Transformer-based models for cardiovascular disease risk prediction: what is the potential?

Lisa Pennells,[1,2,3] Loïc Lannelongue,[1,2,4] and Stephen Kaptoge[1,2] *

[1]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Papworth Road, Cambridge Biomedical Campus, Cambridge CB2 0BB, UK; [2]Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Papworth Road, Cambridge Biomedical Campus, Cambridge CB2 0BB, UK; [3]British Heart Foundation Centre of Research Excellence, University of Cambridge, Papworth Road, Cambridge Biomedical Campus, Cambridge CB2 0BB, UK; and [4]Cambridge Sustainable Computing Lab, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom.

* Corresponding author. Tel: +44 (0)1223 748668, Email: skk22@medschl.cam.ac.uk

Graphical Abstract

<To follow>

Introduction

Artificial intelligence (AI) has potential to improve cardiovascular disease (CVD) prevention in several ways, including improved prediction of disease trajectories in individual patients based on the ever-growing complex mass of data from multiple sources including Electronic Health Records (EHR), imaging, omics and wearables. Transformer-based methods, originally developed for natural language processing, have been heralded as game changes for this purpose.[1]

Published in this issue of the European Heart Journal, authors Cao et al describe the development and validation of China-AIHeart; sex-specific Transformer-based models for predicting 10-year CVD risk among Chinese adults.[2] The developed China-AIHeart models are found to perform favourably in comparison with similar Cox models derived in the same data, and three guideline-recommended CVD risk scores China-PAR,[3] SCORE2 Asia-Pacific[4] and PREVENT.[5] This kind of comparison is a welcome addition to the evidence base on such methods, but what does the current work really tell us about the potential for Transformer-based modelling? How do these methods work? Has their potential been realised here? Do these methods really provide improved risk prediction over the long-standing Framingham-style models?

Advances in artificial intelligence

Transformers have been a key methodological development underpinning recent advances in generative artificial intelligence.[1,6] Although they share fundamental components with other deep learning architectures, namely that they are neural networks operating on learned data representations, they differ in several important respects. At a high level,

Transformers are a type of AI model that can read and generate text (and increasingly other data like images or signals) by learning patterns from very large datasets. Earlier approaches needed to consider the role of each part of the input fixed (e.g. "age" in column 1) or to process information sequentially (e.g. earlier language models). Transformers, on the other hand, can look at all parts of an input at once and "pay attention" to the most relevant pieces, dynamically deciding which parts of the input matter to each other for the task at hands.[7] This ability lets them capture complex relationships in language, medical notes, or even biological data, and is what underpins modern tools like large language models (e.g., ChatGPT).

The principal distinction between Transformers and traditional machine learning or earlier deep learning approaches lies in their ability to model complex, context-dependent interactions across all input features simultaneously.[1] This flexibility has driven their success across a range of domains, particularly language and image processing. However, this modelling capacity comes at a cost. Transformers typically require a lot of data, large numbers of parameters, and substantial computational resources to train effectively. Early implementations contained on the order of tens of millions of parameters, whereas contemporary large-scale models may contain tens of billions or more.

This study in context

In the analysis presented by Cao et al, the Transformer-based China-AIHeart models were derived and tested using cohort data on 150,000 individuals without prior CVD from multiple sites across China.[2] The individuals were followed up for a median of 10-years, during which time around 12,000 CVD events were observed. The authors prioritised 22 variables for inclusion as risk predictors based on prior knowledge, all captured at the cohort baseline assessment with no repeat measurements.

Were Transformer models required?

A key question here is whether Transformer models were best suited for the job at hand. The low-dimensional tabular data in this study represents a less natural application domain for Transformers, since there is no sequential or spatial structure that attention mechanisms are designed to exploit. For such tabular data, established methods such as gradient-boosted tree ensemble methods (e.g. XGBoost) remain competitive or superior, particularly when data are limited, owing to their efficiency and inductive biases.[8] Going straight to Transformer-based modelling without considering intermediate solutions to flexibly capture relationships could be considered as using a sledgehammer to crack a nut. Likewise, Cox models remain well suited to this kind low-dimensional data, where it is

feasible to investigate any deviations from the simple structure using traditional statistical approaches. For example, significant non-linear associations could be parsimoniously detected using fractional polynomials or splines and evidence for important interactions principally tested via likelihood ratio or Wald tests. It reassuring that the authors performed parallel comparisons with traditional Cox models, but is unclear whether steps were taken to include non-linear associations and interactions.

Comparing results of Transformer and traditional models

China-AIHeart appeared to perform better than the traditional Cox models in terms of discrimination and other measures of predictive ability, both on internal and external validation. So why is this the case? Can improvement be attributed to the Transformer methodology, or it could other epidemiological factors be relevant?

Diabetes prevalence in the derivation cohort was extremely high at 25.7% in men and 22.1% in women, and even higher in the Xinjiang validation cohort (37.7% and 29.6% respectively). The SHAP values indicate important roles for glucose and kidney function (included as fasting glucose, 2-hour post-load glucose, eGFR and UACR) in China AI-Heart. These variables are most relevant for individuals with diabetes,[5,9] and their inclusion is likely a key driver of improved discrimination. By comparison, no such variables are included in China PAR[3] or SCORE2 Asia-pacific[4] (rather, SCORE2-Diabetes,[9] was developed specifically for individuals with diabetes and includes eGFR, HbA1c and age-of-diabetes-diagnosis) and the basic PREVENT[5] models include only eGFR (extended versions include HbA1c and UACR). Interactions of these variables with diabetes status and with age are commonly important, as are non-linear associations (e.g. for eGFR), and are included in models like PREVENT[5] and SCORE2-Diabetes.[9] However, it is unclear whether these aspects were incorporated in the Cox models derived by the current authors; these models may not represent the state-of-the-art traditional methods that are needed for appropriate comparison.

The SHAP values also suggest an important predictive role for socioeconomic status. Values specific to the Chinese population likely capture substantial regional variation in CVD risk across China.[10] This is not reflected by PREVENT or SCORE2 Asia-pacific which were not developed specifically for China and is also not explicitly considered in China-PAR, which only acknowledges geographical region (north vs south) and urban vs rural environment (for men only).[3] Reporting the hazard ratio for socioeconomic status in the Cox model would have helped quantify its contribution.

Overall, evidence that Transformer methods themselves drive the improved prediction is limited. The inclusion of additional diabetes-related and socioeconomic variables may largely explain the observed gains. Furthermore, the role of overfitting cannot be excluded

given the inherent assumption of complexity of the Transformer models and the similarity between derivation and validation datasets.

Strengths and limitations

A key strength of the current study is analysis of cohort data among people without previous CVD with parallel comparison of Transformer-based versus traditional risk models and adaptations to survival analysis context.

However, there are some notable limitations which may further limit the conclusions that can be drawn. In terms of data quality and handling, around 15% of individuals were excluded from both derivation and validation datasets due to lack of CVD follow-up; it's unclear if they could have been included up to an administrative censoring date as in traditional survival analysis to avoid bias. Furthermore, there were significant missing data in the CHARLS validation dataset, and outcomes here were from self-report only.

The clinical relevance of differences in model performance is also unclear since this is limited to statistical significance of differences in risk prediction measures (C-index, continuous NRI, IDI, F1 and net benefit) rather than translation to meaningful quantities in terms of events identified using relevant risk thresholds. Estimates of population risk distributions and risk classifications were missing yet essential to understand potential clinical implications of the results. The relevance of F1 and SHAP measures are also unclear, as is whether their classification-based calculation accounted for the survival analysis data context.

Relation to previous research and further considerations

Ultimately, this comparison in cohort data is interesting exploration of the use of Transformers with tabular data. This is an active area of research which may offer advantages in settings involving large-scale datasets or complex feature interactions.[1,6] However, it's hard to see that the current application really reflects the potential benefits of the novel approach. Notable previously published uses of Transformers for disease prediction include the Delphi-2M model,[1] a GPT-based model of multi-disease progression which uses longitudinal electronic health records (EHR) data as input. Likewise, the TRISK models use a similar approach,[11] but specifically targeting CVD risk prediction in the primary care context. In both cases Transformers offer real potential in dealing with EHR to uncover individual trajectories based on messy longitudinal data from multiple sources (demographics, symptoms, test results, diagnoses, medication etc). However, future primary prevention-based research should ensure focus on the clinically relevant target population. Much information in EHR is likely to come from prior disease diagnoses and

medication, and for primary prevention populations such information is sparse.[12] Comparative analyses should carefully exclude individuals with prior CVD or related high-risk conditions which preclude them from guideline recommended risk estimation with models. Without this, it remains challenging to unpick whether improvements in performance are really due to advancements in methodology or inclusion of risk predictors such as prior disease and medications. It is perhaps easier to envisage the value Transformers in handling the more data-rich individuals with pre-existing disease to predict future trajectories and to detect important patterns in monitoring and imaging data.

References

1. Shmatko A, Jung AW, Gaurav K, Brunak S, Mortensen LH, Birney E, et al. Learning the natural history of human disease with generative transformers. *Nature* 2025; **647**(8088): 248-56.
2. Cao Q, Xu X, Lin H, Yin Y, Wu S, Li M, et al. Development and Validation of Transformer-Based Models for Predicting Cardiovascular Disease Risk among Chinese Adults. *Eur Heart J* 2026; **EPub**(xxx).
3. Yang X, Li J, Hu D, Chen J, Li Y, Huang J, et al. Predicting the 10-Year Risks of Atherosclerotic Cardiovascular Disease in Chinese Population: The China-PAR Project (Prediction for ASCVD Risk in China). *Circulation* 2016; **134**(19): 1430-40.
4. Hageman SHJ, Huang Z, Lee H, Kaptoge S, Dorresteijn JAN, Pennells L, et al. Risk prediction of cardiovascular disease in the Asia-Pacific region: the SCORE2 Asia-Pacific model. *Eur Heart J* 2025; **46**(8): 702-15.
5. Khan SS, Matsushita K, Sang Y, Ballew SH, Grams ME, Surapaneni A, et al. Development and Validation of the American Heart Association's PREVENT Equations. *Circulation* 2024; **149**(6): 430-49.
6. Langlais É L, Thériault-Lauzier P, Marquis-Gravel G, Kulbay M, So DY, Tanguay JF, et al. Novel Artificial Intelligence Applications in Cardiology: Current Landscape, Limitations, and the Road to Real-World Applications. *J Cardiovasc Transl Res* 2023; **16**(3): 513-25.
7. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention Is All You Need. *arXiv* 2023; **arXiv:1706.03762 [cs.CL]**.
8. Borisov V, Leemann T, Seßler K, Haug J, Pawelczyk M, Kasneci G. Deep Neural Networks and Tabular Data: A Survey. *arXiv* 2022; **arXiv:2110.01889 [cs.LG]**.
9. SCORE2-Diabetes Working Group and the ESC Cardiovascular Risk Collaboration, Pennells L, Kaptoge S, Østergaard HB, Read SH, Carinci F. SCORE2-Diabetes: 10-year cardiovascular risk estimation in type 2 diabetes in Europe. *Eur Heart J* 2023; **44**(28): 2544-56.
10. Zheng C, Wang X, Gu R, Wang W, Tian Y, Pei X, et al. The Burden of Cardiovascular Disease in China: Incidence Estimates From National Surveillance (2023). *J Am Coll Cardiol* 2025.

11. Rao S, Li Y, Mamouei M, Salimi-Khorshidi G, Wamil M, Nazarzadeh M, et al. Refined selection of individuals for preventive cardiovascular disease treatment with a transformer-based risk model. *Lancet Digit Health* 2025; **7**(6): 100873.

12. Mamouei M, Fisher T, Rao S, Li Y, Salimi-Khorshidi G, Rahimi K. A comparative study of model-centric and data-centric approaches in the development of cardiovascular disease risk prediction models in the UK Biobank. *Eur Heart J Digit Health* 2023; **4**(4): 337-46.