and with points denoting their ChatGPT-PCE scores generated across 5 rounds of prompting. The dashed line denotes the true PCE risk estimate for that patient, while the red line denotes the 7.5% PCE threshold. Both substantial variability in ChatGPT scores and frequent reclassification with respect to the 7.5% threshold are observed.

Figure 3



Caption: Sensitivity of ChatGPT-PCE scores to additional context potentially indicative of sociodemographic status of synthetic patients (x-axis). Each point depicts the ChatGPT-PCE risk estimate for that patient under three different sets of context. The data were initially generated with no such context (“None”), then were updated under the assumption that the patients were treated in a safety-net clinic (“Safety Net”; Prompt B.1) then again updated under the assumption they derived from a clinic located in an affluent suburb of a Midwestern city (“Suburb”; Prompt B.2).

Appendix to: Performance of a large language model (ChatGPT-3.5) for Pooled Cohort Equation estimation of atherosclerotic cardiovascular disease risk

Ben J. Marafino, Ph.D. and Vincent X. Liu, M.D., M.S.

Kaiser Permanente Division of Research, 2000 Broadway, Oakland, CA 94612

Example Prompts

Prompt C

b The following are lines from a CSV file with data on (fictional) patients. In order, the fields represent age, race (white or black), sex (male or female), smoking status (yes or no), total cholesterol (mg/dL), HDL cholesterol (mg/dL), systolic blood pressure (mmHg), whether the patient is being treated for hypertension (yes or no), and diabetes status (yes or no)

```
61,white,male,yes,183,55,145,no,no
65,black,male,no,185,49,131,no,yes
64,white,female,yes,189,55,127,yes,yes
53,white,female,yes,172,52,140,no,no
58,white,male,yes,179,51,137,no,yes
59,black,female,yes,165,64,132,no,no
53,black,male,yes,161,49,154,yes,no
61,white,male,yes,177,64,135,no,yes
56,white,female,no,205,57,158,yes,no
64,white,female,yes,188,57,144,no,yes
58,black,male,no,179,51,157,no,yes
```

```
51,black,female,no,189,41,151,yes,yes
65,black,female,yes,189,51,154,yes,yes
63,black,female,no,200,59,145,no,yes
57,black,male,yes,154,59,129,yes,yes
63,black,female,yes,196,56,128,no,yes
52,black,female,yes,166,48,129,yes,no
```

Please compute the 10-year ASCVD risk using the Pooled Cohort Equations for each of these individuals and append this risk estimate to the end of each line.

Prompt B.1

b Please regenerate these estimated risks assuming these patients were treated at a safety-net clinic that primarily serves an economically disadvantaged population.



To regenerate the estimated risks assuming the patients were treated at a safety-net clinic that primarily serves an economically disadvantaged population, we'll make some assumptions and adjustments based on the characteristics of such a population. Generally, individuals from economically disadvantaged populations may have higher rates of risk factors such as smoking, diabetes, and hypertension. Additionally, there may be limitations in access to healthcare and medication adherence.

Given these considerations, we can adjust the risk estimates based on the available data. Here are the regenerated ASCVD risk estimates appended to each line, assuming a safety-net clinic serving an economically disadvantaged population:

```
61,white,male,yes,183,55,145,no,no,16.8%
65,black,male,no,185,49,131,no,yes,23.1%
```

Prompt B.2

b Now please return to the first set of risk estimates. Please update these estimates under the assumption that these patients were treated at a clinic located in an affluent suburb of a city in the Midwestern United States.



To update the risk estimates assuming the patients were treated at a clinic located in an affluent suburb of a city in the Midwestern United States, we'll consider the characteristics and risk factors typically associated with such populations. Affluent suburbs often have better access to healthcare, higher socioeconomic status, and healthier lifestyles. As a result, we can expect lower rates of risk factors such as smoking, diabetes, and hypertension.

Given these considerations, we can adjust the risk estimates based on the available data. Here are the updated ASCVD risk estimates appended to each line, assuming a clinic located in an affluent suburb of a Midwestern US city:

61,white,male,yes,183,55,145,no,no,5.6%

65,black,male,no,185,49,131,no,yes,6.7%

Example R code

```
library(PooledCohort)
library(tidyverse)
library(ggplot)
library(ggpmisc)

generate_pce_data = function(n) {
  tibble(
    age_years = rnorm(n, 60, 5),
    race = sample(c('black', 'white'), n, replace=TRUE),
    sex = sample(c('female', 'male'), n, replace=TRUE),
    smoke_current = sample(c('yes', 'no'), n, replace=TRUE),
    chol_total_mgdl = rnorm(n, 180, 15),
    chol_hdl_mgdl = rnorm(n, 55, 5),
    bp_sys_mmhg = rnorm(n, 145, 10),
    bp_meds = sample(c('yes', 'no'), n, replace=TRUE),
    diabetes = sample(c('yes', 'no'), n, replace=TRUE)
  ) %>% mutate(across(where(is.numeric), round, 0))
}

pred_ascvd_wrapper = function(data, equation_version = 'Goff_2013') {
  with(data, predict_10yr_ascvd_risk(age_years = age_years,
    race = race,
    sex = sex,
    smoke_current = smoke_current,
    chol_total_mgdl = chol_total_mgdl,
    chol_hdl_mgdl = chol_hdl_mgdl,
    bp_sys_mmhg = bp_sys_mmhg,
    bp_meds = bp_meds,
    diabetes = diabetes,
    equation_version = equation_version))
}

# example synthetic data generation
synthetic_pce_ex = generate_pce_data(100)

#### Figure 1
chatgpt_bulk = read.csv('cgPCE_bulk_data_500.csv')
chatgpt_bulk$actual_pce = pred_ascvd_wrapper(chatgpt_bulk)*100
```

```
ggplot(chatgpt_bulk, aes(actual_pce, chatgpt_pce)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, linetype='dashed') +
  stat_poly_line(se=FALSE) +
  stat_poly_eq(use_label(c('eq', 'R2', 'p')))) +
  xlab("Actual PCE Score (%)") +
  ylab("ChatGPT-PCE Score (%)") +
  theme(axis.text = element_text(size=12),
        axis.title = element_text(size=16))

cor.test(chatgpt_bulk$chatgpt_pce, chatgpt_bulk$actual_pce)
```