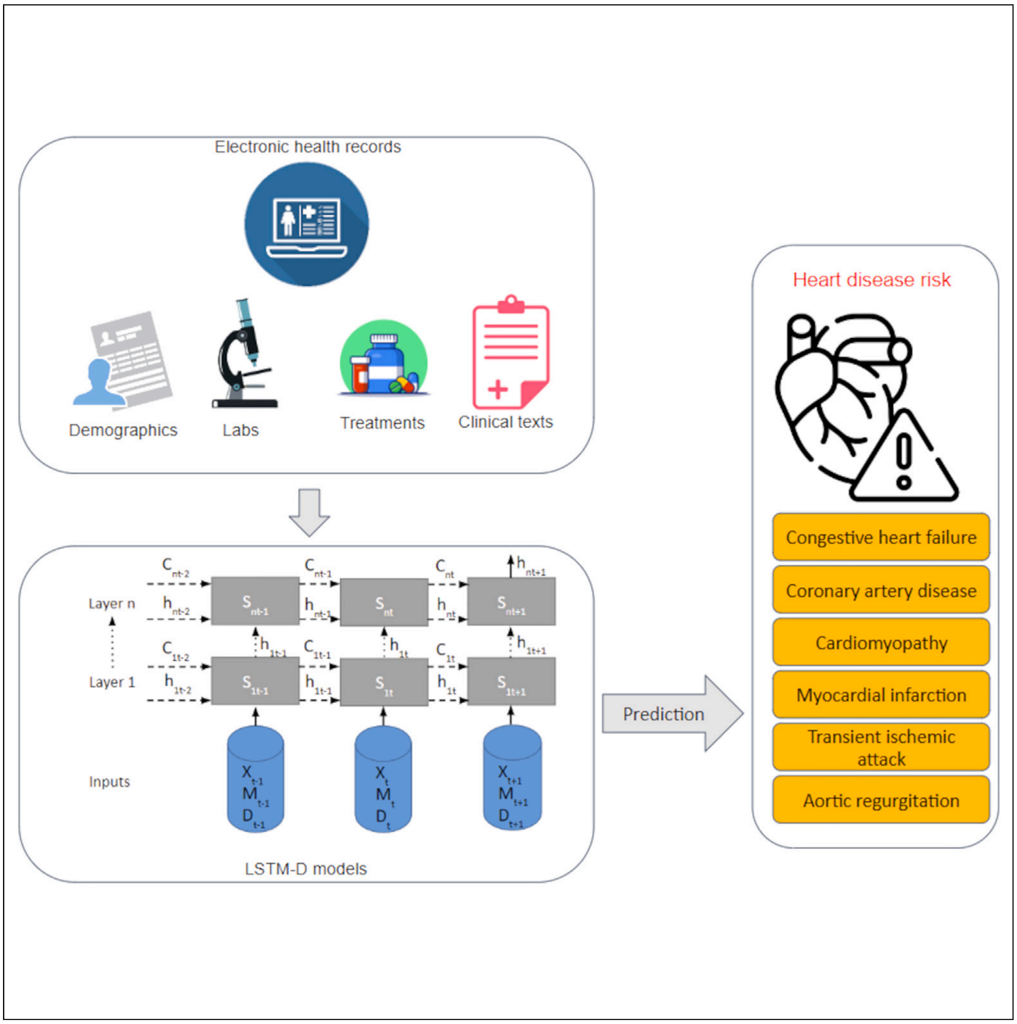
iScience

CellPress
OPEN ACCESS

Article

Risk prediction of heart diseases in patients with breast cancer: A deep learning approach with longitudinal electronic health records data



Sicheng Zhou, Anne Blaes, Chetan Shenoy, Ju Sun, Rui Zhang

zhan1386@umn.edu

Highlights

Developed LSTM-D models to predict six heart diseases in patients with breast cancer

Enhanced heart disease prediction using NLP-extracted breast cancer phenotypes

Identified optimal observation windows for heart disease prediction

Zhou et al., iScience *27*, 110329
July 19, 2024 © 2024 Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.110329

Article

Risk prediction of heart diseases in patients with breast cancer: A deep learning approach with longitudinal electronic health records data

Sicheng Zhou,[1] Anne Blaes,[2] Chetan Shenoy,[3] Ju Sun,[4] and Rui Zhang[5,6,*]

SUMMARY

Accurately predicting heart disease risks in patients with breast cancer is crucial for clinical decision support and patient safety. This study developed and evaluated predictive models for six heart diseases using real-world electronic health records (EHRs) data. We incorporated a trainable decay mechanism to handle missing values in the long short-term memory (LSTM) model, creating LSTM-D models to predict heart disease risk based on longitudinal EHRs data. Additionally, we deployed NLP methods to extract breast cancer phenotypes from clinical texts, integrating unstructured and structured data to enhance predictions. Our LSTM-D models outperformed baseline models in predicting congestive heart failure, coronary artery disease, cardiomyopathy, myocardial infarction, transient ischemic attack, and aortic regurgitation, with AUC scores ranging from 0.7189 to 0.9548. Observation windows of 12–24 months were found optimal for model performance. This research advances precise, personalized care strategies, enabling early intervention and improved management of cardiovascular risks in breast cancer survivors.

INTRODUCTION

The incidence of female breast cancer has been steadily increasing since the mid-2000s, accounting for approximately 32% of all new cancer diagnoses among women.[1,2] Cardiotoxicity and cardiovascular diseases are significant problems associated with breast cancer treatments. It is one of the leading causes of death for patients with breast cancer, with rates ranging from 7.4% to 13.3%, according to different studies.[3,4] Cardiotoxicity can present acutely or in the long term,[5,6] and breast cancer survivors may have a higher cardiovascular disease risk.[7] Cardiovascular disease risks vary with different cancer treatments, which act through diverse mechanisms. For instance, the anthracyclines and human epidermal growth factor receptor 2 (HER-2) targeted therapy may cause cardiomyocyte damage and heart failure, while radiation therapy may cause coronary and valvular diseases.[3] The incidence of trastuzumab-related cardiotoxicity could range from 2% to 7% for trastuzumab monotherapy, 2%–13% for trastuzumab combined with paclitaxel, and can reach 27% when combined with anthracyclines.[8] Given these risks, early identification of cardiotoxicity and cardiovascular diseases is paramount for proactive patient management and care. This is especially crucial since cardiotoxicity not only causes cardiovascular diseases but also delays breast cancer treatment, potentially leading to further harm to patients. However, we currently do an inadequate job of identifying patients with cancer who are at high risk for cardiotoxicity and cardiovascular disease. Identifying high-risk patients is critical to testing and identifying successful interventions to prevent adverse cardiovascular outcomes and applying them in clinical practice. There have been several clinical trials of cardiovascular medications such as angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, beta-blockers, and statins in high-risk patients receiving anthracyclines and/or anti-HER2 therapies, and most of them have been unimpressive or negative.[9–12] The main reason for the lack of benefit has been the low rates of adverse outcomes identified in the supposed high-risk patients, highlighting our current inability to reliably identify patients with cancer who are at high risk for cardiotoxicity and cardiovascular disease.[13] Thus, there is an urgent need for accurate methods to identify patients with cancer at high risk for a broad array of cardiovascular diseases.

Electronic Health Records (EHRs) have revolutionized healthcare, offering a wealth of patient data that can be harnessed to predict health outcomes and guide clinical decision-making. Traditional machine learning models have been widely used in healthcare for heart disease prediction.[14–16] They have focused on aggregated features, such as event counts and averages. However, these models often fall short when dealing with complex, high-dimensional, and temporal data typically found in EHRs.[17] In contrast, time-series models, such as recurrent neural networks (RNN) based models, provide significant advantages in handling such data. Unlike traditional models that treat data points

[1] Institute for Health Informatics, University of Minnesota, Minneapolis, MN, USA
[2] Division of Hematology, Oncology and Transplantation, University of Minnesota, Minneapolis, MN, USA
[3] Cardiovascular Division, Department of Medicine, University of Minnesota Medical School, Minneapolis, MN, USA
[4] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA
[5] Division of Computational Health Sciences, Department of Surgery, University of Minnesota, Minneapolis, MN, USA
[6] Lead contact
*Correspondence: zhan1386@umn.edu

https://doi.org/10.1016/j.isci.2024.110329

independently, temporal models are designed to recognize and utilize the temporal dependencies and patterns in time-series data, making them particularly suited for EHRs data analysis.

Previous studies have developed predictive models for heart diseases or other clinical events using EHRs data, with the approach shifting from traditional machine learning models to deep learning models. Ezaz et al. conducted a study that sought to develop a proportional hazards model to identify older women with breast cancer who are at higher risk of HF or CM after trastuzumab-based therapy.[18] Candidate predictors of HF and CM were identified, including age, adjuvant chemotherapy, coronary artery disease, atrial fibrillation or flutter, diabetes mellitus, hypertension, and renal failure. The model could classify HF/CM risk into low, medium, and high-risk groups. Yang et al. developed a gradient-boosting model to predict heart failure in patients with cancer. A series of structured EHRs features, including demographics, diagnoses codes, medications, and procedures, were used, and the model achieved an area under the Receiver Operating Characteristic curve (AUC) score of 0.9077.[14] Liu et al. leveraged the unstructured clinical texts of patients to build a convolutional neural networks (CNN) model that predicts and classifies patients with a high risk of HF and achieved an F-1 score of 0.756.[19] The study shows the feasibility of using unstructured EHR data for heart disease prediction. Several studies have explored RNN-based models to take advantage of the temporal patterns hidden in the longitudinal EHRs data. Choi et al. investigated the gated recurrent unit (GRU) model to predict heart failure in patients using the structured longitudinal medication, diagnosis, and procedure codes information data.[17] The results indicate that the GRU model significantly outperformed the traditional machine learning models, including the support vector machine (SVM), linear regression (LR) and multilayer perceptron (MLP) models. Chu et al. proposed a hybrid deep learning approach that combined the RNN model with the generative adversarial network to predict heart failure for general patients and obtained an improved AUC score compared to the traditional GRU model.[20] One major issue of the longitudinal EHRs data is the missing values, which could cause the underperformance of the model, loss of power, biased estimates, and so forth.[21–23] Various approaches have been explored to mitigate its influence on developing predictive models using EHRs data. Perez-Lebel compared the performance of data imputation methods that use mean, median, or k-nearest neighbor to build the predictive models and found these imputation methods could improve the performance to different extents, and using a mask to indicate which values have been imputed is essential.[24] Che et al. proposed a trainable decay mechanism that integrates well with GRU-based models, i.e., GRU-D, and achieved remarkable performance in predicting in-hospital mortality and diagnosis of admission for patients. The trainable mechanism was applied to GRU models' inputs and hidden states to capture the missing information and the input observation patterns.[25] Ruan et al. developed similar GRU-D models for the post-surgical complications prediction task. They found that the performance of GRU-D models alternates between outperforming and being on par with the logit model, depending on the specific complication.[26] The efficacy of the trainable decay mechanism is evident in short-term clinical event prediction (typically within days), however, its effectiveness in long-term scenarios (spanning up to years) remains to be elucidated. Given these developments and the challenges inherent in predicting complex clinical outcomes, existing methods have been inadequate for complex and temporal data found in EHRs and cannot effectively identify patients with breast cancer who are at high risk for cardiovascular diseases. And the current evidence-based clinical guidelines for preventing and controlling cardiovascular disease risks for patients with breast cancer are not specific enough. There is a critical need for innovative time-series models capable of harnessing complex, high-dimensional temporal data from EHRs to improve cardiovascular risk prediction and patient management in breast cancer survivors.

In this study, we aim to develop deep learning methods to predict six types of heart diseases among patients with breast cancer using EHRs data. Our approach underscores the potential of advanced time-series models in transforming cardiovascular risk prediction in breast cancer survivors, paving the way for more personalized and effective patient care strategies. The contributions of this study include.

(1) We innovatively developed the LSTM-D that integrated the trainable decay mechanism into LSTM models to predict six types of heart diseases in patients with breast cancer using the longitudinal EHRs data. Heart diseases include congestive heart failure (CHF), coronary artery disease (CAD), cardiomyopathy (CM), myocardial infarction (MI), transient ischemic attack (TIA), and aortic regurgitation (AR). Our LSTM-D models have demonstrated superior performance compared to all baseline models, marking a significant leap in predictive accuracy and model innovation.
(2) We demonstrated substantial enhancement in heart disease prediction attributed to the inclusion of NLP features. We applied NLP methods to extract breast cancer phenotypes from clinical narratives and pathology reports embedded within EHRs. The results underscore the vital role of NLP features in enhancing the accuracy and effectiveness of predictive models in clinical settings.
(3) We explored the influence of different observation windows on the performance of predictive models. Optimized observation windows were identified for various heart diseases. The analysis provides guidance for the application of predictive models in real clinical settings.

RESULTS

Data collections for patients with breast cancer

We identified 3421 patients with breast cancer that met the inclusion criteria. And 3375 patients have unstructured pathology reports or clinical notes that can extract the unstructured features. Major demographics, clinical variables, and heart disease outcomes of patients with breast cancer are shown in Table 1. Also, the NLP extracted cancer related variables were summarized in the supplemental information (Table S1).

Evaluation of the predictive models

We evaluated the all the developed models using the 5-fold cross-validation strategy. The models include baseline models and our LSTM-D models. The AUC scores are shown in Figure 1. We compared the performance of the models with and without the NLP features. In most

Table 1. Summary of major demographics, clinical variables, and heart disease outcomes of patients with breast cancer

Variables	Total patients (*n* = 3375)
Age (years)	57.57
Race	
White	3022 (89.5%)
Asian	87 (2.58%)
Black or African American	88 (2.61%)
American Indian or Alaska Native	17 (0.50%)
Native Hawaiian or Other Pacific Islander	2 (0.06%)
Chemotherapy (binary)	727 (21.5%)
Targeted therapy (binary)	627 (18.6%)
Radiation therapy (binary)	2066 (61.2%)
Heart disease outcomes	
CAD (binary)	162 (4.80%)
CHF (binary)	193 (5.72%)
CM (binary)	165 (4.89%)
MI (binary)	48 (1.42%)
TIA (binary)	48 (1.42%)
AR (binary)	489 (14.5%)

cases, the NLP features improved the performance of the models. Overall, the LSTM-D models with NLP features obtained the best performance for all six diseases. The AUC scores for CHF, CAD, CM, MI, AR, and TIA are 0.9334, 0.9439, 0.9548, 0.8989, 0.7694, and 0.7189, respectively. The detailed AUC scores and corresponding standard deviations can be found in supplemental information (Tables S2 and S3).

We further evaluated the influence of the decaying mechanism on the LSTM model. Figure 2 shows the AUC scores of the LSTM-D and regular LSTM models (without decay mechanism). The LSTM-D models performed better than the LSTM models for predicting all types of heart diseases, indicating the effectiveness of the decaying mechanism in handling long-term EHRs data. The AUC scores of regular LSTM models can be found in supplemental information (Table S4).

Figure 3 shows the AUC scores of the LSTM-D models for different observation windows. Overall, the models could achieve better performance along with the increasing observation windows, especially for the AR and TIA. For CHF, CAD, CM and MI, the performance of the models remains stable when the observation window is longer than 12 months. The AUC scores of different observation windows can be found in supplemental information (Table S5).

Figure 4 shows the top 20 features of the predictive models for the six heart diseases. The complete ranking of the importance of the feature can be found in supplemental information (Table S6).

DISCUSSIONS

In this study, we developed LSTM-D models that leverage time-series data from EHRs to predict heart diseases in patients with breast cancer. The results in Figure 1 demonstrate that the LSTM-D models outperformed all the baseline models across all types of heart diseases. This superior performance can be attributed to the LSTM-D models' ability to effectively capture and utilize the temporal dependencies inherent in EHRs time-series data. For the patients with breast cancer with heart diseases in our study, the average time between the cancer diagnosis and heart disease diagnosis is 989 days. Traditional atemporal models using the aggregated features cannot effectively handle the longitudinal patterns and time-dependent relationships in the data. Inspired by the GRU-D models,[25] we integrated the trainable decay mechanism into three parts of the LSTM models, i.e., inputs, hidden states, and cell states to address the challenge of missing values in time series data. Figure 2 shows the efficacy of the trainable decay mechanism in our heart disease prediction tasks by comparing the LSTM-D models with classic LSTM models. The decay mechanism provides an effective way to integrate incomplete data, a common issue in EHRs, thereby improving the robustness and reliability of the model in real-world clinical settings. Traditional models often require extensive pre-processing to handle such data inconsistencies, which can lead to loss of information and potential biases.

A significant enhancement in our models' predictive capability was observed by integrating unstructured features extracted using NLP methods from clinical texts. Clinical texts are often underutilized in the development of clinical predictive modeling. Our results indicate the critical role that unstructured clinical texts can play in improving the performance of disease prediction models. The inclusion of NLP-based features suggests that valuable patient information residing in unstructured EHRs data can be effectively utilized for improved predictive outcomes. Furthermore, we explored the impact of different observation windows on model performance. This analysis is crucial for clinical applications as it informs the decision-making process regarding the amount of historical patient data that should be considered for

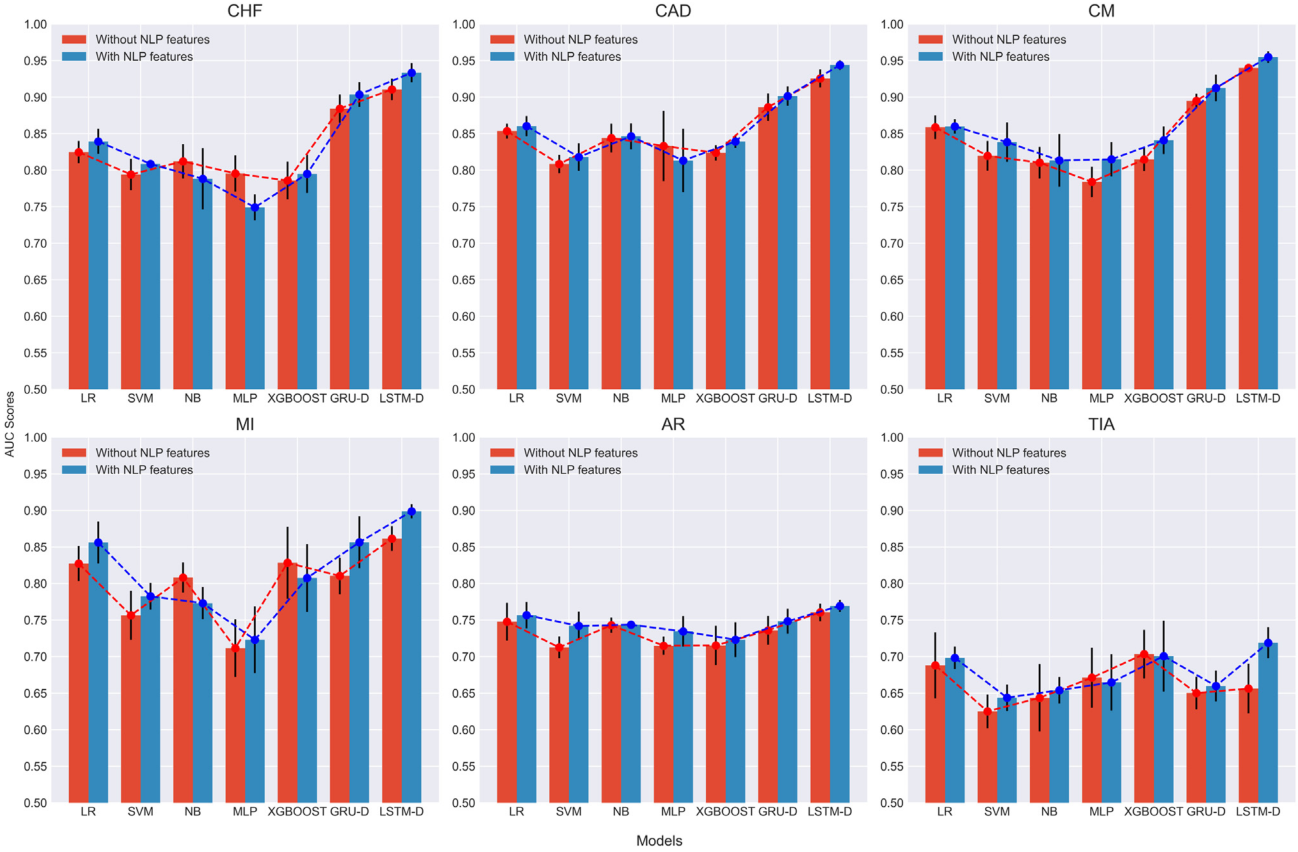


Figure 1. AUC scores of baseline models and LSTM-D models for predicting different heart diseases

The black error bars indicate the standard deviations.

accurate disease prediction. As shown in Figure 3, for CHF, CM, and CAD, there is no significant improvement in performance after 12 months of the observation window. For MI, the optimal observation window is > 24 months, while for AR and TIA, a full timeline of patients' data is needed to obtain the best performance. From a clinical point of view, the model performance is related to the timing of the cardiac events. The data close to the cardiac events may have more predictive power, this might explain the performance at the 1-month observation window already achieved AUC >0.9 for CHD, CAD, and CM. Overall, a longer observation window is beneficial for model performance. However, there are tradeoffs, longer observation windows indicate more efforts for data collection and preprocessing. Also, the model performance won't keep increasing along with the increase of observation windows. Thus, for the model's implementation in clinical practice, the choice of the optimal observation windows may need further validation, and we should keep monitoring the performance, as the optimal observation windows may change over time for different disease outcomes.

We delved into the feature importance of the LSTM-D models in predicting the six types of heart diseases. Figure 4 shows the relative performance change of each feature and ranks the importance of features from high to low. These top features are highly consistent with the risk factors identified from the epidemiological work,[27–29] such as BMI, SBP, target therapy, and radiation therapy. The NLP features such as cancer laterality, cancer stage, and cancer grade were also found to play essential roles during the risk predictions. This analysis not only offers an understanding of the model's decision-making process but also highlights key clinical factors that are most indicative of each heart disease. Such insights are invaluable for clinicians, as they provide a deeper understanding of disease mechanisms and can guide targeted interventions.

In a comparative analysis, the overall performance of our models surpasses other models in previous studies. For the prediction of CAD, our model achieved an AUC of 0.944, which outperforms the previous benchmark of 0.930 reported in a meta-analysis study.[30] In a similar study that developed machine learning methods to predict a series of cardiovascular diseases,[31] they achieved an AUC of 0.821 for CAD, o.882 for CHF, and 0.807 for MI, while our models achieved an AUC of 0.944, 0.933, and 0.899 for the three diseases respectively, which indicates significant improvement. We have internally evaluated the feasibility of integrating the models into the clinical workflow. The implementation of the models has two stages, in stage 1 we only include the structured data, and in stage 2 we will implement the complete models by including the NLP features. There are several steps before it is fully implemented such as setting up a data pipeline that can automatically collect, store, and pre-process the EHR data for patients, and prospective evaluation and safety monitoring of the models.

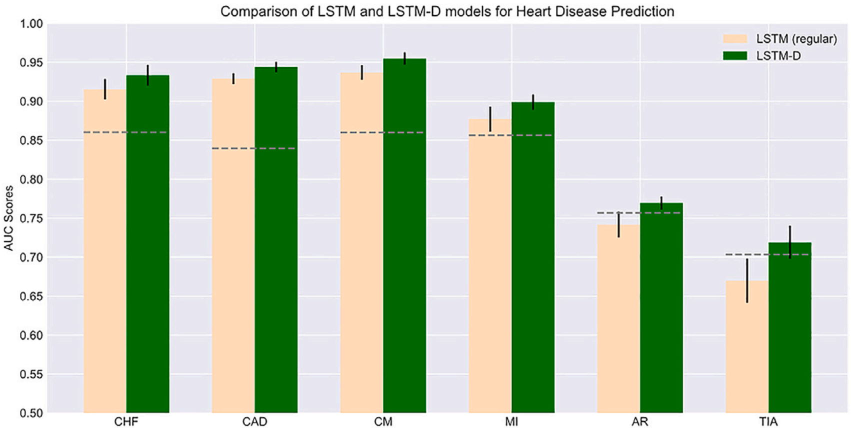


Figure 2. AUC scores of LSTM and LSTM-D models for different heart diseases

The black error bars indicate the standard deviations. The gray dashed lines in bar clusters indicate the best performance of baseline models.

Limitations of the study

This study presents several limitations that suggest avenues for future research. Firstly, the dataset used to train our models was restricted in size, particularly for MI and TIA. We anticipate that enlarging the dataset will enhance model performance. Additionally, due to limitations in data availability, the inclusion of all potential risk factors, such as dosage and frequency of various cancer therapies, was not feasible. Beyond the six cardiovascular diseases analyzed, other conditions, such as ischemic stroke, warrant further investigation. Furthermore, our data were exclusively sourced from the M Health Fairview, the models' generalizability across different clinical environments needs to be further validated. Our research will proceed in three main directions in the future. We intend to focus on verifying the generalizability of our models by testing them in varied clinical contexts and enhancing their explainability in collaboration with clinicians. Concurrently, we plan to refine the models by incorporating additional pertinent risk factors and broadening their scope to encompass a more extensive array of cardiovascular diseases. Additionally, we will assess the fairness of our AI systems, examining whether race and other social determinants of health impact model performance. These efforts will help ensure that our predictive tools are both effective and equitable across different patient demographics.

In summary, we developed LSTM-D models that effectively predict six types of heart diseases in patients with breast cancer, with AUC scores ranging from 0.7189 to 0.9548. Our analysis highlights the importance of incorporating the NLP features and collecting data in an adequate time range to build predictive disease risk models based on longitudinal EHRs data. The implications of this study offer a pathway toward more personalized, accurate, and timely predictions of heart diseases in patients with breast cancer, ultimately contributing to better patient outcomes and healthcare delivery.

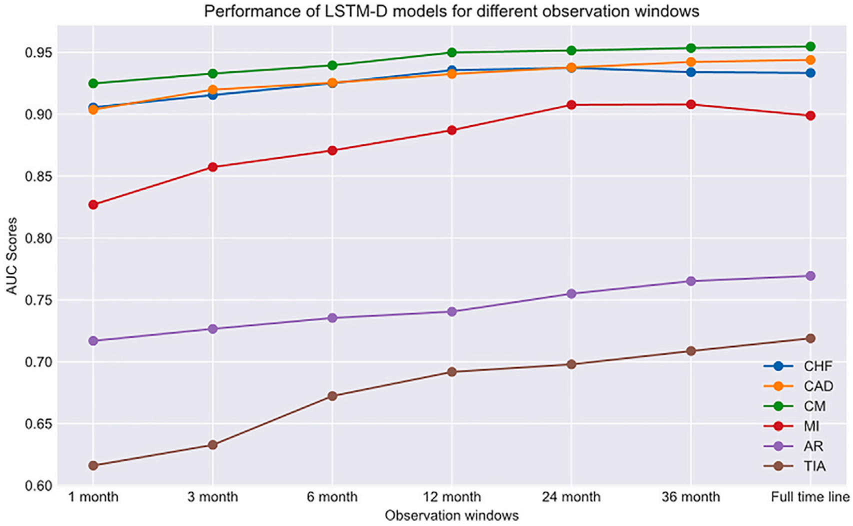


Figure 3. AUC scores of LSTM-D models for different observation windows, ranging from one month to the full timeline

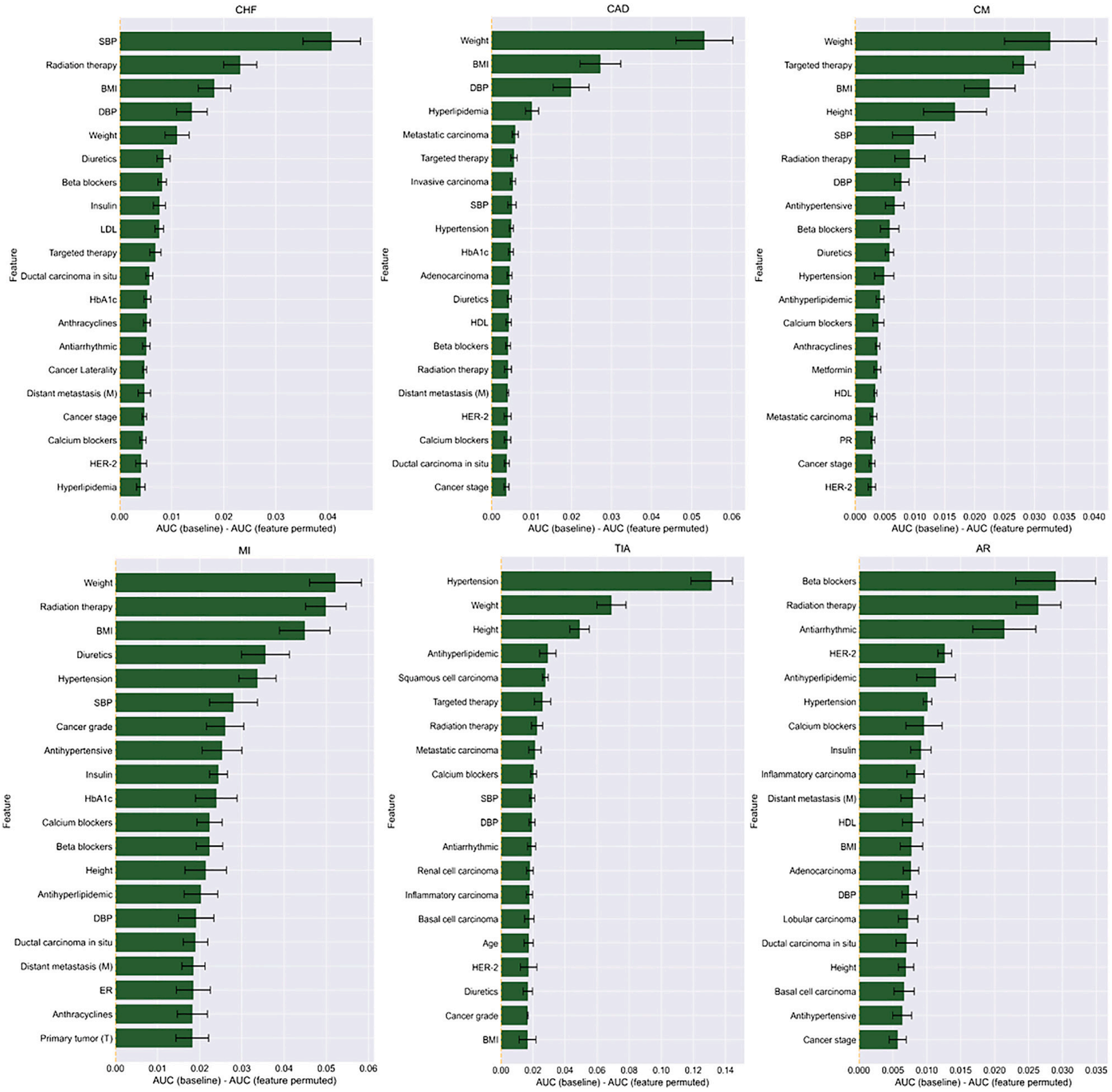


Figure 4. Top 20 predictors of LSTM-D models for different heart diseases

The X axis indicates the difference in AUC scores between original and permuted features. The black error bars indicate the standard deviations.

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
  - Human participants

- METHOD DETAILS
    - Structured clinical risk factors relevant to heart diseases
    - Cancer phenotypes extracted by NLP
    - LSTM-D model for heart disease prediction
    - Baseline models
    - Data preparation
    - Model evaluations
    - Observation windows
    - Permutation feature importance

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110329.

ACKNOWLEDGMENTS

This work is partially supported by the National Institute on Minority Health and Health Disparities under grant number 1R21MD019134 (Zhang, Wang, Hou); the National Cancer Institute under grant number 1R01CA287413 (Sun, Cui & Zhang); and the University of Minnesota Clinical and Translational Science Institute under grant number UL1TR002494. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We thank support from the University of Minnesota (UMN) Center for Learning Health System Sciences.

AUTHOR CONTRIBUTIONS

SZ conducted the main experiments. RZ guided the study design, data collection, and analysis. All authors participated in the writing of the article and critical revisions of the article for important intellectual content.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 29, 2024
Revised: May 12, 2024
Accepted: June 18, 2024
Published: June 20, 2024

REFERENCES

1. Giaquinto, A.N., Sung, H., Miller, K.D., Kramer, J.L., Newman, L.A., Minihan, A., Jemal, A., and Siegel, R.L. (2022). Breast cancer statistics, 2022. CA. Cancer J. Clin. *72*, 524–541.
2. Siegel, R.L., Giaquinto, A.N., and Jemal, A. (2024). Cancer statistics, 2024. CA. Cancer J. Clin. *74*, 12–49.
3. Cardinale, D., Colombo, A., Bacchiani, G., Tedeschi, I., Meroni, C.A., Veglia, F., Civelli, M., Lamantia, G., Colombo, N., Curigliano, G., et al. (2015). Early detection of anthracycline cardiotoxicity and improvement with heart failure therapy. Circulation *131*, 1981–1988.
4. Bria, E., Cuppone, F., Fornier, M., Nisticò, C., Carlini, P., Milella, M., Sperduti, I., Terzoli, E., Cognetti, F., and Giannarelli, D. (2008). Cardiotoxicity and incidence of brain metastases after adjuvant trastuzumab for early breast cancer: the dark side of the moon? A meta-analysis of the randomized trials. Breast Cancer Res. Treat. *109*, 231–239.
5. Broder, H., Gottlieb, R.A., and Lepor, N.E. (2008). Chemotherapy and cardiotoxicity. Rev. Cardiovasc. Med. *9*, 75–83.
6. Cai, F., Luis, M.A.F., Lin, X., Wang, M., Cai, L., Cen, C., and Biskup, E. (2019). Anthracycline-induced cardiotoxicity in the chemotherapy treatment of breast cancer: Preventive strategies and treatment. Mol. Clin. Oncol. *11*, 15–23.
7. Bradshaw, P.T., Stevens, J., Khankari, N., Teitelbaum, S.L., Neugut, A.I., and Gammon, M.D. (2016). Cardiovascular disease mortality among breast cancer survivors. Epidemiol. Camb. Mass *27*, 6–13.
8. Chavez-MacGregor, M., Zhang, N., Buchholz, T.A., Zhang, Y., Niu, J., Elting, L., Smith, B.D., Hortobagyi, G.N., and Giordano, S.H. (2013). Trastuzumab-related cardiotoxicity among older patients with breast cancer. J. Clin. Oncol. *31*, 4222–4228.
9. Bisceglia, I., Mistrulli, R., Cartoni, D., Matera, S., Petrolati, S., and Canale, M.L. (2023). Cardiac toxicity of chemotherapy for breast cancer: do angiotensin-converting enzyme inhibitors and beta blockers protect? Eur. Heart J. Suppl. *25*, B25–B27.
10. Kimmick, G., Dent, S., and Klem, I. (2019). Risk of cardiomyopathy in breast cancer: how can we attenuate the risk of heart failure from anthracyclines and anti-HER2 therapies? Curr. Treat. Options Cardiovasc. Med. *21*, 1–17.
11. Mauro, C., Capone, V., Cocchia, R., Cademartiri, F., Riccardi, F., Arcopinto, M., Alshahid, M., Anwar, K., Carafa, M., Carbone, A., et al. (2023). Cardiovascular Side Effects of Anthracyclines and HER2 Inhibitors among Patients with Breast Cancer: A Multidisciplinary Stepwise Approach for Prevention, Early Detection, and Treatment. J. Clin. Med. *12*, 2121.
12. Dempsey, N., Rosenthal, A., Dabas, N., Kropotova, Y., Lippman, M., and Bishopric, N.H. (2021). Trastuzumab-induced cardiotoxicity: a review of clinical risk factors, pharmacologic prevention, and cardiotoxicity of other HER2-directed therapies. Breast Cancer Res. Treat. *188*, 21–36.
13. Virizuela, J.A., García, A.M., de Las Peñas, R., Santaballa, A., Andrés, R., Beato, C., De La Cruz, S., Gavilá, J., González-Santiago, S., and Fernández, T.L. (2019). SEOM clinical guidelines on cardiovascular toxicity (2018). Clin. Transl. Oncol. *21*, 94–105.
14. Yang, X., Gong, Y., Waheed, N., March, K., Bian, J., Hogan, W.R., and Wu, Y. (2019). Identifying Cancer Patients at Risk for Heart Failure Using Machine Learning Methods (American Medical Informatics Association), p. 933.
15. Chang, W.-T., Liu, C.-F., Feng, Y.-H., Liao, C.-T., Wang, J.-J., Chen, Z.-C., Lee, H.-C., and Shih, J.-Y. (2022). An artificial intelligence approach for predicting cardiotoxicity in breast cancer patients receiving anthracycline. Arch. Toxicol. *96*, 2731–2737.

16. Du, Z., Yang, Y., Zheng, J., Li, Q., Lin, D., Li, Y., Fan, J., Cheng, W., Chen, X.-H., and Cai, Y. (2020). Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: model development and performance evaluation. JMIR Med. Inform. *8*, e17257.
17. Choi, E., Schuetz, A., Stewart, W.F., and Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. J. Am. Med. Inform. Assoc. *24*, 361–370.
18. Ezaz, G., Long, J.B., Gross, C.P., and Chen, J. (2014). Risk prediction model for heart failure and cardiomyopathy after adjuvant trastuzumab therapy for breast cancer. J. Am. Heart Assoc. *3*, e000472.
19. Liu, X., Chen, Y., Bae, J., Li, H., Johnston, J., and Sanger, T. (2019). Predicting heart failure readmission from clinical notes using deep learning (IEEE), pp. 2642–2648.
20. Chu, J., Dong, W., and Huang, Z. (2020). Endpoint prediction of heart failure using electronic health records. J. Biomed. Inform. *109*, 103518.
21. Li, J., Yan, X.S., Chaudhary, D., Avula, V., Mudiganti, S., Husby, H., Shahjouei, S., Afshar, A., Stewart, W.F., Yeasin, M., et al. (2021). Imputation of missing values for electronic health record laboratory data. NPJ Digit. Med. *4*, 147.
22. Stiglic, G., Kocbek, P., Fijacko, N., Sheikh, A., and Pajnkihar, M. (2019). Challenges associated with missing data in electronic health records: a case study of a risk prediction model for diabetes using data from Slovenian primary care. Health Inf. J. *25*, 951–959.
23. Ayilara, O.F., Zhang, L., Sajobi, T.T., Sawatzky, R., Bohm, E., and Lix, L.M. (2019). Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. Health Qual. Life Outcomes *17*, 106.
24. Perez-Lebel, A., Varoquaux, G., Le Morvan, M., Josse, J., and Poline, J.-B. (2022). Benchmarking missing-values approaches for predictive models on health databases. GigaScience *11*, giac013.
25. Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. Sci. Rep. *8*, 6085.
26. Ruan, X., Fu, S., Storlie, C.B., Mathis, K.L., Larson, D.W., and Liu, H. (2022). Real-time risk prediction of colorectal surgery-related post-surgical complications using GRU-D model. J. Biomed. Inform. *135*, 104202.
27. Johnson, C.B., Davis, M.K., Law, A., and Sulpher, J. (2016). Shared risk factors for cardiovascular disease and cancer: implications for preventive health and clinical care in oncology patients. Can. J. Cardiol. *32*, 900–907.
28. Koene, R.J., Prizment, A.E., Blaes, A., and Konety, S.H. (2016). Shared risk factors in cardiovascular disease and cancer. Circulation *133*, 1104–1114.
29. Meijers, W.C., and de Boer, R.A. (2019). Common risk factors for heart failure and cancer. Cardiovasc. Res. *115*, 844–853.
30. Krittanawong, C., Virk, H.U.H., Bangalore, S., Wang, Z., Johnson, K.W., Pinotti, R., Zhang, H., Kaplin, S., Narasimhan, B., Kitai, T., et al. (2020). Machine learning prediction in cardiovascular diseases: a meta-analysis. Sci. Rep. *10*, 16057.
31. Zhou, Y., Hou, Y., Hussain, M., Brown, S.A., Budd, T., Tang, W.H.W., Abraham, J., Xu, B., Shah, C., Moudgil, R., et al. (2020). Machine learning–based risk assessment for cancer therapy–related cardiac dysfunction in 4300 longitudinal oncology patients. J. Am. Heart Assoc. *9*, e019628.
32. Sun, D., Simon, G.J., Skube, S., Blaes, A.H., Melton, G.B., and Zhang, R. (2017). Causal Phenotyping for Susceptibility to Cardiotoxicity from Antineoplastic Breast Cancer Medications (American Medical Informatics Association), p. 1655.
33. Zhou, S., Wang, N., Wang, L., Liu, H., and Zhang, R. (2022). CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. J. Am. Med. Inform. Assoc. *29*, 1208–1216.
34. Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Comput. *9*, 1735–1780.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Python	https://www.python.org/	Version 3.8.12
Scikit-learn	https://scikit-learn.org/	Version 1.3.2
Pytorch	https://pytorch.org/	Version 1.8.1
XGBoost	https://xgboost.readthedocs.io/en/stable/#	Version 1.7.3
CancerBERT	Zhou et al.[28]	N/A
LSTM-D	https://github.com/zjsuper/heart_disease_predictions	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Rui Zhang (zhan1386@umn.edu).

Materials availability

This study did not generate any new materials.

Data and code availability

- The data is not publicly available due to HIPPA regulations.
- The codes are available on a public repository (https://github.com/zjsuper/heart_disease_predictions).
- 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

Human participants

This study was approved by the institutional review boards of the UMN. The dataset for this study contains the EHRs data of breast cancer patients extracted from the UMN Clinical Data Repository. The ICD-9 and ICD-10 codes were used to identify patients diagnosed with breast cancer between 2011-2020. All patients have breast cancer treatment records and have a minimum follow-up time of 1 year. The patients with prior cancer history, except for non-melanoma skin cancer or cervical cancer *in situ*, were excluded from the study. We also limited the patients to female adults, and all patients received at least one of the breast cancer treatments:1) Anthracyclines-based chemotherapy, i.e., doxorubicin, epirubicin, daunorubicin, idarubicin, valrubicin; 2) Targeted therapy: limit to trastuzumab as this is a well-known drug causing cardiotoxicity; 3) Radiation therapy: defined by ICD codes. Since the aim of the study is to predict heart diseases, the patients diagnosed with heart diseases before the breast cancer diagnosis were excluded. The summarized age and race of patients is shown in Table 1. The influence of demographics on the model developed in the study requires further investigation.

METHOD DETAILS

Structured clinical risk factors relevant to heart diseases

Comprehensive structured clinical risk factors relevant to heart diseases were collected from the EHRs to build the prediction models. We chose these risk factors based on the previous study.[32] We also included some new variables, such as cardiovascular medications, breast cancer treatments, and lab values. The list of variables and potential values is shown in Table S7.

Cancer phenotypes extracted by NLP

Besides, we obtained unstructured cancer phenotypes, i.e., NLP features, from clinical texts in EHRs using the developed CancerBERT model.[33] The cancer phenotype categories include laterality, cancer grade, cancer stages, and histological types. The clinical texts (including pathology reports and clinical notes) were split into sentences and fed into the CancerBERT model. All extracted features were normalized using rule-based methods. If different values of the features were found in clinical texts, we used majority voting to decide the final values. The comprehensive list of variables and potential values is shown in Table S7.

LSTM-D model for heart disease prediction

The proposed LSTM-D models for heart disease prediction are an extension of the RNN framework and were inspired by previous GRUD models.[25] We designed the multilayer LSTM-D models that integrated the trainable decay mechanisms with the LSTM model for heart disease prediction. Figure S1 shows the architecture of the multilayer LSTM-D model.

Figure S1A shows the high-level structure of the model. It has the multilayer LSTM as the main structure, and a softmax prediction layer was added on top of the LSTM model to predict the Y, i.e., the outcome of heart diseases. Figure S1C shows the format of inputs for the model (include two variables of timeline data as an example). The model inputs contain three parts: $\boldsymbol{X}$ is the time series feature matrix transformed from the EHRs data. $X \in \mathbb{R}^{T \times N}$ where $\boldsymbol{T}$ is the time steps of the time series data and $\boldsymbol{N}$ is the number of features. ***NA*** indicates the missing values in the time series data. $\boldsymbol{M} \in \mathbb{R}^{T \times N}$ is the masking matrix that indicates the positions of missing values in $\boldsymbol{X}$, where $\mathbf{m}_\mathbf{t}^\mathbf{n}$ is 1 if $\mathbf{x}_\mathbf{t}^\mathbf{n}$ is not NA, or 0 otherwise. $\boldsymbol{D} \in \mathbb{R}^{T \times N}$ is the matrix that measures the time interval (days) between each time point. Time points are the time (days) since the first time the time series data was collected for each patient. The calculation of $\boldsymbol{D}$ is adapted from **Che** et al.[25] and shown in Equation 1, where $\mathbf{d}_\mathbf{t}^\mathbf{n}$ is the time interval of feature **n** since its last observation, and $\boldsymbol{T_t}$ is the time of $\boldsymbol{t}$th time step.

$$d_t^n = \begin{cases} (T_t - T_{t-1}) \times 0.01 + d_{t-1}^n, t > 1, X_{t-1}^n = NA \\ (T_t - T_{t-1}) \times 0.01, t > 1, X_{t-1}^n \neq NA \\ 0, t = 1 \end{cases} \quad \text{(Equation 1)}$$

Figure S1B shows the structure of a cell of the model. The feature vector at time step $\boldsymbol{t}$ ($\boldsymbol{X_t}$), hidden state ($\boldsymbol{h_{t-1}}$) and cell state ($\boldsymbol{C_{t-1}}$) of time step $\boldsymbol{t-1}$ are transformed to the decayed $\boldsymbol{X_{dt}}$, $\boldsymbol{h_{dt-1}}$ and $\boldsymbol{C_{dt-1}}$ respectively before calculating the LSTM cell. The calculation of $\boldsymbol{X_{dt}}$ is shown in Equation 2.[25]

$$x_{dt}^n = m_t^n x_t^n + \left(1 - m_t^n\right)\left(\alpha_{x_t}^n x_{t'}^n + \left(1 - \alpha_{x_t}^n\right)\overline{x}^n\right) \quad \text{(Equation 2)}$$

$\boldsymbol{x_{dt}^n}$ and $\boldsymbol{x_t^n}$ are the decayed and original features $\boldsymbol{n}$ at time point t. $\mathbf{m}_\mathbf{t}^\mathbf{n}$ is the missing indicator of feature **n** at time point **t**. $\boldsymbol{\overline{x}^n}$ is the mean value of feature n of training data. $\boldsymbol{\alpha_{x_t}^n}$ is the decay rate of feature n at time point **t**, which is determined by Equation 3.[25] $\boldsymbol{W_\alpha}$ and $\boldsymbol{b_\alpha}$ are the trainable decay weight matrix and bias, $\boldsymbol{d_t}$ is the time interval vector from **D**.

$$\alpha_t = e^{-\max\left(0, W_\alpha d_t + b_\alpha\right)} \quad \text{(Equation 3)}$$

The $\boldsymbol{h_{dt-1}}$ and $\boldsymbol{C_{dt-1}}$ are determined by the Equations 4 and 5, where $\boldsymbol{\alpha_{ht}}$ and $\boldsymbol{\alpha_{Ct}}$ are decay rate of the hidden states and cell states. They are also calculated by Equation 3 with their own $\boldsymbol{W_\alpha}$ and $\boldsymbol{b_\alpha}$ trainable decay weight parameters. The calculation of $\boldsymbol{C_t}$ and $\boldsymbol{h_t}$ are same as normal LSTM cell.[34]

$$h_{dt-1} = \alpha_{ht} \odot h_{t-1} \quad \text{(Equation 4)}$$

$$C_{dt-1} = \alpha_{Ct} \odot C_{t-1} \quad \text{(Equation 5)}$$

Baseline models

We also developed a series of baseline models for the heart disease prediction task to compare with our proposed LSTM-D models. The models include LR, SVM, NB, XGBoost and MLP. We also investigated the performance of GRU-D models that achieved state-of-the-art performances in previous clinical event prediction tasks.[25,26] The GRU-based models were trained and evaluated on the same time series data as the LSTM-based models.

Data preparation

For the GRU and LSTM-based time series models, we collected the multivariate time series data from patients. All variables in Table S7 were collected since the breast cancer diagnosis date for patients. The outcomes are the six types of heart diseases, i.e., CHF, CAD, CM, MI, TIA and AR. For other baseline models, i.e., LR, SVM, NB, XGBoost, and MLP, that are not designed for time series data, we organized the extracted variables into a tabular format that can be handled by these models. The index date was defined as the first date of any breast cancer treatment (chemotherapy, radiation, or targeted therapy). Prescriptions of cardiovascular medications and heart diseases (outcomes) were extracted from the follow-up period (after the index date). For all other variables, the longitudinal observations before index data were summarized as the value before and closest to the index date. Missing values for continuous variables were imputed using either average or normal values. For triglyceride, BMI, DBP and SBP, the mean values were used for imputation. For HDL, LDL and Hba1c, the missing values were set to 55 mg/dL, 115 mg/dL, and 6%, respectively.

Model evaluations

We independently trained all the models for different heart diseases and framed the prediction into a binary classification task. 5-fold cross-validation was used for model training and evaluation. We took four chunks in each fold as the training and validation sets (in a 4:1 ratio) and one chunk as the test set. We trained the models for 20 epochs. After each epoch, we assessed the model performance using the AUC score on the validation set. After all the training epochs, we chose the model that obtained the highest AUC on the validation set and further evaluated its AUC on the held-out test set. This process was replicated across all five folds, and the mean test AUC and the corresponding standard deviation were calculated as the final model performance metric.

Observation windows

We evaluated the LSTM-D models with different observation windows (i.e., the period to collect patients' time series data before diagnosing any heart disease) to explore the effect of observation window lengths on model performance. We set the observation windows to 3, 6, 12, 18, 24, 36 months and extracted the features within different observation windows. We trained the models separately for each observation window and evaluated the models to obtain the AUC scores.

Permutation feature importance

We conducted the permutation feature importance test for the LSTM-D models to determine the importance of individual features of a model. We initially trained the models and obtained the baseline AUC scores. We then randomly shuffled one feature's values across the samples in the test set. This shuffling breaks the relationship between that feature and the dependent variable, effectively rendering the feature as noise. The models were re-evaluated and we obtained the AUC scores after the feature permutations. This process was iteratively applied to each feature, and the change in the AUC scores, calculated as the difference between the pre-permutation AUC score and the post-permutation AUC score, indicates the feature's importance. The decreased AUC score suggests the feature is important for the model's predictions; otherwise, it implies the feature is not critical.