PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Predicting incident heart failure from population-based nationwide electronic health records: protocol for a model development and validation study
AUTHORS	Nakao, Yoko; Nadarajah, Ramesh; Shuweihdi, Farag; Nakao, Kazuhiro; Fuat, Ahmet; Moore, Jim; Bates, Christopher; Wu, Jianhua; Gale, Chris

VERSION 1 – REVIEW

REVIEWER	Nunes, Rafael Amorim Belo Hospital Alemao Oswaldo Cruz
REVIEW RETURNED	23-Apr-2023

GENERAL COMMENTS	In this study, the authors aim to develop and validate a model for predicting early heart failure at a population level, using two cohorts based on electronic health records data. Predicting models are essential for helping diagnosis and decision-making processes, so the question to be answered by the study appears pertinent. The study design is also well described. However, some points should be reviewed: In the Abstract, the authors justify the study relevance in this way: ``HF could reduce downstream healthcare impact, but predicting incident HF is challenging and statistical models are limited by performance and scalability in routine clinical practice. A HF prediction model developed in nationwide electronic health records (EHRs) could provide a scalable solution´´. Could you improve this statement? In the Abstract, the authors stated: ``Both comprise a large, representative population of England linked at patient-level to secondary care and mortality data´´. Could you clarify better this phrase? In page 7 , the 2nd paragraph seems to me not very clear. Are current guidelines for screening and diagnosing heart failure based only on natriuretic peptides? Could you recheck the current recommendations? In page 7, 3rd paragraph, the authors described the following: ``Previous models applicable to community-based EHRs to predict HF risk have been limited. Models have seldom been externally validated,18 19 which prohibits an understanding of their generalizability. Many have been developed in curated prospective cohort studies, and their performance may not translate to EHR data´´. It would be interesting to discuss deeply how each of these

	models was developed and how they differ from the model proposed by the authors. In the Methods, page 13, 4th paragraph, I would like to understand better and in an easier way, how the use of machine learning techniques could improve the predicting model. Is it will be used with logistic regression or will be performed in a separate way? In the Conclusion section, it could be put, from a practical point of view, how the development of the prediction model could be translated into clinical practice (creating an algorithm available to general practitioners?)

REVIEWER	Mpanya, Dineo University of the Witwatersrand Johannesburg, Internal Medicine
REVIEW RETURNED	24-Apr-2023

GENERAL COMMENTS	Nakao et al. prepared a protocol for a study on “Predicting incident heart failure from population-based nationwide electronic health records” using machine learning algorithms and a traditional statistical method, logistic regression. The authors should consider themselves fortunate to have access and mine such a robust database. This is an excellent project. 1. Page 7, line 56, add “cardiovascular magnetic resonance (CMR) imaging” 2. Page 7, line 54, write laboratory in full. 3. Page 13, line 14, Regarding data cleaning, please elaborate on how categorical and continuous variables will be cleaned. For example, how will you identify outliers? 4. Data cleaning: How will the authors handle missing values? 5. Page 13, line 27: mention that p-values <0.05 will represent statistical significance. 6. Page 13, line 33: Remove “a” 7. Page 14, line 3: consider adding accuracy and the F/F1 score as performance metrics. How will the authors evaluate the performance of the traditional logistic regression model. 8. Page 15, line 24. Please justify (not on the protocol) why you will be using BOTH R and Stata for the data analysis? 9. Change STATA to Stata 10. What proportion of the dataset will be used for training and testing the algorithms? 11. How will the authors ensure that the patients included in the training and testing set are not in the validation set? 12. Consider adding a list of features (predictors) that will be used in the study as an appendix. 13. Please elaborate on why the authors will conduct a systematic review to identify predictors of heart failure. There are many well-described feature selection methods. Why not allow algorithms to select features that best describe the dataset autonomously? 14. With regards to the definition of heart failure, did all patients have baseline echocardiography performed before diagnosing heart failure? 15. How will the 1, 5 and 10-year risk of heart failure be estimated? Please define the start and end-period of each prediction window. 16. I like the choice of classification algorithms selected. Will authors also consider machine learning logistic regression? 17. Will the authors be using grid search to tune the model?

	18. Will medications be included in the analysis? For example, beta-blockers, ACE inhibitors etc.? If so, how will the authors quantify the dosage? Will they use the total daily dose?

REVIEWER	Li, Deng-ao Taiyuan University of Technology
REVIEW RETURNED	01-May-2023

GENERAL COMMENTS	The aim of the study is to develop a model that is widely applicable and scalable in routinely-collected community-based EHRs, test its performance across a range of prediction horizons, and externally validate it in a geographically distinct primary care EHR dataset. The existing problems of this paper are as follows: 1. The content of the article mainly focuses on the description of the dataset itself. And is not closely related to the title of the article. The explanation of the implementation approach of the model is too abstract and vague, and It is not clear how the work will be carried out in the design of the model in the following steps. 2. The author has set forecast windows of 1 year, 5 years, and 10 years. Will the models currently planned to be used be outdated after 10 years? What are your thoughts on this? 3. The article mentions that to ensure the broad applicability of the model, the types of candidate variables will be limited. Will the reduction in the number of feature variables reduce the accuracy of the model's predictions? And is there a pre-planned selection criteria for choosing these variables?

VERSION 1 – AUTHOR RESPONSE

Reviewer Reports:

Reviewer: 1
Dr. Rafael Amorim Belo Nunes, Hospital Alemao Oswaldo Cruz

Comments to the Author:
In this study, the authors aim to develop and validate a model for predicting early heart failure at a population level, using two cohorts based on electronic health records data. Predicting models are essential for helping diagnosis and decision-making processes, so the question to be answered by the study appears pertinent. The study design is also well described. However, some points should be reviewed:
Author reply
On behalf of the authors, I would like to thank the Reviewer for their careful review of the manuscript and extremely helpful suggestions.

In the Abstract, the authors justify the study relevance in this way: ``HF could reduce downstream healthcare impact, but predicting incident HF is challenging and statistical models are limited by performance and scalability in routine clinical practice. A HF prediction model developed in nationwide electronic health records (EHRs) could provide a scalable solution´´. Could you improve this statement?
Author reply
Thank you, we have updated the abstract as follows.
Manuscript change

Predicting incident HF is challenging and statistical models are limited by performance and scalability in routine clinical practice. A HF prediction model implementable in nationwide electronic health records (EHRs) could enable targeted diagnostics to enable earlier identification of HF.

In the Abstract, the authors stated: ``Both comprise a large, representative population of England linked at patient-level to secondary care and mortality data´´. Could you clarify better this phrase?

Author reply

Thank you, we have updated the abstract as follows.

Manuscript change

Both comprise large cohorts of patients, representative of the population of England in terms of age, sex, and ethnicity. Primary care records are linked at patient-level to secondary care and mortality data.

In page 7, the 2nd paragraph seems to me not very clear. Are current guidelines for screening and diagnosing heart failure based only on natriuretic peptides? Could you recheck the current recommendations?

Author reply

We agree that the previous wording may have led to confusion. Thus we have changed it.

Manuscript change

International guidelines define four stages of HF: Stage A HF (at-risk for HF), Stage B HF (pre-HF; structural heart disease without symptoms), Stage C HF (symptomatic HF) and Stage D HF (advanced HF). Mortality increases with progression through the stages. Accordingly, guidelines recommend initiatives to identify individuals with Stage A and B HF as evidence supports that the onset of symptomatic HF can be delayed or prevented by targeting modifiable risk factors.

In page 7, 3rd paragraph, the authors described the following: ``Previous models applicable to community-based EHRs to predict HF risk have been limited. Models have seldom been externally validated,18 19 which prohibits an understanding of their generalizability. Many have been developed in curated prospective cohort studies, and their performance may not translate to EHR data´´. It would be interesting to discuss deeply how each of these models was developed and how they differ from the model proposed by the authors.

Author reply

As the rest of the manuscript details, we aim to overcome the shortfalls of previous models by:

- Deriving a model in electronic health records so that it has a route to implementation.
- our model is derived from real-world EHR data, ensuring its ability to handle the intricacies and heterogeneity of routinely-collected care records.
- Externally validating the model to demonstrate generalisability
- Use only age, sex, and recorded diagnoses to make risk prediction so that the model could be implemented at scale in routinely-collected care records
- Test prediction performance over both short (1-year) and long-term (10-year) prediction horizons so the model could be used both for targeting diagnostic screening and primary prevention interventions

In the Methods, page 13, 4th paragraph, I would like to understand better and in an easier way, how the use of machine learning techniques could improve the predicting model. Is it will be used with logistic regression or will be performed in a separate way?

Author reply

We will use candidate variables to develop a model with logistic regression but also use the same variables to develop models using supervised machine learning techniques (including neural network, random forest, support vector machine). We will compare the prediction performance of models developed from each method. Through this comparative analysis, we will gain insights into the

strengths and weaknesses of different modelling approaches and identify the most effective method for our specific prediction task.

In the Conclusion section, it could be put, from a practical point of view, how the development of the prediction model could be translated into clinical practice (creating an algorithm available to general practitioners?)

Author reply

The conclusion section has now been removed as it is not required in protocol papers. The aim is that a model developed using variables that are not routinely missing in primary care electronic health records could then be programmed into electronic health records by the providers and be available to general practitioners to use in day-to-day practice, or applied at scale to inform national screening or primary prevention initiatives.

Reviewer: 2
Dr. Dineo Mpanya, University of the Witwatersrand Johannesburg

Comments to the Author:
Nakao et al. prepared a protocol for a study on “Predicting incident heart failure from population-based nationwide electronic health records” using machine learning algorithms and a traditional statistical method, logistic regression. The authors should consider themselves fortunate to have access and mine such a robust database. This is an excellent project.

Author reply
On behalf of the authors, I would like to thank the Reviewer for their careful review of the manuscript and extremely helpful suggestions.

1. Page 7, line 56, add “cardiovascular magnetic resonance (CMR) imaging”

Author reply
Thank you. We have updated.

Manuscript change
cardiac magnetic resonance (CMR) imaging

2. Page 7, line 54, write laboratory in full.

Author reply
Thank you. We have updated.

Manuscript change
Laboratory results

3. Page 13, line 14, Regarding data cleaning, please elaborate on how categorical and continuous variables will be cleaned. For example, how will you identify outliers?

Author reply
For categorical variables, we will primarily focus on addressing common data quality issues such as missing values, inconsistent formatting, and encoding errors. We will carefully inspect the data to identify any missing values in categorical variables and determine the most appropriate method for handling them. For medical conditions, we will take positive response approach, ie, if a specific condition is missing or not recorded, we assume the patient did not have the condition. Additionally, we will ensure that the categories are properly defined, and any inconsistencies in their representation will be resolved to maintain data integrity.
Regarding continuous variables, our data cleaning process will involve several steps, including outlier detection. We will employ statistical techniques to identify potential outliers in the continuous variables. One commonly used method is the use of z-scores, where observations that fall outside a certain threshold (typically defined as a z-score of greater than 3 or less than -3) are flagged as potential outliers. We will visually inspect the distribution of the variables and consider other statistical measures, such as the interquartile range (IQR), to identify outliers that may deviate significantly from the majority of the data points.
Once outliers are identified, we will evaluate their impact on the analysis and the predictive modeling process. Depending on the specific circumstances, we may choose to address outliers by either excluding them from the dataset, transforming them through winsorization or log transformations, or employing robust statistical techniques that are less sensitive to outliers.

4. Data cleaning: How will the authors handle missing values?

Author reply
The candidate variables include age, sex, ethnicity and diagnoses. Ethnicity may be missing but we will include an ‘ethnicity unrecorded’ category where it is unavailable as missingness is considered informative. If diagnoses are absent, it is considered that that the patient does not have a diagnosis.

We are not including observations or laboratory results – which are frequently missing in routinely-collected electronic health records. Therefore, we do not expect missing data in the analytical cohort.

Manuscript change

For diagnoses if medical codes are absent in a patient record we will assume that the patient does not have that diagnosis, or that the diagnosis was not considered sufficient-ly important to have been recorded by the GP in case of symptoms.[35] Ethnicity information is routinely collected in the UK NHS and so has increasingly high completeness, and we will include an 'ethnicity unrecorded' category where it is unavailable because missingness is considered to be informative. Accordingly we do not expect any missing data for any of the predictor variables in the analytical cohort.

5. Page 13, line 27: mention that p-values <0.05 will represent statistical significance.

Author reply

Thank you, we have updated.

Manuscript change

We will perform descriptive analyses of all variables and test the statistical difference between cases and controls using the t-test for normally distributed continuous variables, Wilcoxon rank sum test for non-normally distributed a continuous variable (age), and Pearson's Chi-squared test for categorical variables, using a p-value ≤0.05 to represent significance.

6. Page 13, line 33: Remove "a"

Author reply

Thank you – removed.

7. Page 14, line 3: consider adding accuracy and the F/F1 score as performance metrics. How will the authors evaluate the performance of the traditional logistic regression model.

Author reply

To evaluate the performance of the traditional logistic regression model, we will calculate these additional metrics by comparing the predicted outcomes with the true outcomes. Accuracy will help us assess the overall correctness of the model's predictions, while the F1 score will provide insights into the model's ability to balance precision and recall, considering both false positives and false negatives.

By incorporating accuracy and the F/F1 score in our evaluation, we aim to provide a comprehensive assessment of the traditional logistic regression model's performance, alongside the other metrics mentioned in our manuscript. This will allow for a more thorough understanding of the model's strengths and limitations in predicting heart failure risk.

8. Page 15, line 24. Please justify (not on the protocol) why you will be using BOTH R and Stata for the data analysis?

Author reply

We appreciate the reviewer's suggestion and agree that using both R and Stata for the data analysis might create unnecessary redundancy. We understand the concern and would like to clarify our approach.

Upon careful consideration, we have decided to use R as the primary software package for our data analysis. R offers a comprehensive set of tools, libraries, and packages specifically designed for statistical analysis, machine learning, and data visualization. It provides a wide range of functionalities and flexibility, making it well-suited for our study's requirements.

9. Change STATA to Stata

Author reply

Thank you. We have updated.

Manuscript change

Stata

10. What proportion of the dataset will be used for training and testing the algorithms?

Author reply

Preprocessed patient-level data in CPRD-GOLD will be randomly split into an 80:20 ra-tio to create development and internal validation samples using the Mersenne twister pseudorandom number generator.

Manuscript change

Preprocessed patient-level data in CPRD-GOLD will be randomly split into an 80:20 ra-tio to create development and internal validation samples using the Mersenne twister pseudorandom number generator.

11. How will the authors ensure that the patients included in the training and testing set are not in the validation set?

Author reply

CPRD-AURUM and CPRD-GOLD practices are different geographically. There is possibility that patients may transfer from GOLD practices to AURUM practices or vice versa. But the proportion of transfer is limited. In the study, we will ensure that study period for a patient starts from registration with a practice and is censored from the date of transfer out. Therefore there is no overlapping period for the same patient in the training/testing set and the validation set.

12. Consider adding a list of features (predictors) that will be used in the study as an appendix.

Author reply

The list of candidate predictor variables is described in the Methods section: Predictor variables subheading: “The potential predictors will include: age, sex, ethnicity, and all disease conditions during follow-up. Candidate disease conditions will comprise hospitalised diseases, such as other cardiovascular diseases, obesity, diabetes mellitus, thyroid disorders, iron deficiency and anaemia, kidney dysfunction, electrolyte disorders, chronic lung disease, sleep-disordered breathing, hyperlipidaemia, gout, erectile dysfunction, depression, can-cer and infection.”

13. Please elaborate on why the authors will conduct a systematic review to identify predictors of heart failure. There are many well-described feature selection methods. Why not allow algorithms to select features that best describe the dataset autonomously?

Author reply

We are pre-specifying candidate predictor variables using a systematic review in line with recommendations in Prognosis Research in Healthcare : Concepts, Methods, and Impact, and Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research as this incorporates clinical rationale (which helps with later clinical acceptance of a model) as well as statistical techniques for variable inclusion in the prediction model.

14. With regards to the definition of heart failure, did all patients have baseline echocardiography performed before diagnosing heart failure?

Author reply

Heart failure cases refer to coded diagnoses of heart failure in routine care. In routine care individuals may have been given a HF diagnosis but not received an echocardiogram (PMID: 35354658; PMID: 19147462). The positive predictive value of a HF diagnosis in CPRD has previously been validated at 82%.

15. How will the 1, 5 and 10-year risk of heart failure be estimated? Please define the start and end-period of each prediction window.

Author reply

We investigate heart failure risks at 1, 5 and 10 years following the start date (latter of 2 January 1998 or registration date at general practice). We identify all heart failure incidence cases and create binary variables for each of the 1, 5, and 10 years whether incident heart failure or not.

16. I like the choice of classification algorithms selected. Will authors also consider machine learning logistic regression?

Author reply

In our study, we have indeed considered the use of machine learning techniques, including logistic regression, as one of the classification algorithms. Logistic regression is a widely used and well-established method for binary classification tasks, particularly in healthcare and medical research. The term "machine learning logistic regression" may refer to logistic regression implemented within a broader machine learning framework, incorporating techniques such as regularization, feature selection, or ensemble methods to enhance the model's performance.

17. Will the authors be using grid search to tune the model?

Author reply

Thank you for your question regarding the use of grid search for model tuning in our study.

We would like to clarify that in our specific study, we have determined that grid search is not necessary for model tuning. The decision was made based on careful consideration of the objectives, dataset characteristics, and classification algorithms employed in our analysis.
While grid search is a commonly used technique for hyperparameter optimization, it may not always be the most suitable approach for every study.

18. Will medications be included in the analysis? For example, beta-blockers, ACE inhibitors etc.? If so, how will the authors quantify the dosage? Will they use the total daily dose?

Author reply

Medications, such as antihypertensives, are not being included as candidate predictor variables to avoid confounding by indication.

Reviewer: 3
Dr. Deng-ao Li, Taiyuan University of Technology

Comments to the Author:
The aim of the study is to develop a model that is widely applicable and scalable in routinely-collected community-based EHRs, test its performance across a range of prediction horizons, and externally validate it in a geographically distinct primary care EHR dataset. The existing problems of this paper are as follows:
Author reply
On behalf of the authors, I would like to thank the Reviewer for their careful review of the manuscript and extremely helpful suggestions.

1. The content of the article mainly focuses on the description of the dataset itself. And is not closely related to the title of the article. The explanation of the implementation approach of the model is too abstract and vague, and It is not clear how the work will be carried out in the design of the model in the following steps.
Author reply
The format of the article follows previously published protocol papers for prediction model development and validation in BMJ Open (PMID: 34728455) – including sections on the data source, predictor variables, sample size calculations, data pre-processing, prediction model development, internal validation and external validation. The design of the model has been planned to make it easier for EHR system providers to program the model into their EHR systems, but the exact details of how EHR system providers will do this are outside of the scope of this research study, which seeks to develop and validate a prediction model for heart failure.

2. The author has set forecast windows of 1 year, 5 years, and 10 years. Will the models currently planned to be used be outdated after 10 years? What are your thoughts on this?
Author reply
We agree with the reviewer that the algorithm will need to be updated as population characteristics change, data quality of EHRs improves and new or additional risk factors emerge.
Manuscript change
The model will have to be updated as population characteristics change, data quality of EHRs improves and new or additional risk factors emerge.

3. The article mentions that to ensure the broad applicability of the model, the types of candidate variables will be limited. Will the reduction in the number of feature variables reduce the accuracy of the model's predictions? And is there a pre-planned selection criteria for choosing these variables?
Author reply
We are pre-specifying candidate predictor variables using a systematic review in line with recommendations in Prognosis Research in Healthcare: Concepts, Methods, and Impact, and Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research as this incorporates clinical rationale (which helps with later clinical acceptance of a model) as well as statistical techniques for variable inclusion in the prediction model. Variable types such as laboratory results or observations may improve prediction performance, but they limit implementation as they are often missing in routinely-collected electronic health records. Therefore we have made the decision to prioritise the ability to implement the prediction model at scale by limiting the candidate predictors to age, sex, ethnicity and diagnoses.

VERSION 2 – REVIEW

REVIEWER	Nunes, Rafael Amorim Belo Hospital Alemao Oswaldo Cruz
REVIEW RETURNED	25-Jun-2023

GENERAL COMMENTS	Considering the modifications implemented by the authors, the article is able to be published.

REVIEWER	Mpanya, Dineo University of the Witwatersrand Johannesburg, Internal Medicine
REVIEW RETURNED	22-Jun-2023

GENERAL COMMENTS	Well done to the authors. I have no further comments. Good luck.

REVIEWER	Li, Deng-ao Taiyuan University of Technology
REVIEW RETURNED	29-Jun-2023

GENERAL COMMENTS	The purpose of this study is to develop a widely applicable and scalable model for predicting heart failure events using routinely collected community electronic health records (EHRs). After carefully reviewing your manuscript, I believe you have made detailed revisions in data processing and experimental design based on the initial review comments. Based on these points, I consider the article to be mature enough for acceptance and to proceed with the next steps of the publication process.