BMJ Open

Artificial intelligence-based mining of electronic health record data to accelerate the digital transformation of the national cardiovascular ecosystem: design protocol of the CardioMining study

Athanasios Samaras,[1] Alexandra Bekiaridou,[1,2] Andreas S Papazoglou,[1] Dimitrios V Moysidis,[1] Grigorios Tsoumakas,[3] Panagiotis Bamidis,[4] Grigorios Tsigkas,[5] George Lazaros,[6] George Kassimis ,[1,7] Nikolaos Fragakis,[7] Vassilios Vassilikos,[8] Ioannis Zarifis,[9] Dimitrios N Tziakas,[10] Konstantinos Tsioufis,[6] Periklis Davlouros,[5] George Giannakoulas ,[1] CardioMining Study Group

To cite: Samaras A, Bekiaridou A, Papazoglou AS, *et al*. Artificial intelligence-based mining of electronic health record data to accelerate the digital transformation of the national cardiovascular ecosystem: design protocol of the CardioMining study. *BMJ Open* 2023;**13**:e068698. doi:10.1136/bmjopen-2022-068698

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-068698).

Received 28 September 2022
Accepted 12 March 2023

For numbered affiliations see end of article.

Correspondence to
Dr George Giannakoulas; ggiannakoulas@auth.gr

ABSTRACT

Introduction Mining of electronic health record (EHRs) data is increasingly being implemented all over the world but mainly focuses on structured data. The capabilities of artificial intelligence (AI) could reverse the underusage of unstructured EHR data and enhance the quality of medical research and clinical care. This study aims to develop an AI-based model to transform unstructured EHR data into an organised, interpretable dataset and form a national dataset of cardiac patients.

Methods and analysis CardioMining is a retrospective, multicentre study based on large, longitudinal data obtained from unstructured EHRs of the largest tertiary hospitals in Greece. Demographics, hospital administrative data, medical history, medications, laboratory examinations, imaging reports, therapeutic interventions, in-hospital management and postdischarge instructions will be collected, coupled with structured prognostic data from the National Institute of Health. The target number of included patients is 100 000. Natural language processing techniques will facilitate data mining from the unstructured EHRs. The accuracy of the automated model will be compared with the manual data extraction by study investigators. Machine learning tools will provide data analytics. CardioMining aims to cultivate the digital transformation of the national cardiovascular system and fill the gap in medical recording and big data analysis using validated AI techniques.

Ethics and dissemination This study will be conducted in keeping with the International Conference on Harmonisation Good Clinical Practice guidelines, the Declaration of Helsinki, the Data Protection Code of the European Data Protection Authority and the European General Data Protection Regulation. The Research Ethics Committee of the Aristotle University of Thessaloniki and Scientific and Ethics Council of the AHEPA University Hospital have approved this study. Study findings will be disseminated through peer-reviewed medical journals and international conferences. International collaborations with other cardiovascular registries will be attempted.

Trial registration number NCT05176769.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ Discharge letters and prognostic data for 100 000 cardiac patients will be collected.
- ⇒ Natural language processing techniques will facilitate automated clinical data mining from unstructured electronic health records.
- ⇒ The accuracy of the automated model will be compared with the manual data extraction by study investigators.
- ⇒ Machine learning tools will provide data analytics.
- ⇒ Generalising natural language processing models across languages still remains challenging.

INTRODUCTION

The combination of medicine with computer science and artificial intelligence (AI) techniques, including machine learning (ML) and natural language processing (NLP), is promising and is going to rapidly change the future of medicine in the upcoming years.[1–3] NLP aims to interpret human language and quantify aspects of medical practice that were previously amenable only to laborious and costly work.[4] ML focuses on the interpretation of data and has been used in different settings in medicine, including the automated interpretation of ECGs, image classification and risk stratification.[3 5] Mainly at a research level, these methods are already offering novel clinical practice approaches and could have an impact on a plethora of cardiac diseases, including heart failure and coronary artery disease.[6] Nevertheless, real-world clinical implementation remains a challenge and this lack of impact on everyday clinical practice stands in stark contrast to the enormous progress in research.

With the growing number of patients and their concentration in large tertiary centres, it becomes attractive to collect large amounts of clinical data systematically. Such registries are essential for exploring the characteristics of different comorbidities and understanding real-world cardiac patients. However, with the unprecedented amount of data, manual collection and traditional processing methods become a challenge, as it is time-consuming and costly for healthcare systems.[7] Other significant difficulties one faces are the unstructured free-form text of the electronic health records (EHRs) and the need for deidentification and safety of the vast amount of patient data. AI methods could fill this void in medical records, enabling the ability to analyse large amounts of information efficiently.[8]

The use of AI automated processes constitutes a novelty in big data configuration, offering a quick, reliable and fully deidentified data extraction for further processing.[9 10] The results from its efficient use can be easily extended to different healthcare systems, amplifying the produced knowledge and improving diagnostic and therapeutic accuracy, transforming the current clinical care practice.[11] Transferring AI methods from the laboratory to everyday clinical practice is a difficult task that necessitates a high level of specialisation, financial resources and cross-disciplinary collaboration among academia, industry and clinical institutions.[12]

This study aims to contribute to the development of a clinically useful and feasible AI model for accurate automated extraction and processing of large volumes of raw and unstructured clinical data from EHRs. The information acquired from automated procedures will form the largest national database of cardiac patients, derived from unstructured data. Ultimately, this study aspires to encourage the digital transformation of the national cardiovascular ecosystem, by improving clinical documentation towards an automated, rapid recording and utilisation of clinical data.

METHODS AND ANALYSIS

CardioMining study design

CardioMining is the first nationwide study involving AI for the automated data extraction from EHRs of patients discharged from Greece's largest Cardiology Departments of tertiary hospitals. This ongoing, retrospective, multicentre, observational cohort study aims to use novel NLP and ML techniques to efficiently extract and process large volumes of unstructured clinical data from electronic clinical narratives, forming the most extensive national dataset.

The target cohort size is 100 000 consecutively enrolled adult patients hospitalised for any reason across all participating study sites. The study is active since January 2022 and is expected to be completed by January 2025. Inclusion and exclusion criteria are presented in box 1.

Box 1 Inclusion and exclusion criteria

Inclusion criteria

⇒ Patients discharged from cardiology departments of tertiary hospitals in Greece.

⇒ Patients whose medical records are electronically stored in each hospital's electronic information systems.

Exclusion criteria

⇒ Patients who died during hospitalisation, and thus no discharge letter was issued.

The resulting database will be a springboard for clinically valuable conclusions, such as outcome prediction, risk stratification and clinical decision support systems. Our protocol has been developed according to the Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence extension (online supplemental appendix).[13]

Details of data

Electronically registered medical records of patients discharged from Cardiology wards of tertiary hospitals in Greece from the day that each hospital developed EHRs until 2022 will be retrospectively collected from hospital discharge letters. Each discharge letter includes demographics, discharge diagnoses, medications, diagnostic examinations, therapeutic interventions, in-hospital management and postdischarge instructions in the Greek language and in unstructured form (box 2). Baseline clinical data are neither coded nor in structured form, apart from the laboratory exams which are structured in a predefined format but will, nevertheless, also require development of automated algorithms to be integrated in the final dataset. Patients who died during hospitalisation were excluded from the study, since no electronically registered discharge letter is issued for these patients in Greek hospitals. Hence, information for these patients is not available, apart from a handwritten administrative document that displays the reason of death as an International Classification of Diseases 10th Revision (ICD-10) code.

Data deidentification

All information that could potentially be used to identify a person, such as names, postal codes, places of residence and occupation, will be deleted from these electronic files before data extraction. Thus, the data will not be able to be assigned to a specific subject, as no additional information or identifiers will be collected for the subjects. After the files are deidentified, each patient's clinical note will be linked with a specific key ('identifier'). The electronic file that contains the correlation of the 'identifier' with the patient's clinical note will be stored in a secure hospital electronic location. Data will be centrally stored in a structured electronic database and only accessible by study staff. Strict subject confidentiality will be maintained through subject identification codes.

Box 2 Extracted data for the discharged patients

1. Patient characteristics
⇒ Demographics.
⇒ History—comorbidities.
⇒ Clinical presentation.
2. Discharge diagnosis
⇒ Coronary artery disease.
⇒ Acute coronary syndrome
  ⇒ST-elevation myocardial infarction.
  ⇒Non-ST-elevation myocardial infarction.
  ⇒Unstable angina.
⇒ Heart failure.
⇒ Sinus tachycardia.
⇒ Sinus bradycardia.
⇒ Supraventricular arrhythmia.
⇒ Ventricular arrhythmia.
⇒ Bradyarrhythmias.
⇒ Cardiac arrest.
⇒ Cardiogenic shock.
⇒ Device implantation or malfunction.
⇒ Endocarditis.
⇒ Myocarditis.
⇒ Pericarditis.
⇒ Pericardial effusion.
⇒ Congenital heart disease.
⇒ Presyncope/syncope.
⇒ Valvular heart disease.
⇒ Cardiomyopathy.
⇒ Acute pulmonary oedema.
⇒ Pulmonary embolism.
⇒ Arterial hypertension.
⇒ Amyloidosis.
3. Electrocardiography on admission and on discharge.
4. Chest X-ray.
5. Echocardiographic reports (quantitative and qualitative data).
6. Catheterisation lab reports.
7. Laboratory examination (serial measurements).
8. In-hospital clinical course and medical management.
9. Discharge medication.

Data and safety monitoring

At multiple time points, a data and safety monitoring board consisting of study investigators and an independent statistician will review accumulating data for quality and safety and report back to the study's steering committee. A simple data use agreement and verification that the researcher has undergone human subjects training will be required to limit access to the clinical database to authorised medical researchers.

AI-model development for automated extraction of data from EHRs

A sample of the fully deidentified files will undergo manual extraction (figure 1). It will serve as a data set for training and evaluating NLP techniques to extract cardiology entities from the records automatically. Of these records, 70% will be used for training, 15% for validation and 15% for testing the developed models. As a baseline, we will use a dictionary-based method containing various forms of the entities we aim to extract. We will employ two main approaches for automating the data extraction process, one operating at the level of the whole record and one operating at the finer-grain level of each entity mention. For the first one, we will investigate the state-of-the-art neural architecture of transformers, as well as more classical linear models and support vector machines, based on the bag-of-words representation of the records, treating the entities as labels in a multilabel classification task. For the second one, we will use the baseline to automatically tag particular words and phrases corresponding to entity mentions or alternatively employ manual annotation. Then we will use a transformer architecture to perform sequence tagging, that is, outputting the particular tokens in the record that correspond to each recognised entity, apart from the recognised entity. A variation of this second approach is first to use a Named Entity Recognition model for generic cardiology entities, followed by an entity normalisation model that links the entity mentions to the particular entity.[14] In all cases, we will exploit information concerning the structure of the record into meaningful sections (figure 2).

Assuming our final model manages to achieve high accuracy on the test set, we will apply it to the complete set of records that have not been manually processed to extract all Cardiology entities from them automatically. This structured information from the discharge letters will be integrated with laboratory measurements and imaging data using a multimodal deep learning method to improve risk stratification and prognosis estimation for patients with different cardiac diseases and treatment recommendations. Following the development of the digital research infrastructure, a pilot prospective study in the participating cardiology departments will be performed to demonstrate the feasibility and accuracy of the AI-algorithm for automated extraction and processing of unstructured clinical data from EHRs. A successful validation of this tool will enable its rapid implementation in the clinical practice.

Study endpoints

The primary and secondary endpoints of this study are summarised in box 3.

The primary endpoint is the measurement of the 'test error'; the model's accuracy to automatically extract clinical data from patients' medical records for further processing and analysis compared with traditional human intervention-based data extraction methods. The obtained baseline clinical data will be merged with the study's secondary endpoints that will be explored on the resulting dataset over the follow-up period of each patient. The endpoints include all-cause mortality, thromboembolic events, number and cause of rehospitalisation, development of new-onset cardiovascular diseases and postdischarge modifications in medication. Secondary endpoints will be either provided in a structured form or extracted from electronic healthcare systems using the aforementioned text mining capabilities of the deployed

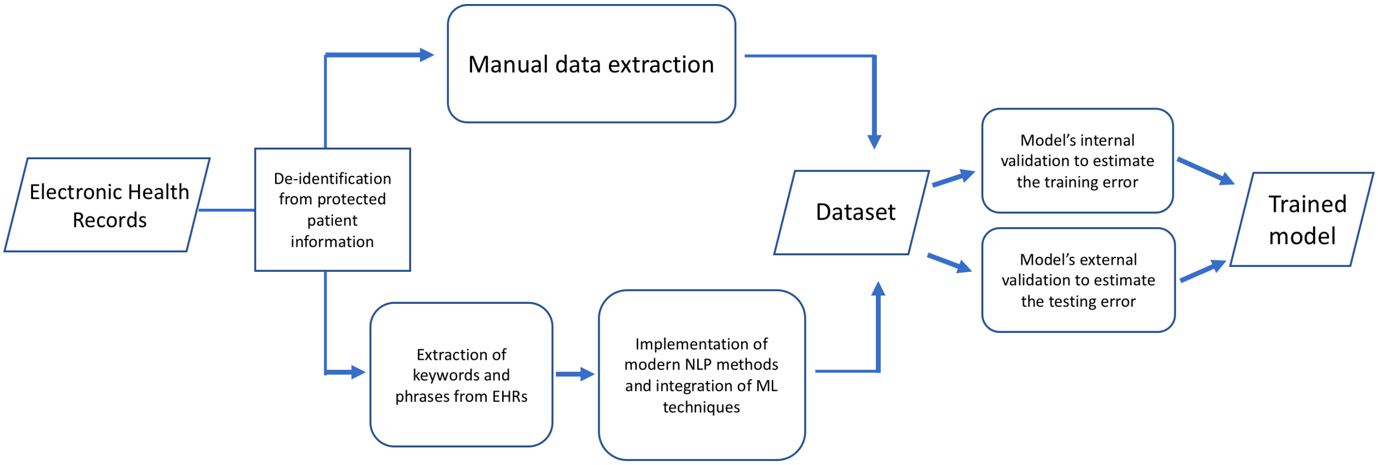
Electronic Health Records
De-identification from protected patient information
Manual data extraction
Extraction of keywords and phrases from EHRs
Implementation of modern NLP methods and integration of ML techniques
Dataset
Model's internal validation to estimate the training error
Model's external validation to estimate the testing error
Trained model

Figure 1 Data extraction method. Data from deidentified electronic health records will undergo both manual and automated extraction. The results of the automated extraction using NLP techniques will be validated on the manually organised dataset using accuracy metrics of the NLP model. The obtained knowledge from the manual dataset will help with the development of an accurate trained AI-model for automated data extraction. EHR, electronic health record; ML, machine learning; NLP, natural language processing.

digital research infrastructure. These endpoints will be integrated in ML models to enable postprocessing data analytics, such as risk stratification for each clinical condition, phenotyping and patient clustering. Hence, the purpose of secondary endpoints is not to test the accuracy of the AI model but rather to provide clinical implications in the digital research infrastructure.

In most healthcare systems, a large volume of clinical data is stored in electronic hospital systems which function as data repositories with no functionalities in terms of data analysis. All these stored data cannot be automatically reshaped in a structured format for analytical purposes and, therefore, require laborious and time-consuming manual extraction by humans. Thus, clinical data are underused and neglected, which results in lack of epidemiological data, research opportunities and loss of valuable clinical information. Optimal utilisation of the increasing volume of clinical data from unstructured clinical notes is a major unmet need in healthcare systems. Hence, the conceptual architecture of our study protocol

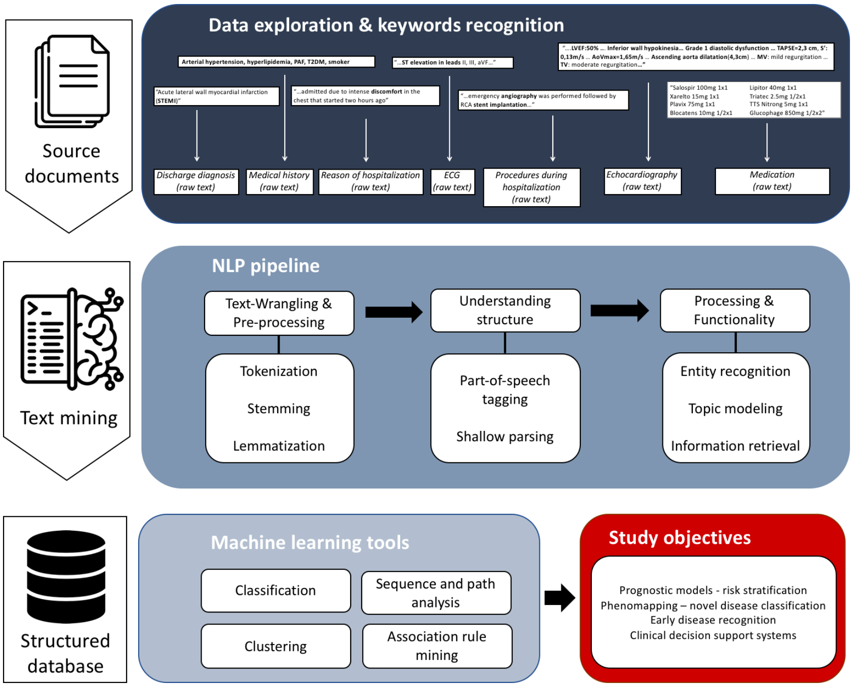
Source documents
Data exploration & keywords recognition
Arterial hypertension, hyperlipidemia, PAF, T2DM, smoker
"…ST elevation in leads II, III, aVF…"
"….LVEF:50% … Inferior wall hypokinesia… Grade 1 diastolic dysfunction … TAPSE=2,3 cm, S': 0,13m/s … AoVmax=1,65m/s … Ascending aorta dilatation(4,3cm) … MV: mild regurgitation … TV: moderate regurgitation…"
"Acute lateral wall myocardial infarction (STEMI)"
"…admitted due to intense discomfort in the chest that started two hours ago"
"…emergency angiography was performed followed by RCA stent implantation…"
"Salospir 100mg 1x1
Xarelto 15mg 1x1
Plavix 75mg 1x1
Blocatens 10mg 1/2x1
Lipitor 40mg 1x1
Triatec 2.5mg 1/2x1
TTS Nitrong 5mg 1x1
Glucophage 850mg 1/2x2"
Discharge diagnosis (raw text)
Medical history (raw text)
Reason of hospitalization (raw text)
ECG (raw text)
Procedures during hospitalization (raw text)
Echocardiography (raw text)
Medication (raw text)
Text mining
NLP pipeline
Text-Wrangling & Pre-processing
Understanding structure
Processing & Functionality
Tokenization
Stemming
Lemmatization
Part-of-speech tagging
Shallow parsing
Entity recognition
Topic modeling
Information retrieval
Structured database
Machine learning tools
Classification
Sequence and path analysis
Clustering
Association rule mining
Study objectives
Prognostic models - risk stratification
Phenomapping – novel disease classification
Early disease recognition
Clinical decision support systems

Figure 2 Utilisation text mining techniques to extract knowledge from unstructured clinical notes. Keyword recognition will provide the baseline information for treating the entities as labels in a multilabel classification task. Text mining techniques will allow the development of a structured database for further processing through machine learning models. PAF, paroxysmal atrial fibrillation; T2DM, Type 2 diabetes mellitus; aVF, augmented Vector Foot; LVEF, left ventricular ejection fraction; TAPSE, Tricuspid annular plane systolic excursion; RCA, right coronary artery; TTS, transdermal therapeutic system.

Box 3 Primary and secondary endpoints

Primary endpoint

⇒ The accuracy of an artificial intelligence-based model to automatically extract clinical data from patients' medical records for further processing and analysis compared with traditional human intervention-based data extraction methods.

Secondary endpoints

⇒ All-cause mortality.
⇒ Thromboembolic events.
⇒ Number and cause of rehospitalisations.
⇒ Development of major cardiovascular diseases (eg, heart failure, coronary artery disease, diabetes mellitus).
⇒ Postdischarge modifications in medical therapy.
⇒ Prescription and use of guideline-recommended drugs in various cardiovascular diseases.

(figure 3) includes the development of an AI-model that enables automated data extraction from unstructured EHRs, which will contribute to the rapid development of a structured cardiovascular database to facilitate further data processing and provide useful data analytics (prevalence of diseases, risk stratification, early diagnosis, clinical decision support systems, minimisation of human error).

Follow-up

The follow-up period in this retrospective study will be between the initial hospital discharge and the end of the follow-up, either the date of outcome occurrence or the current date (in outcome-free patients). Updated information for this study concerning secondary endpoints will be obtained from central databases and prescription registers, managed by the Hellenic Ministry of Health services.

Statistical analysis

Given the size of the participating clinics and the years during which the recording of EHRs in electronic form was applied, it is estimated that the sample of patient records will be about 100 000. Continuous variables will be tested for normality with the Kolmogorov-Smirnov test and presented as a mean SD or medians, with comparisons between groups made using the Wilcoxon rank-sum test. Categorical variables will be expressed as frequencies (%), with comparisons made using the Pearson's χ^2 test. Outcome analysis will be performed using ML algorithms, such as regression, decision trees, random forests, support vector machines, extreme gradient boosting. Statistical analysis will be performed using SPSS V.27 (SPSS), Stata V.15.1 (StataCorp) and R packages.

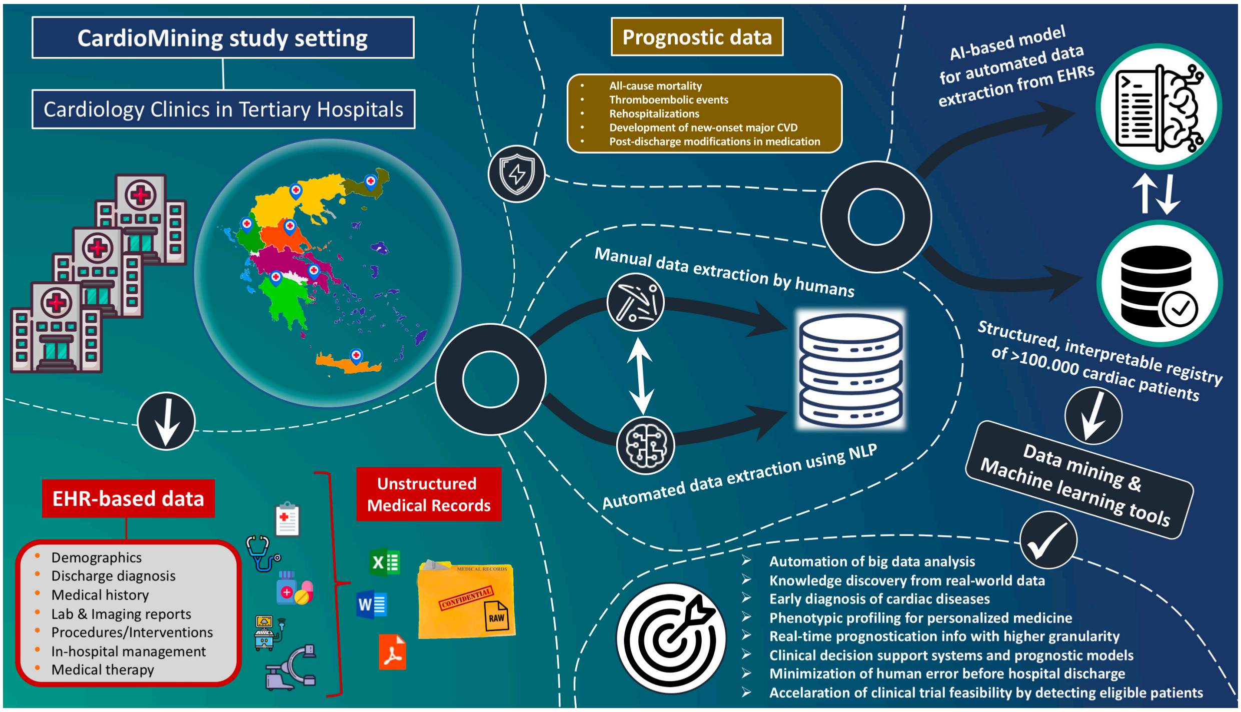
CardioMining study setting
Cardiology Clinics in Tertiary Hospitals
Prognostic data
All-cause mortality
Thromboembolic events
Rehospitalizations
Development of new-onset major CVD
Post-discharge modifications in medication
AI-based model for automated data extraction from EHRs
Manual data extraction by humans
Automated data extraction using NLP
Structured, interpretable registry of >100.000 cardiac patients
Data mining & Machine learning tools
EHR-based data
Demographics
Discharge diagnosis
Medical history
Lab & Imaging reports
Procedures/Interventions
In-hospital management
Medical therapy
Unstructured Medical Records
CONFIDENTIAL
RAW
Automation of big data analysis
Knowledge discovery from real-world data
Early diagnosis of cardiac diseases
Phenotypic profiling for personalized medicine
Real-time prognostication info with higher granularity
Clinical decision support systems and prognostic models
Minimization of human error before hospital discharge
Accelaration of clinical trial feasibility by detecting eligible patients

Figure 3 The roadmap of the CardioMining study towards digital transformation of the national cardiovascular ecosystem. The digital transformation of a healthcare system at a national level is a great challenge but also a complex and difficult task. The low digital maturity of the health sector in Greece coupled with the ongoing rapid technological changes worldwide demand urgent action through the implementation of a paradigm shift. The CardioMining study will retrospectively collect unstructured data derived from electronic health records of cardiology departments. Data extraction will be performed both manually by humans and automatically using natural language processing algorithms. The validated artificial intelligence models will contribute to the development of a structured registry of cardiac patients to provide data analytics through machine learning models. CVD, cardiovascular disease.

Patient and public involvement

There has been no patient involvement in the design, or conduct, or reporting, or dissemination plans of our research.

Ethics and dissemination

All participating sites will obtain approval from appropriate independent ethics committees or institutional review boards prior to the initiation of the study. The Research Ethics Committee of the Aristotle University of Thessaloniki and Scientific and Ethics Council of the AHEPA University Hospital have approved this study. This study will be conducted in keeping with the International Conference on Harmonisation Good Clinical Practice guidelines, the Declaration of Helsinki, the Data Protection Code of the European Data Protection Authority and the European General Data Protection Regulation or otherwise that may replace it. To maintain the patient's confidentiality, no demographic and personal identification data will be collected (eg, first name, date of birth). Data deidentification will be performed by the lead researcher. Only deidentified data will be obtained, which will be completely disconnected from the personal data of each patient. Study findings will be published in peer-reviewed medical journals and presented at international conferences. Collaborations with study groups sharing the same research focus will be attempted.

Author affiliations

[1]1st Department of Cardiology, University General Hospital of Thessaloniki AHEPA, Thessaloniki, Greece
[2]Institute of Bioelectronic Medicine, Feinstein Institutes for Medical Research, New York, New York, USA
[3]School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
[4]Medical Physics and Digital Innovation Laboratory, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece
[5]Department of Cardiology, University Hospital of Patras, Rio Patras, Greece
[6]1st Cardiology Department, "Hippokration" General Hospital, University of Athens Medical School, Athens, Greece
[7]2nd Cardiology Department, Hippokrateion General Hospital, Aristotle University of Thessaloniki, Thessaloniki, Greece
[8]3rd Cardiology Department, Hippokrateion General Hospital, Aristotle University of Thessaloniki, Thessaloniki, Greece
[9]Department of Cardiology, "George Papanikolaou" General Hospital, Thessaloniki, Greece
[10]Department of Cardiology, Democritus University of Thrace, University Hospital of Alexandroupolis, Alexandroupolis, Greece

Twitter Alexandra Bekiaridou @ampekiaridou

Acknowledgements We would like to acknowledge the following investigators: Data collection: Petrina Miltiadous, Eleni Aintinidou, Anastasia Chatzisavvidou, Giannis Papamichail, Giannis Giannou, Konstantinos Konstantinidis, Andromachi Keskinidou, Konstantinos Rizos, Christos Petridis, Isidoros Karamanidis, Nikolaos Mantzou, Christina Moutafi, Zeta Kousourna, Theocharis Spiritinoudis, Mavridou Christina, Ioannis Botis, Eirini Lazari, Evaggelia Mermigka, Christina-Angeliki Papadaki, Ntalaoutis Andreas, Androniki Papadopoulou, Ioannis Leventis, Stylianos Zervakis, Anthi Pleuritaki, Eirini Savva, Spiridon Kassotakis, Vaso Papacosta, Themistoklis Pateromichelakis, Panagiota Serafeim, Eleni Sidiropoulou, Despoina Ntiloudi, Nikolaos Papakonstantinou. Data analysis: Evangelos Logaras, Dimitris Dimitriadis, Parmenion Charistos.

Collaborators CardioMining Study Group AHEPA University Hospital of Thessaloniki (Greece): Athanasios Feidakis, Anastasios Kartas, Vasiliki Patsiou, Eirinaios Tsiartas, Antonios Orfanidis, Triantafyllia Grantza, Chrysanthi Ioanna Lampropoulou, Dimitrios Kostakakis, Olga Kazarli, Maria Ioannou, Maria Eirini Kiriakideli, Melina Kyriakou, Dimitra Kontopyrgou, Martha Zergioti, Eleftherios Gemousakakis, Amalia Baroutidou, Alexios Vagianos, Alexandros Liatsos, Konstantinos Barmpagiannos, George Tyrikos, George Konstantinou, Anthi Vasilopoulou, Marina Spaho, Eleni Manthou, Panagiotis Zymaris, Eleni Baliafa, Maria Baloka, Iasonas Dermitzakis, Vasiliki Anagnostopoulou, Chrysi Solovou, Anna Maria Louka, Aliki Iliadou, Ioanna Filimidou, Aspasia Kyriafini, Odysseas Kamzolas, Ioannis Vouloagkas, Despoina Nteli, Nikolaos Outountzidis, Athanasia Vathi, Anastasia Foka, Michael Botis, Anastasia Christodoulou, George Vogiatzis, Eleni Vrana, Maria Nteli, Stefanos Antoniadis, Foteini Charisi, Mairifylli Vamvaka, Dimitrios Triantis, Efi Delilampou, Vaggelis Axarloglou, Georgios Charistos, George Anagnostou, Sofia Christodoulou, Anastasios Papanastasiou, Eleni Tziona, Nikolaos Batis, Katerina Gakidi, Artemis Iosifidou, Andreanna Moura, Christos Alexandropoulos, Theoni Exintaveloni, Asterios Karakoutas, Damianos Porfyropoulos, Michail Bountas, Athanasios Pachoumis, Eleftherios Markidis, Maria Sitmalidou, Athanasia Pappa, Konstantinos C Theodoropoulos, George Rampidis, Apostolos Tzikas, Stylianos Paraskevaidis, Georgios Efthimiadis, Antonios Ziakas. University Hospital of Patras (Greece): Theofilatos Athinagoras, Christoforos Travlos, Nikolaos Vythoulkas-Biotis, Kassiani Maria Nastouli, Nikolaos Kartas, Angeliki Vakka, Theoni Theodoropoulou, Maria Bozika, Virginia Anagnostopoulou, Georgios Tsioulos. Hippokration General Hospital of Athens (Greece): Emilia Lazarou, Panagiotis Tsioufis, Ioannis Kachrimanidis, Nick Argyriou. University Hospital of Heraklion (Greece): Emmanouil Kampanieris, Alexandros Patrianakos, George Kochiadakis. Alexandra General Hospital of Athens (Greece): Ioannis Kanakakis. University Hospital of Alexandroupolis (Greece): Marios Vasileios Koutroulos, Georgios K Chalikias. Hippokrateion University Hospital of Thessaloniki (Greece): Sophia Alexiou, Athena Nasoufidou, Panagiotis Stachteas, Constantinos Bakogiannis. University Hospital of Larissa (Greece): Tsantikos Christos, Grigorios Giamouzis, John Skoularigis. Evangelismos General Hospital of Athens (Greece): Ioannis Alexanian. Papageorgiou General Hospital of Thessaloniki (Greece): Dimitrios Farmakis, Ioannis Styliadis. Hellenic Red Cross Hospital of Athens (Greece): George Fotos, Nikolaos Bourboulis. Tzaneio General Hospital of Piraeus (Greece): Despoina Ntiloudi, Evangelos Pisimisis. Hygeia Hospitals Group (Greece) & European Heart Agency, European Society of Cardiology (Belgium): Panos E Vardas. Aristotle University of Thessaloniki (Greece): Antonis Billis, Ilias Kyparissidis. University of Western Macedonia (Greece): Dimitrios Tsalikakis. Beuth University of Applied Sciences (Berlin, Germany): Jens-Michael Papaioannou, Alexander Löser.

Contributors Substantial contributions to the conception or design of the work (AS, GG); or the acquisition, analysis or interpretation of data for the work (AS, AB, ASP, DVM, GTso, PB, GTsi, GL, GK, NF, VV, IZ, DNT, KT, PD, GG, CardioMining Study Group); and Drafting the work (AS, AB) or revising it critically for important intellectual content (ASP, DVM, GTso, PB, GTsi, GL, GK, NF, VV, IZ, DNT, KT, PD, GG, CardioMining Study Group); AND Final approval of the version to be published (AS, AB, ASP, DVM, GTso, PB, GTsi, GL, GK, NF, VV, IZ, DNT, KT, PD, GG, CardioMining Study Group); and Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved (AS, AB, ASP, DVM, GTso, PB, GTsi, GL, GK, NF, VV, IZ, DNT, KT, PD, GG, CardioMining Study Group).

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs

George Kassimis http://orcid.org/0000-0003-4485-5054
George Giannakoulas http://orcid.org/0000-0001-7491-6319

REFERENCES

1 Shivade C, Raghavan P, Fosler-Lussier E, *et al*. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2014;21:221–30.
2 Krittanawong C, Zhang H, Wang Z, *et al*. Artificial intelligence in precision cardiovascular medicine. *J Am Coll Cardiol* 2017;69:2657–64.
3 Madani A, Arnaout R, Mofrad M, *et al*. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit Med* 2018;1:6.
4 Chary M, Parikh S, Manini AF, *et al*. A review of natural language processing in medical education. *West J Emerg Med* 2019;20:78–86.
5 Deo RC. Machine learning in medicine. *Circulation* 2015;132:1920–30.
6 Nasir K, DeFilippis A. Big data and ASCVD risk prediction: building a better mouse trap? *J Am Coll Cardiol* 2022;79:1167–9.
7 Shickel B, Tighe PJ, Bihorac A, *et al*. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform* 2018;22:1589–604.
8 Mehta N, Devarakonda MV. Machine learning, natural language programming, and electronic health records: the next step in the artificial intelligence journey? *J Allergy Clin Immunol* 2018;141:2019–21.
9 Boag W, Doss D, Naumann T, *et al*. What's in a note? Unpacking predictive value in clinical note representations. n.d. Available: http://www.github.com/wboag/wian
10 Hashir M, Sawhney R. Towards unstructured mortality prediction with free-text clinical notes. *J Biomed Inform* 2020;108:103489.
11 Diller G-P, Kempny A, Babu-Narayan SV, *et al*. Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: data from a single tertiary centre including 10 019 patients. *Eur Heart J* 2019;40:1069–77.
12 Kim J. Big data, health informatics, and the future of cardiovascular medicine. *J Am Coll Cardiol* 2017;69:899–902.
13 Cruz Rivera S, Liu X, Chan A-W, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health* 2020;2:e549–60.
14 Ji Z, Wei Q, Xu H. *BERT-based ranking for biomedical entity normalization*.

SPIRIT

STANDARD PROTOCOL ITEMS: RECOMMENDATIONS FOR INTERVENTIONAL TRIALS

Section/item	**Item No**	**Description**	**page**
Administrative information			
Title	1	Descriptive title identifying the study design, population, interventions, and, if applicable, trial acronym	1
		Indicate that the intervention involves artificial intelligence/machine learning learning and specify the type of model. Specify the intended use of the AI intervention.	1
Trial registration	2a	Trial identifier and registry name. If not yet registered, name of intended registry	6
	2b	All items from the World Health Organization Trial Registration Data Set	-
Protocol version	3	Date and version identifier	6
Funding	4	Sources and types of financial, material, and other support	18
Roles and responsibilities	5a	Names, affiliations, and roles of protocol contributors	18
	5b	Name and contact information for the trial sponsor	N/A
	5c	Role of study sponsor and funders, if any, in study design; collection, management, analysis, and interpretation of data; writing of the report; and the decision to submit the report for publication, including whether they will have ultimate authority over any of these activities	N/A
	5d	Composition, roles, and responsibilities of the coordinating centre, steering committee, endpoint adjudication committee, data management team, and other individuals or groups overseeing the trial, if applicable (see Item 21a for data monitoring committee)	N/A
Introduction			

Background and rationale	6a	Description of research question and justification for undertaking the trial, including summary of relevant studies (published and unpublished) examining benefits and harms for each intervention *Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (eg, healthcare professionals, patients, public). Describe any pre-existing evidence for the AI intervention.*	8-9 8-9
	6b	Explanation for choice of comparators	9-16
Objectives	7	Specific objectives or hypotheses	9
Trial design	8	Description of trial design including type of trial (eg, parallel group, crossover, factorial, single group), allocation ratio, and framework (eg, superiority, equivalence, noninferiority, exploratory)	9
Methods: Participants, interventions, and outcomes			
Study setting	9	Description of study settings (eg, community clinic, academic hospital) and list of countries where data will be collected. Reference to where list of study sites can be obtained *Describe the onsite and offsite requirements needed to integrate the AI intervention into the trial setting.*	9
Eligibility criteria	10	Inclusion and exclusion criteria for participants. If applicable, eligibility criteria for study centres and individuals who will perform the interventions (eg, surgeons, psychotherapists) *State the inclusion and exclusion criteria at the level of participants. State the inclusion and exclusion criteria at the level of the input data.*	10
Interventions	11a	Interventions for each group with sufficient detail to allow replication, including how and when they will be administered *State which version of the AI algorithm will be used. Specify the procedure for acquiring and selecting the input data for the AI intervention. Specify the procedure for assessing and handling poor quality or unavailable input data. Specify whether there is human-AI interaction in the handling of the input data, and what level of expertise is required for users. Specify the output of the AI intervention. Explain the procedure for how the AI intervention's output will contribute to decision-making or other elements of clinical practice.*	N/A 13-14

	11b	Criteria for discontinuing or modifying allocated interventions for a given trial participant (eg, drug dose change in response to harms, participant request, or improving/worsening disease)	N/A
	11c	Strategies to improve adherence to intervention protocols, and any procedures for monitoring adherence (eg, drug tablet return, laboratory tests)	N/A
	11d	Relevant concomitant care and interventions that are permitted or prohibited during the trial	N/A
Outcomes	12	Primary, secondary, and other outcomes, including the specific measurement variable (eg, systolic blood pressure), analysis metric (eg, change from baseline, final value, time to event), method of aggregation (eg, median, proportion), and time point for each outcome. Explanation of the clinical relevance of chosen efficacy and harm outcomes is strongly recommended	14-15
Participant timeline	13	Time schedule of enrolment, interventions (including any run-ins and washouts), assessments, and visits for participants. A schematic diagram is highly recommended (see Figure)	N/A
Sample size	14	Estimated number of participants needed to achieve study objectives and how it was determined, including clinical and statistical assumptions supporting any sample size calculations	10,16
Recruitment	15	Strategies for achieving adequate participant enrolment to reach target sample size	N/A
Methods: Assignment of interventions (for controlled trials)			**N/A**
Allocation:			
Sequence generation	16a	Method of generating the allocation sequence (eg, computer-generated random numbers), and list of any factors for stratification. To reduce predictability of a random sequence, details of any planned restriction (eg, blocking) should be provided in a separate document that is unavailable to those who enrol participants or assign interventions	
Allocation concealment mechanism	16b	Mechanism of implementing the allocation sequence (eg, central telephone; sequentially numbered, opaque, sealed envelopes), describing any steps to conceal the sequence until interventions are assigned	

Implementation	16c	Who will generate the allocation sequence, who will enrol participants, and who will assign participants to interventions	
Blinding (masking)	17a	Who will be blinded after assignment to interventions (eg, trial participants, care providers, outcome assessors, data analysts), and how	
	17b	If blinded, circumstances under which unblinding is permissible, and procedure for revealing a participant’s allocated intervention during the trial	
Methods: Data collection, management, and analysis			
Data collection methods	18a	Plans for assessment and collection of outcome, baseline, and other trial data, including any related processes to promote data quality (eg, duplicate measurements, training of assessors) and a description of study instruments (eg, questionnaires, laboratory tests) along with their reliability and validity, if known. Reference to where data collection forms can be found, if not in the protocol	12-15
	18b	Plans to promote participant retention and complete follow-up, including list of any outcome data to be collected for participants who discontinue or deviate from intervention protocols	N/A
Data management	19	Plans for data entry, coding, security, and storage, including any related processes to promote data quality (eg, double data entry; range checks for data values). Reference to where details of data management procedures can be found, if not in the protocol	12-17
Statistical methods	20a	Statistical methods for analysing primary and secondary outcomes. Reference to where other details of the statistical analysis plan can be found, if not in the protocol	16
	20b	Methods for any additional analyses (eg, subgroup and adjusted analyses)	16
	20c	Definition of analysis population relating to protocol non-adherence (eg, as randomised analysis), and any statistical methods to handle missing data (eg, multiple imputation)	16
Methods: Monitoring			

Data monitoring	21a	Composition of data monitoring committee (DMC); summary of its role and reporting structure; statement of whether it is independent from the sponsor and competing interests; and reference to where further details about its charter can be found, if not in the protocol. Alternatively, an explanation of why a DMC is not needed	13
	21b	Description of any interim analyses and stopping guidelines, including who will have access to these interim results and make the final decision to terminate the trial	N/A
Harms	22	Plans for collecting, assessing, reporting, and managing solicited and spontaneously reported adverse events and other unintended effects of trial interventions or trial conduct *Specify any plans to identify and analyse performance errors. If there are no plans for this, explain why not.*	N/A
Auditing	23	Frequency and procedures for auditing trial conduct, if any, and whether the process will be independent from investigators and the sponsor	N/A
Ethics and dissemination			
Research ethics approval	24	Plans for seeking research ethics committee/institutional review board (REC/IRB) approval	16,17
Protocol amendments	25	Plans for communicating important protocol modifications (eg, changes to eligibility criteria, outcomes, analyses) to relevant parties (eg, investigators, REC/IRBs, trial participants, trial registries, journals, regulators)	N/A
Consent or assent	26a	Who will obtain informed consent or assent from potential trial participants or authorised surrogates, and how (see Item 32)	N/A
	26b	Additional consent provisions for collection and use of participant data and biological specimens in ancillary studies, if applicable	N/A
Confidentiality	27	How personal information about potential and enrolled participants will be collected, shared, and maintained in order to protect confidentiality before, during, and after the trial	16
Declaration of interests	28	Financial and other competing interests for principal investigators for the overall trial and each study site	12,13

Access to data	29	Statement of who will have access to the final trial dataset, and disclosure of contractual agreements that limit such access for investigators	12-17
		State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use.	N/A
Ancillary and post-trial care	30	Provisions, if any, for ancillary and post-trial care, and for compensation to those who suffer harm from trial participation	N/A
Dissemination policy	31a	Plans for investigators and sponsor to communicate trial results to participants, healthcare professionals, the public, and other relevant groups (eg, via publication, reporting in results databases, or other data sharing arrangements), including any publication restrictions	16-17
	31b	Authorship eligibility guidelines and any intended use of professional writers	N/A
	31c	Plans, if any, for granting public access to the full protocol, participant-level dataset, and statistical code	N/A
Appendices			
Informed consent materials	32	Model consent form and other related documentation given to participants and authorised surrogates	N/A
Biological specimens	33	Plans for collection, laboratory evaluation, and storage of biological specimens for genetic or molecular analysis in the current trial and for future use in ancillary studies, if applicable	N/A

*It is strongly recommended that this checklist be read in conjunction with the SPIRIT 2013 Explanation & Elaboration for important clarification on the items. Amendments to the protocol should be tracked and dated. The SPIRIT checklist is copyrighted by the SPIRIT Group under the Creative Commons "Attribution-NonCommercial-NoDerivs 3.0 Unported" license.