Prognostic Risk Prediction using Large Language Models

Aniek Markus*, Tom Seinen*
***Shared first & last author**
Department of Medical Informatics, Erasmus MC, Rotterdam, The Netherlands

Background

Prognostic clinical prediction models help clinicians anticipate patient outcomes and guide interventions [1]. Traditionally, these models use statistical and supervised machine learning (ML) methods, such as logistic regression and random forests, to identify patterns in clinical data. Despite the development of many prognostic models [2], conventional approaches require extensive data processing, lack intuitive explanations, and are seldom implemented in practice [3,4].

The emergence of large language models (LLMs), pretrained on vast general and medical corpora, signals a new era for clinical risk prediction. Unlike ML models that estimate risk directly from observed data, LLMs can reason over patient information to address diverse prognostic tasks, potentially replacing the need for multiple specialized models. However, their effectiveness and reliability for prognostic prediction compared to traditional ML approaches remains unclear.

Recent studies show LLMs' promise for diagnostic prediction [5,6], but non-finetuned LLMs currently underperform compared to locally trained ML models for prognostic tasks, though their accuracy is improving [7,8]. Nevertheless, both proprietary and open-source LLMs demonstrate sociodemographic biases in risk predictions [9], and varied study methodologies hinder direct comparison.

The aim of our study is to evaluate LLM performance on patient-level prognostic prediction tasks using standardized analytics with the OMOP Common Data Model (CDM). We compare supervised ML and zero-shot LLM approaches for two clinical outcomes using primary care EHR data, demonstrate LLM integration into the standardized analytical pipeline, and assess potential biases.

Methods

Study design – An overview of the experimental setup is shown in **Figure 1**.

Data source and setting – We used the Integrated Primary Care Information (IPCI) database [10], containing observational EHR data from Dutch general practitioners (approx. 2.8 million patients, 1992–2022), mapped to the OMOP CDM. Approved by the IPCI governance board under codes 2022-02 and 2023-03.

Prediction task definition – Two tasks were defined (Figure 1, step 1): predicting dementia onset within five years in patients aged 55–85; and predicting five-year risk of heart failure or stroke in adults with type 2 diabetes. Target and outcome cohorts were constructed using ATLAS; detailed definitions are in Supplementary **Table S1**.

Supervised machine learning model – A supervised model was developed using the Patient-Level Prediction (PLP) standardized analytical pipeline [11]. Each patient's 47 clinical phenotype covariates, age, and sex were extracted over a 365-day observation period (see Supplementary **Table S2**). The binary outcome dataset was randomly split into a training set (75%) and a test set (25%). A default LASSO logistic regression model was trained and evaluated on the held-out test set.

Large language model – LLM-based prognostic prediction included several steps. The 47 binary covariates and the demographics were converted into text lists (Figure 1, step 3). The task definitions (cohort JSONs) were summarized into concise prompts using an LLM (Figure 1, step 4) and combined with time-at-risk and observation period details into a modular instruction prompt. The system prompt contained the task-specific instructions, and the user prompt contained the patient covariate lists. All prompts are in Supplementary **Table S3**. Using a custom PLP model extension, the LLM predicted risk probabilities for each patient in the test set (Figure 1, step 5). For this study, we selected the Qwen3-30B-A3B [12] LLM model because of its speed, reasoning capabilities, and relatively small size with strong performance.

Evaluation and bias assessment – Both approaches were evaluated on the same test set using receiver operating characteristic (ROC) curves and the area under the ROC curve (AUROC). Calibration was examined with calibration plots, and the predictive agreement was assessed by the correlation between predicted probabilities. We evaluated the potential bias by comparing differences in true positive rate (equal opportunity) and false positive rate (mistreatment) averaged over clinically relevant thresholds (<0.2) between male and female patients.

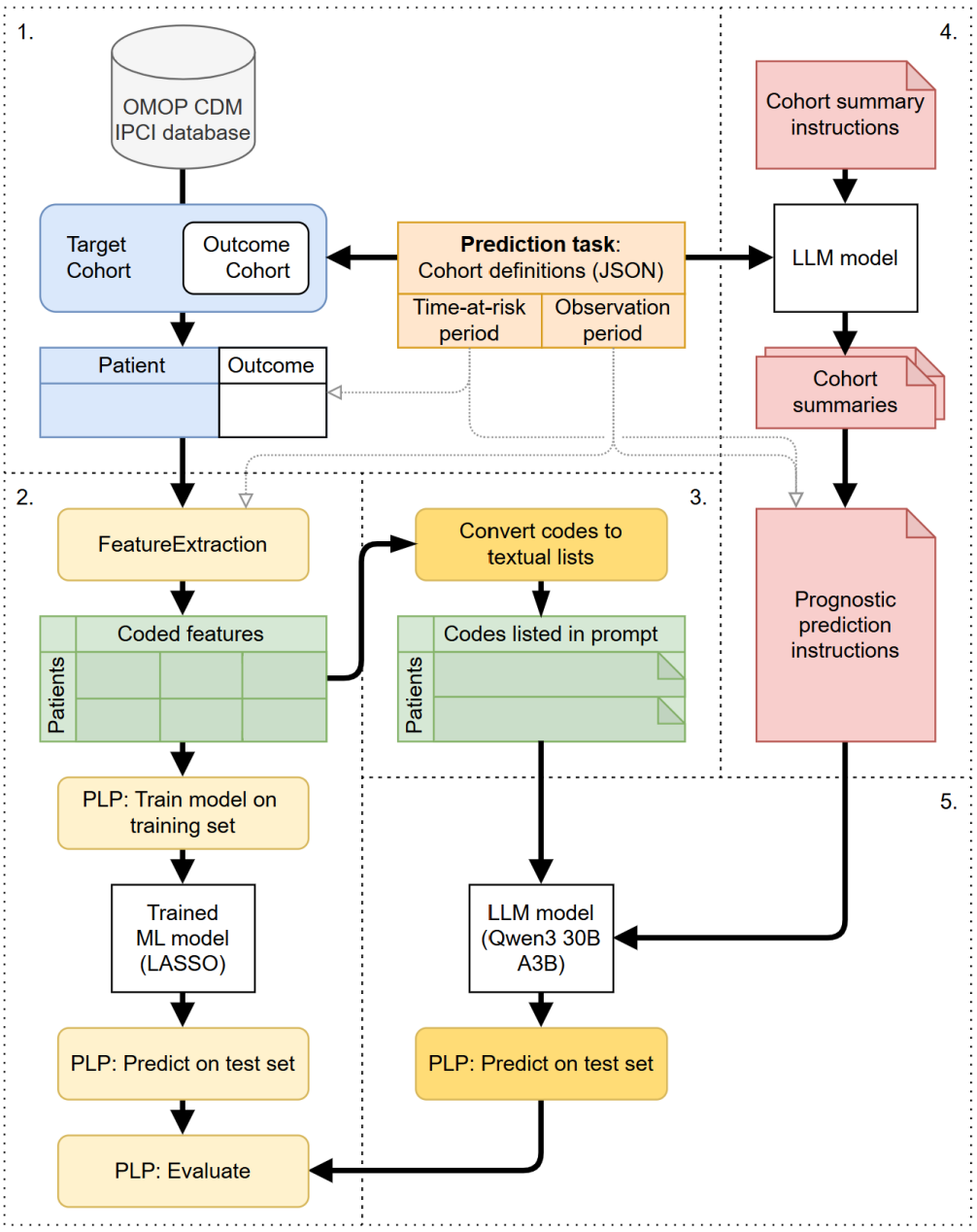

Figure 1. *Overview of the experimental setup. 1) We defined the prediction tasks by specifying the target and outcome cohorts, the time-at-risk, and extracting relevant patient data from the OMOP CDM database. 2) We applied the standardized analytical pipeline using the FeatureExtraction and PatientLevelPrediction (PLP) R packages to derive coded features for each patient, train supervised ML models, generate predictions, and evaluate model performance. 3) We transformed each patient's coded features into text-based prompts, listing all relevant codes for that individual. 4) We generated a concise cohort summary for each prediction task using an LLM, which was used to construct the prediction instruction prompt. 5) We prompted an LLM with both the patient-specific code lists and the task-specific prediction instructions, and evaluated the LLM's prognostic predictions using the same tools as the supervised ML approach.*

Results

Performance – We identified 30,985 patients for the dementia prediction task and 9,109 patients for the heart failure/stroke prediction task. The observed outcome rates were 2.29% and 6.81%, respectively. The predictive performance of the LASSO model and the LLM is shown in **Figure 2**. In terms of discriminatory power, the LASSO model outperformed the LLM in both prediction tasks. For dementia prediction, the LASSO achieved an AUROC of 0.841 compared to 0.742 for the LLM. Similarly, for heart failure/stroke prediction, the LASSO model attained an AUROC of 0.779, which was again higher than the LLM's AUROC of 0.684. Furthermore, the calibration plot indicates substantially better agreement between predicted and observed risks for the LASSO model compared to the LLM.

Figure 3 shows a moderate correlation (0.53–0.55) between the predicted probabilities of the two models, indicating that while some predictions are similar, there is notable variability. The LLM tended to produce a narrower range of predicted probabilities, often repeating similar values, suggesting a limitation in the model's capacity to generate detailed risk estimates.

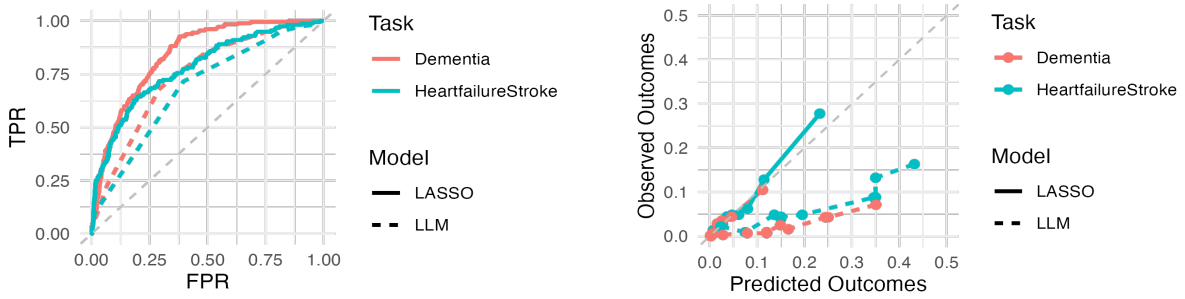

***Figure 2**. ROC curves (left) and calibration plots (right).*

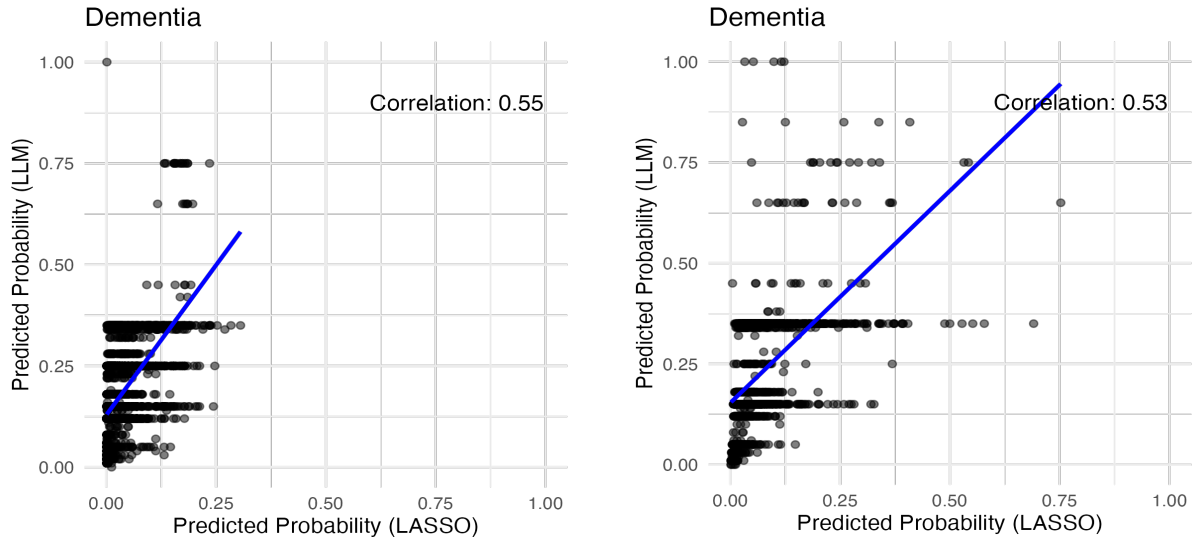

***Figure 3**. Predicted probabilities of LASSO versus LLM for predicting dementia (left) and heart failure/stroke (right).*

Bias – As shown in **Figure 4**, we compared differences in the types of errors across males and females. In general, TPR should be as high as possible (to avoid missed outcomes) and FPR as low as possible (to minimize false alarms). The left upper quadrant indicates positive bias toward females, and the right lower quadrant toward males, measured by equalized odds. The center depicts equal error rates on average. For heart failure/stroke prediction, the FPR was slightly better for females across both models. For dementia, differences in error rates between sexes were larger, with the LLM showing lower FPR/TPR for females and LASSO lower FPR/TPR for males. Overall, there were notable variations in group-fairness across the models, with greater disparity seen in the dementia prediction task.

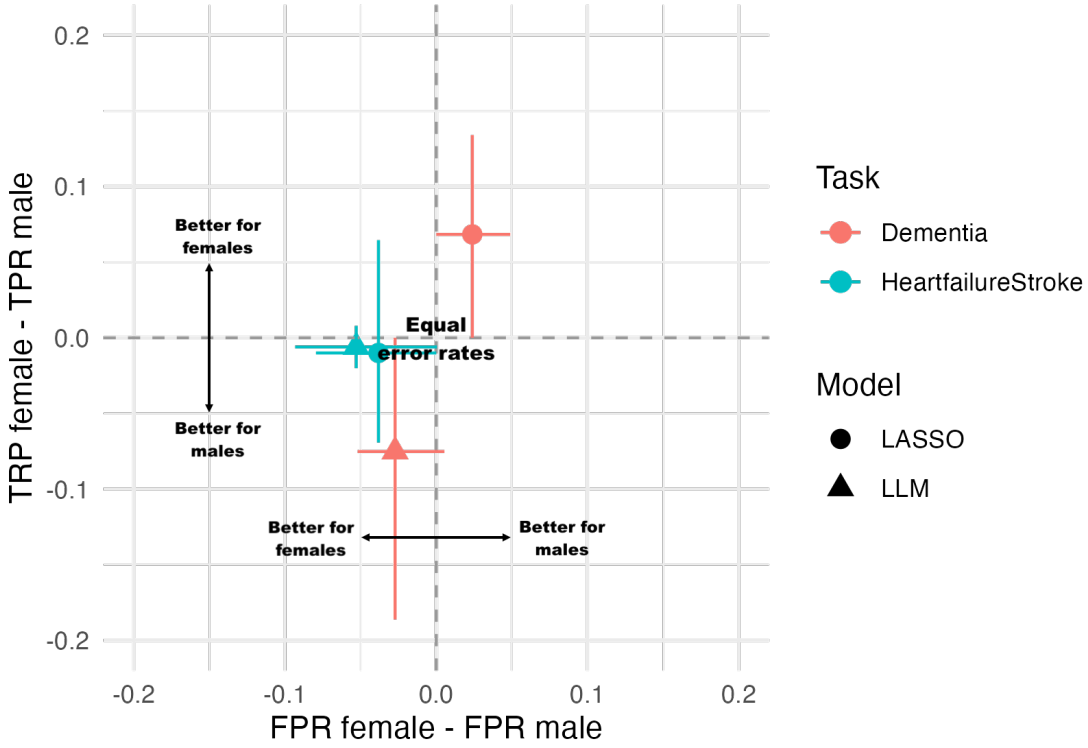

Figure 4*. Group fairness between males and females. Points represent the average difference value between the false positive rates (FPR) on the horizonal axis, and the true positive rates (TPR) on the vertical axis. The bars show the minimum/maximum values across thresholds.*

Conclusion

We integrated a standardized LLM-based prognostic prediction approach into the existing PLP pipeline. While the LLM demonstrated reasonable predictive performance, its results were consistently lower than those of the LASSO model for both prediction tasks, aligning with recent findings [7]. Furthermore, our experiments show bias evaluation is crucial, as error rates may vary by model and prediction task.

Although this exploratory study is limited in scope, model types, and methodological choices, it provides a framework for systematic LLM-based prognostic prediction and evaluation. Our next steps will focus on assessing additional LLMs, experimenting with prompting strategies, and further investigating the explanation capabilities of both approaches.

References

1. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nature medicine 2022;28(1):31-38.
2. Yang C, Kors JA, Ioannou S, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. Journal of the American Medical Informatics Association 2022;29(5):983-89.
3. Dijkland SA, Helmrich IRR, Steyerberg EW. Validation of prognostic models: challenges and opportunities. Journal of Emergency and Critical Care Medicine 2018;2.
4. Hassan M, Kushniruk A, Borycki E. Barriers to and facilitators of artificial intelligence adoption in health care: scoping review. JMIR Human Factors 2024;11:e48633.
5. Glicksberg BS, Timsina P, Patel D, et al. Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room. Journal of the American Medical Informatics Association 2024;31(9):1921-28.
6. Acharya A, Shrestha S, Chen A, et al. Clinical risk prediction using language models: benefits and considerations. Journal of the American Medical Informatics Association 2024;31(9):1856-64.
7. Brown KE, Yan C, Li Z, et al. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. Journal of the American Medical Informatics Association 2025;32(5):811-22.
8. Chen C, Yu J, Chen S, et al. ClinicalBench: Can LLMs Beat Traditional ML Models in Clinical Prediction? arXiv preprint arXiv:2411.06469 2024.
9. Omar M, Soffer S, Agbareia R, et al. Sociodemographic biases in medical decision making by large language models. Nature Medicine 2025:1-9.
10. de Ridder MA, de Wilde M, de Ben C, et al. Data resource profile: the integrated primary care information (IPCI) database, The Netherlands. International Journal of Epidemiology 2022;51(6):e314-e23.
11. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. Journal of the American Medical Informatics Association 2018;25(8):969-75.
12. Yang A, Li A, Yang B, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388 2025.

Supplementary material

Table S1. *Prediction task definitions*

Name	**Dementia task**	**HeartFailureStroke task**
Target definition	First visits of patients between 55 and 85 years old in the year 2019. With no prior dementia, subtypes of dementia, traumatic brain injury, memory or cognitive impairment, or anti-dementia drug prescribed.	First diagnosis of type 2 diabetes mellitus in patients older than 18 years, with no prior type 1 diabetes mellitus, heart failure, or stroke.
Outcome definition	A condition occurrence of dementia for the first time in the patient's history.	A condition occurrence of either heart failure or stroke for the first time in the patient's history.
Time at risk	5 years / 1825 days	5 years / 1825 days
Observation period	1 year / 365 days	1 year / 365 days
Washout period	365 days	365 days
Minimum time at risk	1825 days	1825 days

Table S2. *Names of the 47 phenotypes used as covariates.*

Phenotype names	
Acetaminophen prescribed	Hypertension
Acute gastrointestinal bleeding	Hypothyroidism
Alcoholism	Inflammatory bowel disease
Anemia	Low back pain
Angina	Major depressive disorder
Antibiotics	Neuropathy
Antiepileptics prescribed	Obesity
Anxiety	Opioids prescribed
Any cancer	Osteoarthritis
Aspirin	Osteoporosis
Asthma	Peripheral vascular disease
Atrial fibrillation	Pneumonia
Chronic hepatitis	Psychotic disorder
Chronic kidney disease or end stage renal disease	Rheumatoid arthritis
Chronic obstructive pulmonary disease	Seizure
Coronary artery disease	Skin ulcer
Deep vein thrombosis	Sleep apnea
Dyspnea	Smoking
Edema	Steroids prescribed
Gastroesophageal reflux disease	Stroke or transient ischaemic attack
Heart failure	Type 1 diabetes
Heart valve disorder	Type 2 diabetes
Hormonal contraceptives prescribed	Urinary tract infections
Hyperlipidemia	

Table S3. *System and user prompts and expected assistant response for LLM cohort summary creation and prognostic prediction.*

Message type	**Message content**
Cohort summarization	
System	Given an OMOP CDM cohort definition in JSON format, provide a short and concise summary describing this cohort.
User	[The cohort JSON generated by ATLAS]
Assistant	*[Generated cohort summary]*
Modular instructions for prognostic prediction	

System	# DEFINITIONS
	## TARGET COHORT
	[*Generated target cohort summary*]
	## CLINICAL OUTCOME
	[*Generated outcome cohort summary*]
	## TIME-AT-RISK
	[*Time-at-risk*]
	
	# INSTRUCTIONS
	Given the clinical data of a patient who meets the criteria for the TARGET COHORT, estimate the risk probability that this patient will develop the specified CLINICAL OUTCOME within the defined TIME-AT-RISK.
	
	- The clinical data are the patient DEMOGRAPHICS and recorded CLINICAL EVENTS (diagnoses and drugs prescribed) in the last [*observation window*].
	- Your response must contain:
	  - The estimated risk probability between 0.00 and 1.00, rounded to two decimal places.
	  - A short sentence to motivate your decision, no longer than 150 characters.
	
	Output format:
	{
	  "reasoning": "[short one sentence reasoning.]",
	  "probability": [0\.\d{2}]
	}

User	# PATIENT DATA ## DEMOGRAPHICS [*List of demographic data, example:*] age in years = 46 gender = MALE ## CLINICAL EVENTS [*List of present covariates, example:*] Asthma Steroids prescribed Dyspnea
Assistant	*{* * Reasoning: [free text]* * Result: [probability between 0 and 1.]* *}*