BMJ Open

The Salmanticor Study. Rationale and Design of a Population-based Study to Identify Structural Heart Disease Abnormalities: a Spatial and Machine Learning Analysis

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-024605
Article Type:	Protocol
Date Submitted by the Author:	11-Jun-2018
Keywords:	structural heart disease, population, rural, urban, spatial analysis, machine learning

Complete List of Authors:
Melero-Alegria, Jose Ignacio; Hospital Universitario de Salamanca-IBSAL, Cardiology
Cascon, Manuel; Hospital Universitario de Salamanca-IBSAL, Cardiology
Romero, Alfonso; Miguel Armijo Primary Care Center
Vara, Pedro P; Hospital Universitario de Salamanca-IBSAL, Cardiology
Barreiro-Perez, Manuel; Hospital Universitario de Salamanca-IBSAL, Cardiology
Vicente-Palacios, Victor; Hospital Universitario de Salamanca-IBSAL, Cardiology
Perez-Escanilla, Fernando; San Juan Primary Care Center
Hernandez-Hernandez, Jesus; Hospital Universitario de Salamanca-IBSAL
Garde, Beatriz; Hospital Universitario de Salamanca-IBSAL, Cardiology
Cascon, Sara; Robleda Primary Care Center
Martin-Garcia, Ana; Hospital Universitario de Salamanca-IBSAL, Cardiology
Díaz-Pelaez, Elena; Hospital Universitario de Salamanca-IBSAL, Cardiology
de Dios, Jose Maria; Salamanca Primary Care Center Management
Uribarri, Aitor; Hospital Universitario de Salamanca-IBSAL
Jimenez-Candil, Javier; Hospital Universitario de Salamanca-IBSAL
Cruz-Gonzalez, Ignacio; Hospital Universitario de Salamanca-IBSAL, Cardiology
Blazquez, Baltasara; Miranda del Castañar Primary Care Center
Hernandez, Jose Manuel; Miranda del Castañar Primary Care Center
Sanchez-Pablos, Clara; Hospital Universitario de Salamanca-IBSAL
Santolino, Inmaculada; Santa Marta Primary Care Center
Ledesma, Maria Concepcion; Peñaranda de Bracamonte Primary Care Center
Muriel, Paz; Miguel Armijo Primary Care Center
Dorado-Diaz, P. Ignacio; Hospital Universitario de Salamanca-IBSAL
Sanchez, Pedro L; University of Salamanca, Cardiology

THE SALMANTICOR STUDY. RATIONALE AND DESIGN OF A POPULATION-BASED STUDY TO IDENTIFY STRUCTURAL HEART DISEASE ABNORMALITIES: A SPATIAL AND MACHINE LEARNING ANALYSIS

José Ignacio Melero-Alegría, RN[1]
Manuel Cascón, MD, PhD[1]
Alfonso Romero, MD[2]
Pedro Pablo Vara, DCS[1]
Manuel Barreiro-Pérez, MD, PhD[1]
Victor Vicente-Palacios, PhD[1]
Fernando Pérez-Escanilla, MD, PhD[3]
Jesús Hernández-Hernández, MD[1]
Beatriz Garde, BPharm[1]
Sara Cascón, MD, PhD[4]
Ana Martín-García, MD, PhD[1]
Elena Díaz-Peláez, MD[1]
José María de Dios, MD[5]
Aitor Uribarri, MD, PhD[1]
Javier Jiménez-Candil, MD, PhD[1]
Ignacio Cruz-González, MD, PhD[1]
Baltasara Blazquez, MD[6]
José Manuel Hernández, MD[6]
Clara Sánchez-Pablos, RN[1]
Inmaculada Santolino, MD[7]
María Concepción Ledesma, MD[8]
Paz Muriel, MD[2]
P. Ignacio Dorado-Díaz, MD[1]
Pedro L Sanchez, MD, PhD[1§]

From the
[1]Department of Cardiology, Hospital Universitario de Salamanca, Instituto de Investigación Biomédica de Salamanca (IBSAL), Facultad de Medicina, Universidad de Salamanca, and CIBERCV, Salamanca, Spain
[2]Miguel Armijo Primary Care Center, Salamanca, Spain
[3]San Juan Primary Care Center, Spain
[4]Robleda Primary Care Center, Salamanca, Spain
[5]Salamanca Primary Care Center Management, Salamanca, Spain
[6]Miranda del Castañar Primary Care Center, Salamanca, Spain
[7]Santa Marta Primary Care Center, Salamanca, Spain
[8]Peñaranda de Bracamonte Primary Care Center, Salamanca, Spain

BRIEF TITLE: The SALMANTICOR Study

Address for Correspondence[§]:

Pedro L Sanchez, MD, PhD.
Cardiology Department. Hospital Universitario de Salamanca-IBSAL.
Paseo de San Vicente 58-187. 38007 Salamanca. SPAIN.
Telephone: 34-923291100 (ext 55356).
e-mail: pedrolsanchez@secardiologia.es

Financial Support:
This study was supported by national (PI14/00695, Institute of Health Carlos III, Spanish Ministry of Economy and Competitiveness) and community (GRS1030/A/14, SACYL, Junta Castilla y León) competitive grants and by the Spanish Cardiovascular Network (RIC and CIBERCV) from the Institute of Health Carlos III, Spanish Ministry of Economy and Competitiveness, Obra Social "la Caixa" and Philips Ibérica Healthcare division.

Acknowledgements:
We thank all primary care physicians and personnel helping with the development of the study. We thank Philips Iberica and Obra Social "La Caixa" for their support. We specially thank participants in the study and apologize for any inconvenient we could cause. We thank the involvement of the Salamanca patient organisation "El paciente experto", for providing counselling to SALMANTICOR and for further promoting the dissemination of the results to the society and to the regional government.

Potential Conflicts of Interest: None to disclose.

Word Count: 4715 (excluding references)

Abstract

Objectives. To obtain data on the prevalence and incidence of structural heart disease in a population setting, and to analyze and present those data on the application of spatial and machine learning methods that, although known to geography and statistics, need to become used from healthcare research and from political commitment to obtain resources and support effective public health program implementation.

Methods and analysis. A cross-sectional survey of randomly selected residents of Salamanca (Spain)

Population. 2400 individuals, stratifies by age and sex and by place of residence (rural and urban) will be studied

Measurements. The variables to analyze will be obtained from the clinical history, different surveys including social status, Mediterranean diet, functional capacity, electrocardiogram, echocardiogram, VASERA and biochemical and genetic analysis.

Ethics and dissemination. The study has been approved by the clinical research ethics committee of the health care community. All study participants will sign and informed consent to agree to participate in the study. The results of this study will allow the understanding of the relationship of the different influencing factors and their relative weight in the development of structural heart disease. For the first time, a detailed cardiovascular map showing the spatial distribution and a predictive machine learning system of different structural heart diseases and associated risk factors will be created and will be used as a regional policy to stablish effective public health programs to fight heart disease. At least ten publications in the first-quartile scientific journals are planned.

Trial registration number. NCT03429452; Pre-results.

Abstract word count: 248

Strengths and limitations

- To obtain data on the prevalence and incidence of structural heart disease in the setting of a population-based study and primary care assistance that will enroll a total of 2400 individuals, stratifies by age, sex and by place of residence (rural and urban), in a Spanish community: Salamanca.
- To create a population-based established control group providing availability of normative reference values quantification for echocardiographic, electrocardiographic, VASERA, biochemical and genetics parameters.
- To show the spatial distribution different patterns of structural heart disease through the spectrum of age and sex and between urban and rural residences.
- To develop a predictive model of structural heart disease using cardiovascular heterogeneous data (images including) and machine learning techniques
- The study will be established as the global observatory on cardiovascular health research and development of the regional healthcare government to support effective public health program implementation.

Keywords (MeSH terms)

Structural heart disease · population · rural · urban · spatial analysis · Multiple factor analysis · Principal component analysis · multivariate statistics · Cokriging · geostatistics · machine learning

Abbreviations

ABI	ankle-brachial index
ACE	angiotensin-converting enzyme
ba-PWV	brachial ankle pulse wave velocity
CA	correspondence analysis
CAVI	cardio-ankle vascular index
CEIC	clinical research ethics committee
ECG	electrocardiogram
GP	Gaussian process
MCA	multiple correspondence analysis
MFA	multiple factor analysis
ML	machine learning
NSAIDs	nonsteroidal anti-inflammatory drugs
PACS	picture archiving and communication system
PCA	principal component analysis
RAAS	renin-angiotensin-aldosterone system
VNP	virtual private network
2D	two dimensional

Introduction

Each year heart disease causes almost 4 million deaths in Europe and the United States; that´s 1 in every 4 deaths.[1 2] Although, number of deaths from heart disease has decreased, the burden of heart disease is increasing. In 2015, more than 85 million people in Europe were living with cardiovascular disease.[2] The increase in the prevalence of classical cardiovascular risk factors, dietary factors, physical activity and probably other social factors make the largest contribution to the risk of heart disease. Overall cardiovascular disease health care costs in the European Union and the United States have increased rapidly over the last ten years; currently overpassing €200 billion a year.[2 3]

In this sense, public health delivery planning requires reliable information about contemporary population-level disease prevalence and incidence. Furthermore, community healthcare systems should obtain and provide their own data before implementing any effective health program as these regional systems are highly influenced by geographic diversity, the availability of resources and infrastructure, and the characteristics of healthcare systems and patterns of reimbursement.[4] This is well illustrated by some heart disease examples as the attention of myocardial infarction, where communication of accurate and timely information to the health care community, decision makers, and the public program effects, have been gaining momentum in the recent decade.[5-8]

Policies need to consider both standardized rates, which describe disease prevalence and incidence independently of changes in populations, and absolute numbers of patients affected, which describe the impact of the disease on the population, political commitment, resources and services of interest.[4 9] Limited data exist on estimation of

heart disease prevalence in a population setting. Previous studies have frequently been based on selected cohorts, which may not represent the general population.[10-13] Other studies have restricted case identification to those made in general practice consultations or hospital admissions.[14-16] However, it is only by considering presentations across the whole spectrum of structural heart disease that the full burden of disease can be captured and an accurate distinction made between incident and prevalent cases. Thus, contemporary population-based studies of heart disease prevalence and incidence are needed to inform resource planning and research prioritization but current evidence is scarce.

Spatial analysis are great tools to investigate population behavior, relations and consequently determine future action plans or policies. Spatial methods are varied, ranging from descriptive spatial analysis to complex interpolation algorithms. Gaussian Process (GP) procedures, such as cokriging, have distinct advantages over conventional spatial prediction techniques.[17] They allow researchers to include measured spatial variability in the geostatistical estimation process and they smooth predicted values based on the proportion of total sample variability accounted by random noise. Furthermore, GP helps mitigate the effect of variable sample density caused by hot spots (some zones are usually oversampled). Hence, geostatistics techniques are suitable methods to apply on population studies.

Furthermore, the volume of quantitative and imaging data, generated by population studies, will also be a big driver in the future for research and how we provide care. In this sense, machine learning (ML) to train algorithms to recognize cardiac damage at a better level, avoiding diagnostic errors and improving the early identification of the disease offers new approaches to leveraging the growing volume of data available for

analyses[18-21]. Thus, we are convinced that ML can play a key role in population-based epidemiological studies when trying to early recognize patients-disease vulnerability.

The objectives of this study are: to obtain data on the prevalence and incidence of structural heart disease in a population setting; to show the spatial distribution different patterns of structural heart disease through the spectrum of age and sex and between urban and rural; to develop a predictive model of structural heart disease using cardiovascular heterogeneous data (images including) and ML techniques and; to generate new hypotheses which might serve to healthcare research and to political commitment to obtain resources and support effective public health program implementation.

We describe the design, data and imaging acquisition, analysis methods and quality assurance metrics for the SALMANTICOR study.

Methods

Study Design and Participants

The SALMANTICOR study is a cross-sectional descriptive population-based study of the prevalence of structural heart disease and their risk factors that will enroll a total of 2400 individuals, stratifies by age, sex and by place of residence (rural and urban), in a Spanish community: Salamanca. Structural heart disease refers to any of the following heat abnormalities including congenital heart disease, cardiomyopathies, valvar heart disease, ischemic heart disease, pericardial diseases and rhythm or conduction disorders.

The province of Salamanca is located on the western Spain, bordered in the west by Portugal. It has an area of 12.349 km2 and in 2014 had a population of 342,857 people; 167,459 (49%) male and 175.398 (51%) female people. It is divided into 362

municipalities; more than half are villages with fewer than 300 people. In fact, 227,878 (67%) people live in 10 municipalities of more than 5,000 individuals that will be considered for future analysis as urban areas and 114,581 (33%) people live in the rest of municipalities and consequently will be considered as rural areas.

Spain´s and consequently Salamanca healthcare system is public, guaranteeing universal coverage. In total, 98.7 percent of the population are insured for this public Spanish healthcare system. In Salamanca, a total of 35 primary health centers throughout the province provide healthcare services to the overall population: 18 to the urban-considered municipalities and 17 to the rural-considered municipalities (**Figure 1**).

Individuals aged ≥18 years included in the lists of all primary healthcare facilities of the province of Salamanca represented the reference population of 295,975 subjects: mean age 52.9±19.8 years; 52.4% females; 61.3% residing in urban areas. A sample size of 2400 subjects is calculated based on an expected prevalence of structural heart disease of 6% with a confidence interval of 95% and a 1% precision. In order to obtain the necessary sample size, 35% more requests for participation will be made, estimating errors of location from the healthcare database or refuses to participate in the study. Thus, 3564 people will be randomly selected from the primary care lists.

Cohort participants will undergo a basal examination visit, in these primary healthcare centers, between 2015 and 2018. Surviving participants are expected to return for a 5 and 10-year follow-up visit. Institutional review committee approval was obtained and all participants will provide informed consent. The SALMANTICOR study is designed to provide echocardiographic parameters characterizing cardiac structure and function in all individuals. SALMANTICOR participants will undergo

surveillance for cardiovascular events, including heart failure, incident coronary heart disease, and all-cause mortality.

Medical investigation process

Medical history, surveys completion, and examinations will be obtained at the subject´s primary care referral center and will be analyzed and interpreted centrally at University Hospital of Salamanca. A complete medical history, physical examination and the surveys completion checkout will be performed by a cardiologist in a separate office to where examinations and blood sample extraction will be performed. Echocardiographic measures will be initially performed. Participant blood pressure and VASERA measures will be taken within 30 minutes of starting the echocardiographic exam and after the subject will be resting for 10 minutes. ECG will be performed after VASERA to finalize with the blood sample extraction.

Questionnaires

After obtaining written informed consent, trained interviewers will use a structured questionnaire to collect baseline data in face-to-face interviews at the time of physical examination. Self-reported diseases will be verified by individuals' primary care doctors according to recognized international standards. The questionnaire collected information on demographics and cardiovascular risk factors, cardiovascular and non-cardiovascular medical history, physical examination, medication, socio-economic status, dietary habits and life-style and physical activity. (**Table 1**)

Echocardiographic Assessment

A standardized echocardiography ultrasound examination, including M-mode, 2D, spectral, color flow and tissue Doppler will be performed by a certified technical professional using Philips CX-50 scanner with a standard 2.5-3.5-MHz phased-array probe. Image acquisition will be performed using a preprogrammed acquisition

protocol, following American and European Society of Echocardiography recommendations,[22-24] which guided sonographer through each protocol required view as outlined in **Table 2**. All studies will be acquired and stored digitally on a local PACS and transferred from field primary care centers to a secure server at the Salamanca University Hospital on the same day via a dedicated VPN connection. Development of the imaging and analysis protocol, field center echocardiography manual of operations, reading center manual of operations, field center sonographer, training of sonographer occurred from July 2015 to October 2015, followed by the initiation of the SALMANTICOR visit in November 2015, which is expected to continue until May 2018.

For patients in sinus rhythm, >3 full cardiac cycles will be recorded for each view, with recording beginning once the view is optimized. For subjects in atrial fibrillation, >5-second acquisitions per view will be recorded. Sonographers are instructed to continuously optimize both imaging depth and sector width to maintain a frame rate of 50 to 80 frames per second. Sonographers are also instructed to adjust 2D gain and compression, when necessary, to optimally demonstrate left ventricle endocardial borders. The color Doppler Nyquist limit will be set at 64 cm/s. Color Doppler gain will be set just below the level at which random background noise will be seen. Sonographers will optimally align spectral Doppler parallel to the direction of the blood flow of interest. Sonographers will optimize the baseline shift and velocity range so that the spectral envelope will occupy approximately three fourths of the display. All spectral Doppler acquisitions will be performed with a sweep speed between 75 to 100 cm/s, and a sample volume length of 3 mm for pulsed-wave Doppler. The tissue Doppler sample volume will be placed at the level of annulus (mitral and tricuspid) and

the baseline shift and velocity range optimized. All tissue Doppler acquisitions will be performed with similar acquisitions of spectral Doppler with a filter setting of 100 Hz.

Echocardiograms will be obtained at the subject´s primary care referral center and sonographers will not perform any measurements on the images obtained because all measurements will be analyzed and interpreted centrally at University Hospital of Salamanca. All SALMANTICOR echocardiograms will be read by a certified cardiologist and over-read by a board-certified cardiologist with expertise in echocardiography (Dr. Barreiro-Pérez) assessing **Table 3** variables. Over-reads of echocardiograms will be performed to confirm the accuracy of key quantitative measurements and to identify clinically important findings. Inter and intra-reader reproducibility was assessed before initiating the trial. For inter-reader productibility, intra-class correlation values ranged from 0.85 to 0.99 with left atrial volume and LV end-diastolic volumes having the highest intra-class correlation values (0.97-0.99). Intra-class correlation values were slightly better from intra-reader assessments for all measures.

Vascular Function Assessment

Cardio-ankle vascular index (CAVI), brachial ankle pulse wave velocity (ba-PWV) and ankle-brachial index (ABI) will be estimated using the VaSera VS-1500® device (Fukuda Denshi) as described by our group.[25] The ba-PWV will be calculated, as well as CAVI, which gives a more accurate estimation of the atherosclerosis degree. CAVI integrates cardiovascular elasticity derived from the aorta to the ankle pulse velocity through an oscillometric method; it is used as a good measure of vascular stiffness and does not depend on blood pressure.[26] CAVI values will be automatically calculated by substituting the stiffness parameters in the following equation to detect the vascular elasticity and the ba-PWV: stiffness parameter $\beta=2p \times 1/ (Ps-Pd) \times In (Ps/Pd) \times$ ba-

PWV^2, where p is the blood density, Ps and Pd are systolic blood pressure and diastolic blood preassure in mm Hg, respectively; and ba-PWV is measured between the aortic valve and ankle. The average coefficient of the variation of CAVI is <5%, which is small enough for clinical use and confirms that CAVI has favorable reproducibility.[27 28] CAVI and ABI will be measured in the resting position. ba-PWV is estimated using the following equation: ba-PWV=(0.5934 x height [cm] + 14.4724) / tba, where tba is the time the same waves were transmitted to the ankle. For the study, the lowest ABI and the highest CAVI and ba-PWV obtained will be considered. CAVI is classified as normal (CAVI<8), borderline (8≤CAVI<9) and abnormal (CAVI≥9). Abnormal CAVI represents subclinical atherosclerosis, and ba-PWV ≥17.5 is considered abnormal.[29 30] ABI ≤ 0.9 was considered abnormal.

Electrocardiographic examination

Electrocardiographic examination will be performed using a General Electric MAC 3500 ECG System (Niskayuna, New York, USA), which automatically measures wave voltage and duration. ECG will be performed by the same nurse trained to carefully standardized procedures for ECG acquisition. The standard 12-lead ECGs will be obtained at a paper speed of 25 mm/sec, amplitude of 10 mm/1mV, and a filter range 0.04 to 40 Hz from all patients. ECG tracing will be interpreted in a similar way to the echocardiographic protocol by independent cardiologist and over-read by a board-certified cardiologist with expertise in electrocardiography (Dr. Jesús Hernández) at the University Hospital of Salamanca. ECG measurements and interpretations will be done following standard methods,[31 32] (**Table 4**).

Laboratory test

Venous blood sampling will be performed at the end of the examination after participants have fasted and abstained from smoking and consumption of alcohol and

caffeinated beverages for 12 hours, following the protocol used in our hospital for other multidisciplinary projects.[25] A total of 20 mL of venous blood will be drawn for research testing. Blood will be drawn as follows: EDTA 10 mL and serum 10 mL. Aliquots of plasma (3 x 2 mL), serum (4 x 2 mL) and white cell pellet (3 x 2 mL) will be stored in freezers (-80ºC) until analysis. All biomaterial (serum, plasma and white blood cells) will be stored in the IBSAL biobank. Referral for biobanking is carried out through a specific electronic database. Biochemical tests include NT-proBNP, troponin, haemoglobin, blood cell count, thrombocytes, ferritin and iron, transferrin and iron saturation, potassium, sodium and creatinine, glycated haemoglobin, plasma glucose, aspartate aminotransferase, alanine aminostrasferase, total cholesterol, triglycerides, HDL and LDL, uric acid, high-sensitive C-reactive protein, thyroid-stimulating hormone. Further, biomarkers indicative of different pathophysiological mechanisms relevant to heart disease analyzed. White cell pellet will be used for genotyping.

Results and Outcomes

After the clinical history is performed and the echocardiogram and electrocardiogram interpreted, a clinical report is sent to the patient and to the primary care medical doctor. Individuals needing a further evaluation will be sent to the Cardiology Department through a preference standardized protocol.

Individuals will be contacted at 5-years intervals to ascertain the clinical status and to repeat the described basal evaluations. Clinical outcomes will include cardiovascular MACE, commencing dialysis and first hospitalization.

Statistical Analysis

Casual and multivariate inference

Data input will be stored in a database designed for the project. Normal distribution of variables will be verified using the Kolmogorov-Smirnov test. Quantitative variables

will be displayed as mean ± standard deviation if normally distributed or as the median (interquartile range) if asymmetrically distributed and qualitative variables will be expresses as frequencies. Analysis of difference of means between variables of two categories will be carried out using a Student´s t test or a Mann-Whitney U test, as appropriate, while qualitative variables will be analyzed using a χ^2 test. To analyze the relationship between qualitative variables of more than two categories and quantitative variables, an analysis of variance and the least significant difference test will be used in the post-hoc tests. The relationship of quantitative variables to each other will be tested using Pearsons or Spearmans correlation as appropriate. ANCOVA (covariance analysis) will be performed to adjust for the variables that can affect the results as confounders. A multivariate analysis of variance (MANOVA) will be performed in cases with more than one dependent variable to identify whether changes in the independent variables have significant effects on the dependent variables. The association between the variables studied will be performed by multiple linear regression. Data will be analyzed using the SPSS version 23.0 statistical package (SPSS Inc., Chicago, Illinois, USA). A value of $p < 0.05$ will be considered statistically significant.

Spatial analysis

In addition, this research aims for having a spatial understanding of the structural heart disease abnormalities in the province of Salamanca. Such demanding task will be carried out by applying different statistic procedures as Multiple Factor Analysis (MFA) and Cokriging.

MFA is an extension of Principal Component Analysis (PCA) tailored to handle distinct variables (quantitative, categorical or frequency) and different data tables collected on the same observations.[33] MFA is put into practice depending on the data

tables and the variables types: in the case of quantitative variables a PCA is applied; Multiple Correspondence Analysis (MCA) is applied in case of categorical variables [34]; and Correspondence Analysis (CA) for frequency variables.[35] Cokriging is a multivariate geostatistical procedure used for interpolation purposes.[36] This method is a generalization of a multivariate linear-weighted regression model, which weights depend on distance, direction and orientation of the neighboring data to the unsampled location.

In the SALMANTICOR study, we will further combine MFA and Cokriging. In our case, we have two different levels of observations, participants and municipalities. As a mathematical comparison, municipalities contain participants, therefore if we want to extend our investigation to a spatial analysis we need to utilize the resulting MFA projections over their corresponding municipality areas and then apply a Cokriging analysis over the unsampled municipalities (**Figure 2**) (**supplementary data**). This combination will provide a spatial understanding of the Salamanca population and will cover the whole analysis, however if we want to focus on a specific questionnaire we could skip the MFA and just get the results obtained from the MCA, PCA or CA and then apply a Cokriging analysis. In addition, if we require analyzing a particular item from a questionnaire we could also perform the analysis. In summary, we have a versatile methodology that permit to study as concrete aspects as wider analysis of our study.

The R packages FactoMineR and Gstat would be used in order to apply MFA and Cokriging, respectively.[37] [38] Additional code would be shared in a public Github repository.

Machine learning

The SALMANTICOR study will also be analyzed following the ML pipeline represented in **Figure 3**. ML first step will consist in the development of scalable methods for ML optimization with the aim to develop a first approach to the predictive structural heart disease model. Our ML model will start from ingesting raw data, leveraging data processing techniques to wrangle and, process and engineer meaningful features and attributes from this data (feature engineering). The derived features are attributes or properties shared by all the independent units on which analysis or prediction is to be done. In our case, clinical variables, variables quantified from imaging data and, deep learning image segmentation data will be chosen. Features will be combined with scalable ML algorithms, including deep learning process and automatic extraction of data functionalities, in order to develop the model (fit model). The model´s basic behavior and functionalities will be tested to develop a robust and reliable model (training model). We will validate, train and improve the ML model in a trial an error process until satisfactory model performance (validation). The SALMANTICOR study sample will be randomly divided into training (70% of sample) and validation (30% of sample), following previous published ML models.[39] We will build our predictor models using: random forest, gradient boosting, logistic regression, K-nearest neighbors, support vector machine, linear discriminant analysis and, naive Bayesian network models (**supplementary data**).

For the realization of this ML models we will use free software (Python) and free open-source unified workbench such as Scikit-learn.[40]

Quality control

Different processes will be carried out to guarantee study data quality and thus maximize the validity and reliability of measurements of the results. To this effect, field work operation manuals have been prepared. These documents specify the adequate

procedure for performing each test. All of these actions will confirm adequate performance of each procedure. Monthly meetings will be held with the principal investigator of the study to analyze the entire process, and an annual report on study progress will be prepared.

Ethical Review Board and dissemination plan

The study has been approved by the clinical research ethics committee (CEIC) of the health area of Salamanca ('CEIC of Salamanca Health Area, 9/29/2014). Participants will be required to sign an informed consent form prior to inclusion in the study, in accordance with the declaration of Helsinki and the WHO standards for observational studies. The study has been registered in ClinicalTrials.gov with identifier NCT03429452. Participants will be informed of the objectives of the project and of the risks and benefits of the examinations made. None of the examinations pose life-threatening risks for the type of participants to be included in the study. The study includes the obtaining of biological samples (including genetics analysis); the study participants therefore will be informed in detail. The confidentiality of the recruited participants will be ensured at all times in accordance with the provisions of current legislation on personal data protection (15/1999 of December 13, LOPD), and the conditions contemplated by Act 14/2007 on biomedical research.

We will use a variety of methods to ensure that our work will achieve maximum visibility. Publication of our study protocol provides an important first step towards this direction. In this paper, we have sought to offer a comprehensive overview of relevant literature, while underlining current research gaps that necessitated the design and implementation of the SALMANTICOR study. Similarly, the study results, given their applicability and implications for the general population, will be disseminated in

research meetings and in at least ten articles published in scientific journals. Finally, population-based control groups are difficult to obtain, specially in case-control cardiovascular studies where structural heart disease has to be rolled out. The SALMANTICOR study will provide availability of normative reference values quantification for echocardiographic, electrocardiographic, biochemical, genetics, VASERA and other parameters. Thus, international cooperation sharing data and participating in Horizon 2020 programs with the SALMANTICOR population are contemplated.

Patient and public involvement

Patients´ representatives will have an increasingly present voice in the SALMANTICOR study. There is currently an only patient organization for heart disease in the province of Salamanca, "El Paciente Experto". This organization has provided counselling in the design of the study, will jointly interpret the results of the study with the investigators of SALMANTICOR, will help to disseminate them to society, and will be involved when establishing new policies for health improvement and education empowerment with the Administration to halt the epidemic of cardiovascular disease.

Participants in the study will be initially contacted by the investigators through a letter explaining the advantages and disadvantages of the SALMANTICOR study; the importance the study has for a regional health-care policy and, the strategy for disseminating its results. A clinical report will be sent to all participants and their primary care medical doctors immediately after the clinical history is performed and the echocardiogram and electrocardiogram interpreted. Finally, the global and most important observations from the SALMANTICOR study will be also sent by letter to all

participants and to all doctor, primary care and specialists, of the province of Salamanca through the Medical College of Salamanca and our health Administration.

Data statement

Our data will be accessed at the Institute of Research of the University Hospital of Salamanca. Furthermore, our dataset will be published in a public repository. Additional code for our spatial analysis would be shared in a public Github repository.

Discussion

A major strength of the SALMANTICOR study is the selection of a representative population-based cohort across primary care, with a probable significate number of structural heart disease cases in each age, sex and place of residence category to allow overall and subpopulation analyses. This population-based approach increases the generalizability of the finding compared with surveys that addressed cardiovascular risk factors but have never included an echocardiographic assessment.[11 14 41-44] Moreover, in view of the similarity of trends in cardiovascular disease and population ageing from Spain with other developed countries,[45] our findings are likely to be broadly applicable to them.

Echocardiography in the SALMANTICOR study is design to address 3 specific aims. The first is to characterize the abnormalities of cardiac structure and function in a community-based sample and to assess how these abnormalities vary by place of residence (rural or urban), by age and, by sex. The study uses standard and novel echocardiographic techniques to characterize 5 specific domains of cardiac structure. These data will be used to define the population distribution of these measures and to determine their relationship with cardiovascular risk factors, including hypertension,

diabetes mellitus, coronary disease, renal insufficiency, and prognostically relevant biomarkers such as N-terminal pro-brain natriuretic peptide and high-sensitivity troponin. The second aim is to investigate ventricular-arterial coupling in addition to the association of cardiac structure and function with arterial stiffness assessed by CAVI, ba-PWV and ABI. The third aim is to prospectively examine the extent to which these noninvasive measures associate with incidence of adverse cardiovascular outcomes and to determine the degree to which these associations also vary by age, sex and by place of residence (rural or urban). In accomplishing these objectives, this study is developing an echocardiographic imaging database that will facilitate future investigations to compare these echocardiographic measures both with studies previously performed in other Countries,[12 13] and to be used as a very well established control group. Furthermore, our study will provide availability of normative reference values quantification for electrocardiographic, biochemical, genetics, VASERA and other parameters.

Adequate public health and service delivery planning requires reliable information about contemporary population-level disease incidence. SACYL is the regional health-care government authority of Castilla y Leon providing 2,5 million people universal access to health services, which are closely integrated with other public services and policies as part of a holistic approach to improving population health. In this sense, our study data will be used to understand the cardiovascular health needs of our Community population and to improve people´s health and wellbeing, and how they can be developed. SALMANTICOR will be established as the global observatory on cardiovascular health research and development of SACYL, as we will include real-time data about the burden of cardiovascular disease, people´s social circumstances and living conditions, lifestyles and diet, economic factors, access to healthcare and other

services, as well as our genes, age and sex. As well as understating the overall picture of our population´s health, data will be disaggregated to identify inequalities for example by gender, sex, and urban or rural place of residence. This will support the prioritization of interventions depending on the needs of different groups and will require effective actions for the prediction and prevention of cardiovascular disease; from macro-policies down to individuals and families, empowering people to take control of their health. In this sense, two new medical technology research lines have been identified by the SALMANTICOR investigators: exploring the use of spatial methods and exploring modern computational methods developed in the field of ML.

The use of spatial methods in healthcare research enable disease distribution patterns to be identified and have become popular in the field of public health,[46-48] Cancer and other disease mortality atlases have shown us that many risk factors of a territorial nature, influence geographical patterns, making it possible to select disease indicators and so reveal their geographical structure.[49 50] However, the number of spatial analyses published in major epidemiology journals is still very low.[51] One of the reasons is that the application of spatial methods requires specific training and has resulted in their substitution with less optimal methods from healthcare research. Therefore, it is important to promote spatial methods, especially those simple to interpret in the field of population-based studies and which could be potentially used in combination with other computational methods to facilitate interpretation, prediction and healthcare policies. Cardiology spatial analysis have been developed mainly in optimization problems and prevalence prediction. As an example of optimization, travel time isochrones analysis have been deployed in different facilities in order to identify exposed areas and act accordingly.[52] Nevertheless, prevalence prediction are the most

common geostatistical techniques in healthcare and it's not an exception in cardiology.[53 54]

The incorporation of ML in medicine holds promise for substantially improve health-care delivery.[18-21] ML provides methods, techniques, and tools that can help solving diagnostic and prognostic problems in a variety of medical domains. Furthermore, ML offers new approaches to leveraging the growing volume of heterogeneous data, including imaging data, available for analyses. To date, ML has been used in two broad and highly interconnected areas: automation of tasks that might otherwise be performed by a human and generation of clinically important knowledge. However, it is argued that the successful implementation of ML methods can help the integration of computer-based systems in the healthcare environment providing opportunities to really improve the efficiency of medical care and to be used as a regional policy to stablish effective public health programs. In this sense, The SALMANTICOR study represents an excellent opportunity to explore ML algorithms for estimating and ranking the impact of environmental and classical risk factors in the development of structural heart disease in a population-based setting.

References

1. CDC, NCHS. Underlying Cause of Death 1999-2015 on CDC WONDER Online Database, released 2017. Data are from the Multiple Cause of Death Files, 1999-2015, as compiled from data provided by the 56 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed Dec. 6, 2017.

2. European Cardiovascular Disease Statistics 2017 on www.ehnheart.org/cvd:statistics.html, released 2017. Data are from the European Heart Network (EHN), a Brussels-based Alliance of heart foundations and likeminded non-governmental organisations throughout Europe, with member organisations in 25 countries. Accessed Dec. 6, 2017.

3. Mozaffarian D, Benjamin EJ, Go AS, et al. Heart disease and stroke statistics--2015 update: a report from the American Heart Association. *Circulation* 2015;131(4):e29-322. doi: 10.1161/CIR.0000000000000152

4. Pearson TA, Palaniappan LP, Artinian NT, et al. American Heart Association Guide for Improving Cardiovascular Health at the Community Level, 2013 update: a scientific statement for public health practitioners, healthcare providers, and health policy makers. *Circulation* 2013;127(16):1730-53. doi: 10.1161/CIR.0b013e31828f8a94

5. Gerber Y, Weston SA, Enriquez-Sarano M, et al. Contemporary Risk Stratification After Myocardial Infarction in the Community: Performance of Scores and Incremental Value of Soluble Suppression of Tumorigenicity-2. *J Am Heart Assoc* 2017;6(10) doi: 10.1161/JAHA.117.005958

6. Dondo TB, Hall M, Timmis AD, et al. Geographic variation in the treatment of non-ST-segment myocardial infarction in the English National Health Service: a cohort study. *BMJ open* 2016;6(7):e011600. doi: 10.1136/bmjopen-2016-011600

7. Zhang L, Desai NR, Li J, et al. National Quality Assessment of Early Clopidogrel Therapy in Chinese Patients With Acute Myocardial Infarction (AMI) in 2006 and 2011: Insights From the China Patient-Centered Evaluative Assessment of Cardiac Events (PEACE)-Retrospective AMI Study. *J Am Heart Assoc* 2015;4(7) doi: 10.1161/JAHA.115.001906

8. Regueiro A, Bosch J, Martin-Yuste V, et al. Cost-effectiveness of a European ST-segment elevation myocardial infarction network: results from the Catalan Codi Infart network. *BMJ open* 2015;5(12):e009148. doi: 10.1136/bmjopen-2015-009148

9. Conrad N, Judge A, Tran J, et al. Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals. *Lancet* 2017 doi: 10.1016/S0140-6736(17)32520-5

10. Dawber TR, Meadors GF, Moore FE, Jr. Epidemiological approaches to heart disease: the Framingham Study. *Am J Public Health Nations Health* 1951;41(3):279-81.

11. Teo K, Chow CK, Vaz M, et al. The Prospective Urban Rural Epidemiology (PURE) study: examining the impact of societal influences on chronic noncommunicable diseases in low-, middle-, and high-income countries. *Am Heart J* 2009;158(1):1-7 e1. doi: 10.1016/j.ahj.2009.04.019

12. Shah AM, Cheng S, Skali H, et al. Rationale and design of a multicenter echocardiographic study to assess the relationship between cardiac structure and function and heart failure risk in a biracial cohort of community-dwelling elderly persons: the Atherosclerosis Risk in Communities study. *Circulation Cardiovascular imaging* 2014;7(1):173-81. doi: 10.1161/CIRCIMAGING.113.000736 [published Online First: 2013/11/12]

13. Vasan RS, Xanthakis V, Lyass A, et al. Epidemiology of Left Ventricular Systolic Dysfunction and Heart Failure in the Framingham Study: An Echocardiographic Study Over 3 Decades. *JACC Cardiovasc Imaging* 2017 doi: 10.1016/j.jcmg.2017.08.007

14. Yusuf S, Hawken S, Ounpuu S, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet* 2004;364(9438):937-52. doi: 10.1016/S0140-6736(04)17018-9

15. O'Donnell MJ, Xavier D, Liu L, et al. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. *Lancet* 2010;376(9735):112-23. doi: 10.1016/S0140-6736(10)60834-3

16. Chambers J, Kabir S, Cajeat E. Detection of heart disease by open access echocardiography: a retrospective analysis of general practice referrals. *Br J Gen Pract* 2014;64(619):e105-11. doi: 10.3399/bjgp14X677167

17. englund EJ. A variance of statisticians. *Math Geol* 1990;22(4):417-55.

18. Deo RC. Machine Learning in Medicine. *Circulation* 2015;132(20):1920-30. doi: 10.1161/CIRCULATIONAHA.115.001593

19. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* 2016;375(13):1216-9. doi: 10.1056/NEJMp1606181

20. Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. *N Engl J Med* 2017;376(26):2507-09. doi: 10.1056/NEJMp1702071

21. Shameer K, Johnson KW, Glicksberg BS, et al. Machine learning in cardiovascular medicine: are we there yet? *Heart* 2018 doi: 10.1136/heartjnl-2017-311198

22. Lang RM, Badano LP, Mor-Avi V, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *Eur Heart J Cardiovasc Imaging* 2015;16(3):233-70. doi: 10.1093/ehjci/jev014

23. Marwick TH, Gillebert TC, Aurigemma G, et al. Recommendations on the use of echocardiography in adult hypertension: a report from the European Association of Cardiovascular Imaging (EACVI) and the American Society of Echocardiography (ASE)dagger. *Eur Heart J Cardiovasc Imaging* 2015;16(6):577-605. doi: 10.1093/ehjci/jev076

24. American College of Cardiology Foundation Appropriate Use Criteria Task F, American Society of E, American Heart A, et al. ACCF/ASE/AHA/ASNC/HFSA/HRS/SCAI/SCCM/SCCT/SCMR 2011 Appropriate Use Criteria for Echocardiography. A Report of the American College of Cardiology Foundation Appropriate Use Criteria Task Force, American Society of Echocardiography, American Heart Association, American Society of Nuclear Cardiology, Heart Failure Society of America, Heart Rhythm Society, Society for Cardiovascular Angiography and Interventions, Society of Critical Care Medicine, Society of Cardiovascular Computed Tomography, and Society for Cardiovascular Magnetic Resonance Endorsed by the American College of Chest Physicians. *J Am Coll Cardiol* 2011;57(9):1126-66. doi: 10.1016/j.jacc.2010.11.002

25. Gomez-Marcos MA, Martinez-Salgado C, Gonzalez-Sarmiento R, et al. Association between different risk factors and vascular accelerated ageing (EVA study): study protocol for a cross-sectional, descriptive observational study. *BMJ open* 2016;6(6):e011031. doi: 10.1136/bmjopen-2016-011031 [published Online First: 2016/06/09]

26. Takaki A, Ogawa H, Wakeyama T, et al. Cardio-ankle vascular index is a new noninvasive parameter of arterial stiffness. *Circulation journal : official journal of the Japanese Circulation Society* 2007;71(11):1710-4. [published Online First: 2007/10/30]

27. Shirai K, Hiruta N, Song M, et al. Cardio-ankle vascular index (CAVI) as a novel indicator of arterial stiffness: theory, evidence and perspectives. *Journal of atherosclerosis and thrombosis* 2011;18(11):924-38. [published Online First: 2011/06/02]

28. Shirai K. Analysis of vascular function using the cardio-ankle vascular index (CAVI). *Hypertension research : official journal of the Japanese Society of Hypertension* 2011;34(6):684-5. doi: 10.1038/hr.2011.40 [published Online First: 2011/06/07]

29. Hu H, Cui H, Han W, et al. A cutoff point for arterial stiffness using the cardio-ankle vascular index based on carotid arteriosclerosis. *Hypertension research : official journal of the Japanese Society of Hypertension* 2013;36(4):334-41. doi: 10.1038/hr.2012.192 [published Online First: 2013/01/18]

30. Kawai T, Ohishi M, Onishi M, et al. Cut-off value of brachial-ankle pulse wave velocity to predict cardiovascular disease in hypertensive patients: a cohort study.

Journal of atherosclerosis and thrombosis 2013;20(4):391-400. [published Online First: 2012/12/28]

31. Macfarlane PW, Katibi IA, Hamde ST, et al. Racial differences in the ECG--selected aspects. *J Electrocardiol* 2014;47(6):809-14. doi: 10.1016/j.jelectrocard.2014.08.003

32. Rijnbeek PR, van Herpen G, Bots ML, et al. Normal values of the electrocardiogram for ages 16-90 years. *J Electrocardiol* 2014;47(6):914-21. doi: 10.1016/j.jelectrocard.2014.07.022

33. Escofier B, Pages J. Multiple factor for analysis (ALMULT package). *Comput Stat Data Anal* 1994;18:121-40.

34. Guisado-Clavero M, Roso-Llorach A, Lopez-Jimenez T, et al. Multimorbidity patterns in the elderly: a prospective cohort study with cluster analysis. *BMC Geriatr* 2018;18(1):16. doi: 10.1186/s12877-018-0705-7

35. Benzecri JP. L´Analyse des Données. Volume II. L´Analyse des correspondances. *Paris Dunod* 1973

36. Wackermagel H. Multivariate Geostatistics: An Introduction with Applications. *New York, NY: Springer-Verlag* 2003

37. Le S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software* 1990;25(1):1-18.

38. Pebesma EJ. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences* 2004;30:683-91.

39. Frizzell JD, Liang L, Schulte PJ, et al. Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure: Comparison of Machine Learning and Other Statistical Approaches. *JAMA Cardiol* 2017;2(2):204-09. doi: 10.1001/jamacardio.2016.3956

40. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;12:2825-30.

41. Grau M, Elosua R, Cabrera de Leon A, et al. [Cardiovascular risk factors in Spain in the first decade of the 21st Century, a pooled analysis with individual data from 11 population-based studies: the DARIOS study]. *Rev Esp Cardiol* 2011;64(4):295-304. doi: 10.1016/j.recesp.2010.11.005

42. Masia R, Pena A, Marrugat J, et al. High prevalence of cardiovascular risk factors in Gerona, Spain, a province with low myocardial infarction incidence. REGICOR Investigators. *J Epidemiol Community Health* 1998;52(11):707-15.

43. Rigo Carratala F, Frontera Juan G, Llobera Canaves J, et al. [Prevalence of cardiovascular risk factors in the Balearic Islands (CORSAIB Study)]. *Rev Esp Cardiol* 2005;58(12):1411-9.

44. Felix-Redondo FJ, Fernandez-Berges D, Fernando Perez J, et al. [Prevalence, awareness, treatment and control of cardiovascular risk factors in the Extremadura population (Spain). HERMEX study]. *Aten Primaria* 2011;43(8):426-34. doi: 10.1016/j.aprim.2010.07.008

45. Roth GA, Johnson C, Abajobir A, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J Am Coll Cardiol* 2017;70(1):1-25. doi: 10.1016/j.jacc.2017.04.052

46. Elliott P, Wartenberg D. Spatial epidemiology: current approaches and future challenges. *Environ Health Perspect* 2004;112(9):998-1006.

47. Abellan JJ, Richardson S, Best N. Use of space-time models to investigate the stability of patterns of disease. *Environ Health Perspect* 2008;116(8):1111-9. doi: 10.1289/ehp.10814

48. Kontopantelis E, Stevens RJ, Helms PJ, et al. Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study. *BMJ open* 2018;8(2):e020738. doi: 10.1136/bmjopen-2017-020738

49. Ho NT, Thompson C, Nhan LNT, et al. Retrospective analysis assessing the spatial and temporal distribution of paediatric acute respiratory tract infections in Ho Chi Minh City, Vietnam. *BMJ open* 2018;8(1):e016349. doi: 10.1136/bmjopen-2017-016349

50. Lopez-Abente G, Aragones N, Perez-Gomez B, et al. Time trends in municipal distribution patterns of cancer mortality in Spain. *BMC Cancer* 2014;14:535. doi: 10.1186/1471-2407-14-535

51. Auchincloss AH, Gebreab SY, Mair C, et al. A review of spatial methods in epidemiology, 2000-2010. *Annu Rev Public Health* 2012;33:107-22. doi: 10.1146/annurev-publhealth-031811-124655

52. Collaborators GBDRF, Forouzanfar MH, Alexander L, et al. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2015;386(10010):2287-323. doi: 10.1016/S0140-6736(15)00128-2

53. Przybysz R, Bunch M. Exploring spatial patterns of sudden cardiac arrests in the city of Toronto using Poisson kriging and Hot Spot analyses. *PLoS One* 2017;12(7):e0180721. doi: 10.1371/journal.pone.0180721

54. Ogunniyi MO, Holt JB, Croft JB, et al. Geographic variations in heart failure hospitalizations among medicare beneficiaries in the Tennessee catchment area. *Am J Med Sci* 2012;343(1):71-7. doi: 10.1097/MAJ.0b013e318223bbd4

Author statement

Jose Ignacio Melero-Alegría: data acquisition, surveys completion, physical, electrocardiographic and VASERA examinations, design of the work, drafting the work and revising it critically, final approval of the version to be published; Manuel Cascón: data acquisition, surveys completion, conception and design of the work, drafting the work and revising it critically, final approval of the version to be published; Alfonso Romero; conception and design of the work, interpretation of data, drafting the work of revising it critically, primary care coordination, final approval of the version to be published; Pedro Pablo Vara: echocardiographic data acquisition, interpretation of data, final approval of the version to be published; Manuel Barreiro-Pérez: conception and design of the echocardiographic protocol, analysis and interpretation of echocardiographic data, drafting the work and revising it critically for important intellectual content, final approval of the version to be published; Victor Vicente-Palacios: conception and design of the spatial and machine learning analysis, analysis and interpretation of data, drafting the work and revising it critically for important intellectual content, final approval of the version to be published; Fernando Pérez-Escanilla: conception and design of the work, interpretation of data, primary care coordination, final approval of the version to be published; Jesús Hernández-Hernández: conception and design of the electrocardiographic protocol, analysis and interpretation of ECG data, drafting the work and revising it critically for important intellectual content, final approval of the version to be published; Beatriz Garde: conception and design of the lifestyle, Mediterranean and exercise surveys, analysis and interpretation of data, final approval of the version to be published; Sara Cascón: conception and design of the work, coordinator of 5 out of 35 primary care centers, acquisition of data, final approval of the version to be published; Ana Martín-García: analysis and interpretation of echocardiographic data, final approval of the version to be published; Elena Díaz- Peláez: analysis and interpretation of echocardiographic data, final approval of the version to be published; José María de Dios: conception and design of the work, coordinator of 5 out of 35 primary care centers, acquisition of data, final approval of the version to be published; Aitor Uribarri: conception and design of the work (surveys), analysis and interpretation of data, final approval of the version to be published; Javier Jiménez-Candil: conception and design of the work, analysis and interpretation of ECG data, final approval of the version to be published; Ignacio Cruz-González: conception and design of the work (surveys), analysis and interpretation of data, final approval of the version to be published; Baltasara Blazquez: conception and design of the work, coordinator of 5 out of 35 primary care centers, acquisition of data, final approval of the version to be published; José Manuel Hernández: conception and design of the work, coordinator of 5 out of 35 primary care centers, acquisition of data, final approval of the version to be published; Clara Sánchez Pablos: data acquisition, surveys completion, physical, electrocardiographic and VASERA examinations, final approval of the version to be published; Inmaculada Santolino: conception and design of the work, coordinator of 5 out of 35 primary care centers, acquisition of data, final approval of the version to be published; M. Concepción Ledesma: conception and design of the work, coordinator of 5 out of 35 primary care centers, acquisition of data, final approval of the version to be published; Paz Muriel: conception and design of the work, coordinator of 5 out of 35 primary care centers, acquisition of data, final approval of the version to be published;

P. Ignacio Dorado-Díaz: conception and design of the spatial and machine learning analysis, analysis and interpretation of data, drafting the work and revising it critically for important intellectual content, final approval of the version to be published; Pedro L Sánchez: conception and design of the study, interpretation of data, drafting the work, 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.

Tables

Table 1. Questionnaires.

Name of the questionnaire	Number of variables	Principal variables	Time of completion
Demographics & Cardiovascular risk factors	12	Sex, age, residence, smoking, alcohol consumption, hypertension, hypercholesterolemia, diabetes, previous heart disease, family history	5 minutes
Cardiovascular & non-cardiovascular history	23	Coronary heart disease, arrhythmias, valvulopathies, heart failure, cardiac healthcare visits in the past and where (public or private attention), stroke, vascular peripheral disease, bleeding history, chronic kidney disease, chronic lung disease, asthma, rheumatic disease, depressive disorder, dementia, anxiety, dependency	12 minutes
Physical examination	8	Body mass index, abdominal perimeter, heart rate, oxygen saturation, blood pressure, heart murmurs & sounds	8 minutes
Medication	24	Aspirin, clopidogrel, ticagrelor, prasugrel, warfarin, acenocumarol, dabigatran, ribaroxaban, apixaban, edoxaban, betabloquers, ACE inhibitors, RAAS antagonists, calcium channel blocker, diuretics, aldosterone inhibitors, statin, ezetimibe, fibrate, ivabradine, ranolazine, proton-pump inhibitor, NSAIDs, corticoids	10 minutes
Socio-economic status	13	Marital status, education, employment, annual income, homeownership, housing quality, medical coverage	8 minutes
Dietary habits & life-style	39	Number of meals, diet, beverage, salt, bread, olive-oil, coffee, chocolate and potatoes dietary counseling, Mediterranean diet adherence, number of sleeping hours, siesta practice, pet ownership	12 minutes
Physical activity	7	Number of days, number of hours, intensity	5 minutes
Total	**126**		**60 minutes**

Table 2. Echocardiographic imaging protocol required views.

Parasternal position	
Parasternal long axis	2D imaging (at deep depth) 2D imaging (at shallow depth) Color Doppler of the mitral and aortic valves
Parasternal short axis, aortic valve level	2D imaging of AV Color Doppler of AV 2D imaging of RVOT Color Doppler of RVOT PW and CW Doppler of RVOT
Parasternal short axis, mitral valve level	2D imaging
Parasternal short axis, left ventricle apex	2D imaging
Apical position	
Apical 4-chamber view	2D imaging 2D imaging, focused/zoomed of left ventricle 2D imaging, focused on left atrium Color Doppler of mitral valve/left atrium PW Doppler of mitral flow CW Doppler of mitral flow TDI of septal and lateral mitral annulus
Apical 4-chamber view, focused on the RV	2D imaging Color Doppler of tricuspid valve/right atrium CW Doppler of tricuspid regurgitation TDI of lateral tricuspid annulus
Apical 5-chamber view	2D imaging Color Doppler of LVOT PW of LVOT flow CW of transaortic flow
Apical 2-chamber view	2D imaging 2D imaging focused/zoomed on LV 2D imaging focused on left atrium Color Doppler mitral valve/left atrium
Apical 3-chamber view	2D imaging 2D imaging focused/zoomed on LV 2D imaging focused on left atrium Color Doppler mitral valve/left atrium Color Doppler of aortic valve PW of LVOT flow CW of transaortic flow
Subcostal view	
Inferior vena cava	2D imaging (5-s acquisition)

Table 3. Echocardiographic parameters.

Structure and function assessment	Number of variables	Principal variables Time of completion
Aorta & Atrias & ventricles	39	Ascending aorta (mm), LV diastolic dimension (mm), LV systolic dimension (mm), left ventricular mass index (g/m^2), left atrial volume index by biplanar Simpson method (mL/m^2), right ventricular diastolic dimension (mm), right atrial volume index (mL/m^2), biplanar Simpson left ventricular ejection fraction (%), mitral E-wave (cm/s), mitral A-wave (cm/s), mitral E/A, mitral deceleration time (cm/s), pulmonary artery systolic pressure (mm Hg), mitral E/e´septal annulus, mitral E/e´lateral annulus, mitral E/e´average of annulus
Valves	41	Aortic valve jet peak velocity (m/s), aortic mean gradient (mm Hg), aortic cups number, aortic valve calcification, aortic regurgitation presence and grade, mitral valve calcification, mitral mean gradient (mm Hg), mitral pressure half time (msec), mitral prolapse, mitral regurgitation presence and grade, tricuspid regurgitation presence and grade, pulmonary regurgitation presence and grade
Pericardium	3	Pericardial effusion presence and grade

Table 4. 12-lead ECG parameters.

Rhythm	Sinus rhythm Auricular tachycardia Atrial fibrillation Common atrial flutter Uncommon atrial flutter Nodal rhythm Atrial ectopies Ventricular ectopies Atrial paced rhythm Ventricular paced rhythm with sinusal activity Ventricular paced rhythm with atrial fibrillation Atrial and ventricular paced rhythm
Heart rate	
P wave	P duration Sinus P morphology Pulmonary P morphology Interatrial block
PQ time	
AV block	Not present First degree AV block Second degree AV block, Mobitz I Second degree AV block, Mobitz II 2:1 AV block Third degree or complete AV block
QRS duration	
QRS axis	
RR time	
QT time	
QT corrected time	
Brugada pattern	Not present Type I Type II Type III
AV block	Not present First degree AV block Second degree AV block, Mobitz I Second degree AV block, Mobitz II 2:1 AV block Third degree or complete AV block
Early repolarization pattern	Not present Inferior Lateral Inferior & lateral
Bundle branch configuration	Not present Complete left bundle branch block Complete right bundle branch block Incomplete left bundle branch block Incomplete right bundle branch block
Intraventricular conduction disturbances	
Fascicular block configuration	Not present

	Left anterior fascicular block Left posterior fascicular block
Notch QRS presence	
Left ventricular hypertrophy	
Delta waves presence	
Repolarization changes of digitalis	
Pathological Q-waves presence and position	
Significant ST elevation	
Significant ST depression	
Negative T-waves presence and position	

Figure legends

Figure 1. Province of Salamanca map and distribution of the total of 35 primary health centers: 18 in urban-considered municipalities (blue) and 17 in rural-considered municipalities (red). Municipalities of more than 5,000 individuals are considered as urban areas in the SALMANTICOR study.

Figure 2. Left panel represents the spatial analysis pipeline that SALMANTICOR will use for map plotting purposes. We will combine multiple factor analysis (MFA) and Cokriging. We will inquire and analyze participants from municipalities and questionnaires. Initially, for quantitative variables principal component analysis (PCA) is applied; for categorical variables, multiple correspondence analysis (MCA); and for frequency variables, correspondence analysis (CA). We will then ensemble the normalized data in a single table that is analyzed via PCA to describe the spatial behaviors of our samples within crossvariograms (crossvariog). We then will apply a linear model coregionalization (LMC) to finally interpolate the results over the different municipalities of the province of Salamanca using Cokriging. Maps in the right panel represent municipal spatial patterns examples of how we will represent municipal (Salamanca is divided into 362 municipalities) distribution of structural heart disease and dyslipidemia prevalence.

Figure 3. Machine learning (ML) pipeline for the SALMANTICOR study. The learning algorithm will take heterogeneous data that will be preprocessed to create input data for the ML algorithm. Furthermore, raw images will also be used in the ML algorithm using neural network modelling. The output of the ML algorithm will also need to be processed and improved until a satisfactory model is developed.

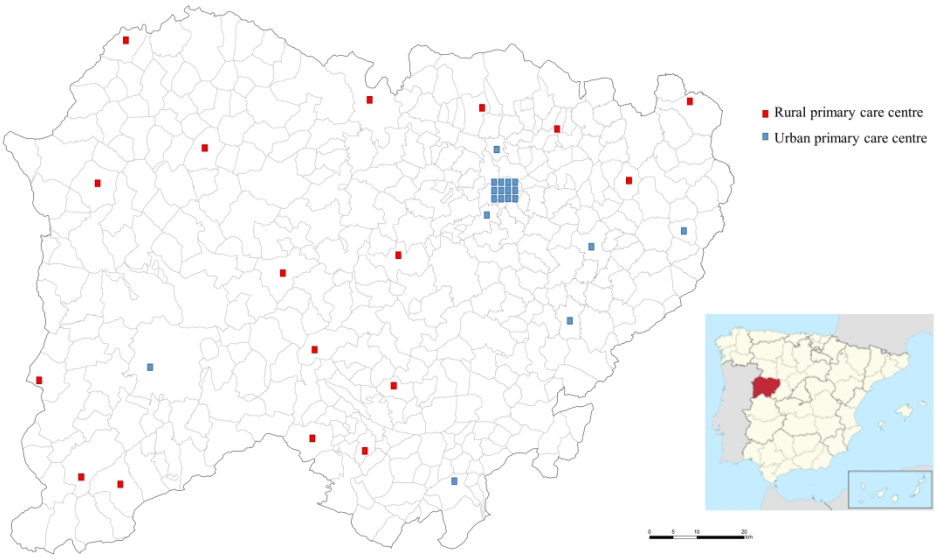

Figure 1

338x190mm (300 x 300 DPI)

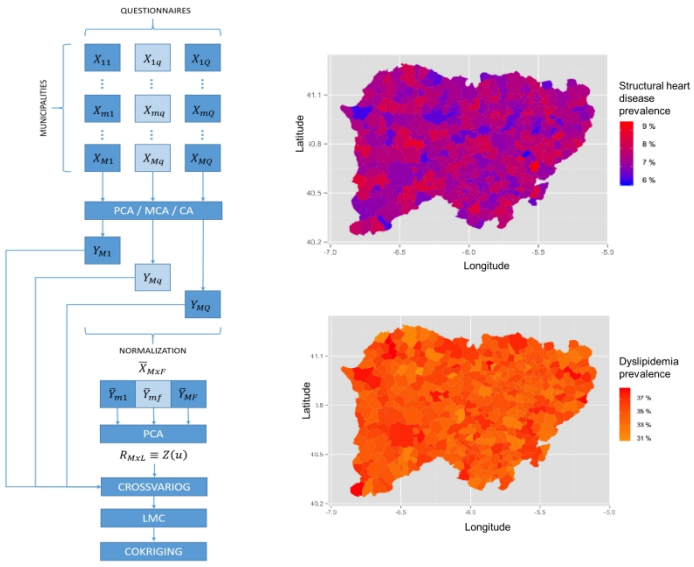

Figure 2

338x190mm (300 x 300 DPI)

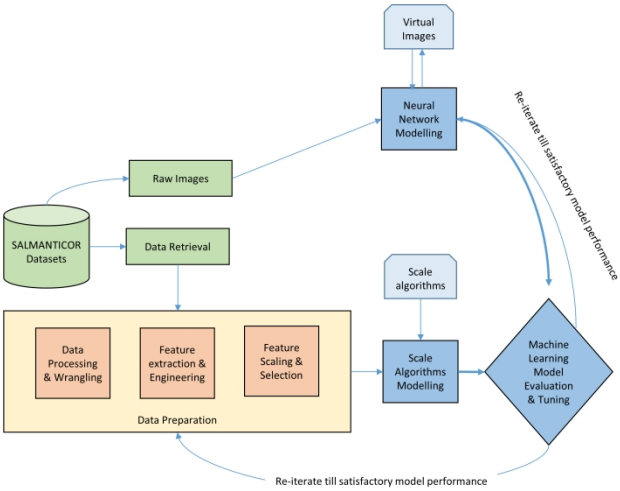

Figure 3

338x190mm (300 x 300 DPI)

Supplementary data of the SALMANTICOR study

Spatial analysis

We will combine multiple factor analysis (MFA) and Cokriging statistics procedures to provide a spatial analysis of the SALMANTICOR population.

Our study will inquire and analyzed N individuals from M municipalities. Q questionnaires were handed to all the participants. Let X_{nmq} be a matrix block where n is the number of participant of a m municipality and k is the correspondent questionnaire of our departing matrix D_{MxQ}.

Therefore, depending on the type of k questionnaire, we will employ a PCA, MCA or CA, to each block X_{nmq} obtaining $\overline{Y}_{mq} = \frac{1}{\lambda_{mq}} Y_{mq}$ where λ_{mq} is its first singular value.

Hence, we join all the resulting $\overline{Y}_{mq}$ forming a $\overline{X}_{MxF}$ matrix where M are the municipalities and F the resulting factors.

$$\overline{X}_{mf} = [\overline{Y}_{m1} | \overline{Y}_{m2} | \dots | \overline{Y}_{mf} | \dots | \overline{Y}_{mF}]$$

Finally, a generalized PCA is applied on $\overline{X}_{MxF}$

After performing MFA we will proceed to project the resulting coordinates that represents our municipalities over the resulting L latent variables obtaining R_{MxL}.

Adding the spatial coordinates u to each municipality we attain $Z(u) = [u|R]$. Once we get the $Z(u)$ matrix, we will apply a spatial interpolator such as Cokriging.

We will then describe the spatial behavior of our samples using variograms. Variograms are illustrations of how the semivariance acts in function of the distance. Semivariance is defined as half the expectation between two different values at two

locations (u and u + h), and is used in univariate analyses. To transfer our analysis to a multivariate problem we will need to build crossvariograms.

A crossvariogram γ_{ij} describes the degree of spatial dependence of our projected variables measuring the variation between two samples depending on the distance h (also known as lag) between them.

After this step, we will define

$$\Gamma(h) = \frac{1}{2}\left[\left(Z_i(u) - Z_i(u+h)\right) \cdot \left(Z_j(u) - Z_j(u+h)\right)\right]$$

with $i, j = 1 \ldots M$ and hence, the crossvariogram

Using a more practical approach, we will need to build a set of experimental crossvariograms based on our matrix $Z(u)$.

Therefore, we will obtaine $\frac{L(L+1)}{2}$ experimental semivariograms, and subsequently these direct and crossvariograms will need to be fitted. The different parts of a theoretical semivariogram are:

Nugget: It represents variability at small distances ($h \approx 0$).

Sill: The semivariance b value at which the semivariogram levels off.

Range: The a distance at which the semivariogram reaches the sill value.

The Linear Model of Coregionalization (LMC) permits all the $\frac{L(L+1)}{2}$ semivariograms to be fitted as linear combinations of S basic semivariogram functions (Gaussian, Exponential, Spherical, etc). The LMC can be expressed as a multivariate nested semivariogram model.

$$\Gamma(h) = \sum_{s=1}^{S} B_s g_s(h)$$

where $\Gamma(h)$ is the S×S matrix of semivariogram values at lag h, and B_s is the S×S matrix of sills of the basic semivariogram function $g_s(h)$. B_s has to be positive semidefinite, to assure that the variance-covariance matrix is also positive.

Once $\Gamma(h)$ is set, we will need to interpolate over the different polygons that represents the municipalities and shape the province of Salamanca. For fulfilling this task, we will apply Cokriging.

Cokriging is the multivariate extension of kriging, whose main purpose is to compute a weighted average of the sample values in close proximity to a grid point, polygon or volume. It searches for the best linear unbiased estimator, based on assumptions on covariances. There are different procedures such as ordinary, universal, or simple Cokriging.

As an example, we present simple Cokriging.

$$\bar{Z}_{i_0}(u_0) = m_{i_0} + \sum_{i=1}^{L}\sum_{\alpha=1}^{M} w_\alpha^i (Z_i(u_\alpha) - m_i)$$

where u_0 is an unsampled municipality and u_α a sample location, w_α^i is the weight and m corresponds to the means of our variables. We can associate a simple cokriging system to this estimator as $C_{ij}w_i = c_{ii_0}$, where C_{ij} is the M×M covariance matrix, and c_{ii_0} is the M_0×M covariance matrix between the unsampled and sample locations.

Machine learning

The following table describes the selected machine learning (ML) algorithms to be used in the SALMANTICOR study.

Algorithm	Type	Description
Random Forest	Combine methods	Classification ensemble through a combination set of non-correlated independently decision trees
Gradient Boosting	Combine methods	Ensemble technique in which decision trees are not independently, but sequentially

Algorithm	Type	Description
Logistic regression	Regression	The go-to method for categorical or binary classification
K-nearest Neighbors	Supervised classification	Classifies each unlabeled example by the majority label among its k-nearest neighbors in the training set
Support Vector Machine	Supervised classification	Classification and regression technique through construction of separating hyperplanes to maximize the margin and to produce a generalization ability
Linear discriminant analysis	Linear discriminant	Searches for directions in the data that have the largest variance and subsequently project the data onto it combining Fisher vectors
Naive Bayes classifier	Probabilistic supervised classification	The Bayesian classification is used as a probabilistic learning method

STROBE statement SALMANTICOR

STROBE Statement—checklist of items that should be included in reports of observational studies

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract: Population-based study
		(b) Provide in the abstract an informative and balanced summary of what was done and what was found: A cross-sectional survey of randomly selected residents of Salamanca (Spain). 2400 individuals, stratifies by age and sex and by place of residence (rural and urban) will be studied. The variables to analyze will be obtained from the clinical history, different surveys including social status, Mediterranean diet, functional capacity, electrocardiogram, echocardiogram, VASERA and biochemical and genetic analysis.
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported: pages 8-9
Objectives	3	State specific objectives, including any prespecified hypotheses: page 10
Methods		
Study design	4	Present key elements of study design early in the paper: The SALMANTICOR study is a cross-sectional descriptive population-based study of the prevalence of structural heart disease and their risk factors that will enroll a total of 2400 individuals, stratifies by age, sex and by place of residence (rural and urban), in a Spanish community: Salamanca
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection: pages 11-17

Participants	6	(*a*) *Cohort study*—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study*—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants: Individuals aged ≥18 years included in the lists of all primary healthcare facilities of the province of Salamanca represented the reference population
		(*b*) *Cohort study*—For matched studies, give matching criteria and number of exposed and unexposed *Case-control study*—For matched studies, give matching criteria and the number of controls per case
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable: The SALMANTICOR study is designed to provide echocardiographic parameters characterizing cardiac structure and function in all individuals. SALMANTICOR participants will undergo surveillance for cardiovascular events, including heart failure, incident coronary heart disease, and all-cause mortality.
Data sources/ measurement	8 *	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group: pages 11-16 and tables
Bias	9	Describe any efforts to address potential sources of bias: Spain´s and consequently Salamanca healthcare system is public, guaranteeing universal coverage. In total, 98.7 percent of the population are insured for this public Spanish healthcare system. In Salamanca, a total of 35 primary health centers throughout the province provide healthcare services to the overall population: 18 to the urban-considered municipalities and 17 to

		the rural-considered municipalities (Figure 1). Individuals aged ≥18 years included in the lists of all primary healthcare facilities of the province of Salamanca represented the reference population of 295,975 subjects: mean age 52.9±19.8 years; 52.4% females; 61.3% residing in urban areas
Study size	10	Explain how the study size was arrived at: A sample size of 2400 subjects is calculated based on an expected prevalence of structural heart disease of 6% with a confidence interval of 95% and a 1% precision. In order to obtain the necessary sample size, 35% more requests for participation will be made, estimating errors of location from the healthcare database or refuses to participate in the study. Thus, 3564 people will be randomly selected from the primary care lists.
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why: pages 16-17
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding: pages 16-19
		(*b*) Describe any methods used to examine subgroups and interactions: pages 16-19
		(*c*) Explain how missing data were addressed: pages 16-19
		(*d*) *Cohort study*—If applicable, explain how loss to follow-up was addressed *Case-control study*—If applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy: pages 16-19
		(*e*) Describe any sensitivity analyses

Continued on next page

Results			
Participants	13*		(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed
			(b) Give reasons for non-participation at each stage
			(c) Consider use of a flow diagram
Descriptive data	14*		(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders
			(b) Indicate number of participants with missing data for each variable of interest
			(c) *Cohort study*—Summarise follow-up time (eg, average and total amount)
Outcome data	15*		*Cohort study*—Report numbers of outcome events or summary measures over time
			Case-control study—Report numbers in each exposure category, or summary measures of exposure
			Cross-sectional study—Report numbers of outcome events or summary measures
Main results	16		(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included
			(*b*) Report category boundaries when continuous variables were categorized
			(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses	17		Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses
Discussion			
Key results	18		Summarise key results with reference to study objectives: pages 20-24
Limitations	19		Discuss limitations of the study, taking into account sources of potential

	9	bias or imprecision. Discuss both direction and magnitude of any potential bias. pages 16-19
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence. pages 16-19
Generalisability	21	Discuss the generalisability (external validity) of the study results. pages 16-19
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based. This study was supported by ~~by~~ national (PI14/00695, Institute of Health Carlos III, Spanish Ministry of Economy and Competitiveness) and community (GRS1030/A/14, SACYL, Junta Castilla y León) competitive grants and by the Spanish Cardiovascular Network (RIC and CIBERCV) from the Institute of Health Carlos III, Spanish Ministry of Economy and Competitiveness, Obra Social "la Caixa" and Philips Ibérica Healthcare division.

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

BMJ Open

The Salmanticor Study. Rationale and Design of a Population-based Study to Identify Structural Heart Disease Abnormalities: a Spatial and Machine Learning Analysis

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-024605.R1
Article Type:	Protocol
Date Submitted by the Author:	06-Oct-2018
Complete List of Authors:	Melero-Alegria, Jose Ignacio; Hospital Universitario de Salamanca-IBSAL, Cardiology Cascon, Manuel; Hospital Universitario de Salamanca-IBSAL, Cardiology Romero, Alfonso; Miguel Armijo Primary Care Center Vara, Pedro P; Hospital Universitario de Salamanca-IBSAL, Cardiology Barreiro-Perez, Manuel; Hospital Universitario de Salamanca-IBSAL, Cardiology Vicente-Palacios, Victor; Hospital Universitario de Salamanca-IBSAL, Cardiology Perez-Escanilla, Fernando; San Juan Primary Care Center Hernandez-Hernandez, Jesus; Hospital Universitario de Salamanca-IBSAL Garde, Beatriz; Hospital Universitario de Salamanca-IBSAL, Cardiology Cascon, Sara; Robleda Primary Care Center Martin-Garcia, Ana; Hospital Universitario de Salamanca-IBSAL, Cardiology Díaz-Pelaez, Elena; Hospital Universitario de Salamanca-IBSAL, Cardiology de Dios, Jose Maria; Salamanca Primary Care Center Management Uribarri, Aitor; Hospital Universitario de Salamanca-IBSAL Jimenez-Candil, Javier; Hospital Universitario de Salamanca-IBSAL Cruz-Gonzalez, Ignacio; Hospital Universitario de Salamanca-IBSAL, Cardiology Blazquez, Baltasara; Miranda del Castañar Primary Care Center Hernandez, Jose Manuel; Miranda del Castañar Primary Care Center Sanchez-Pablos, Clara; Hospital Universitario de Salamanca-IBSAL Santolino, Inmaculada; Santa Marta Primary Care Center Ledesma, Maria Concepcion; Peñaranda de Bracamonte Primary Care Center Muriel, Paz; Miguel Armijo Primary Care Center Dorado-Diaz, P. Ignacio; Hospital Universitario de Salamanca-IBSAL Sanchez, Pedro L; University of Salamanca, Cardiology
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Cardiovascular medicine, Health policy, Health informatics
Keywords:	structural heart disease, population, rural, urban, spatial analysis, machine learning

THE SALMANTICOR STUDY. RATIONALE AND DESIGN OF A POPULATION-BASED STUDY TO IDENTIFY STRUCTURAL HEART DISEASE ABNORMALITIES: A SPATIAL AND MACHINE LEARNING ANALYSIS

José Ignacio Melero-Alegría, RN[1]
Manuel Cascón, MD, PhD[1]
Alfonso Romero, MD[2]
Pedro Pablo Vara, DCS[1]
Manuel Barreiro-Pérez, MD, PhD[1]
Victor Vicente-Palacios, PhD[1]
Fernando Pérez-Escanilla, MD, PhD[3]
Jesús Hernández-Hernández, MD[1]
Beatriz Garde, BPharm[1]
Sara Cascón, MD, PhD[4]
Ana Martín-García, MD, PhD[1]
Elena Díaz-Peláez, MD[1]
José María de Dios, MD[5]
Aitor Uribarri, MD, PhD[1]
Javier Jiménez-Candil, MD, PhD[1]
Ignacio Cruz-González, MD, PhD[1]
Baltasara Blazquez, MD[6]
José Manuel Hernández, MD[6]
Clara Sánchez-Pablos, RN[1]
Inmaculada Santolino, MD[7]
María Concepción Ledesma, MD[8]
Paz Muriel, MD[2]
P. Ignacio Dorado-Díaz, MD[1]
Pedro L Sanchez, MD, PhD[1§]

From the
[1]Department of Cardiology, Hospital Universitario de Salamanca, Instituto de Investigación Biomédica de Salamanca (IBSAL), Facultad de Medicina, Universidad de Salamanca, and CIBERCV, Salamanca, Spain
[2]Miguel Armijo Primary Care Centre, Salamanca, Spain
[3]San Juan Primary Care Centre, Spain
[4]Robleda Primary Care Centre, Salamanca, Spain
[5]Salamanca Primary Care Centre Management, Salamanca, Spain
[6]Miranda del Castañar Primary Care Centre, Salamanca, Spain
[7]Santa Marta Primary Care Centre, Salamanca, Spain
[8]Peñaranda de Bracamonte Primary Care Centre, Salamanca, Spain

BRIEF TITLE: The SALMANTICOR Study

Address for Correspondence[§]:

Pedro L Sanchez, MD, PhD.
Cardiology Department. Hospital Universitario de Salamanca-IBSAL.
Paseo de San Vicente 58-187. 37007 Salamanca. SPAIN.
Telephone: 34-923291100 (ext 55356).
e-mail: pedrolsanchez@secardiologia.es

Financial Support:
This study was supported by national (PI14/00695, Institute of Health Carlos III, Spanish Ministry of Economy and Competitiveness) and community (GRS1030/A/14, SACYL, Junta Castilla y León) competitive grants and by the Spanish Cardiovascular Network (RIC and CIBERCV) from the Institute of Health Carlos III, Spanish Ministry of Economy and Competitiveness, Obra Social "la Caixa" and Philips Ibérica Healthcare division.

Acknowledgements:
We thank all primary care physicians and personnel helping with the development of the study. We thank Philips Iberica and Obra Social "La Caixa" for their support. We especially thank participants in the study and apologise for any inconvenience we could have caused. We thank the involvement of the Salamanca patient organisation "El paciente experto", for providing counselling to SALMANTICOR and for further promoting the dissemination of the results to the society and to the regional government.

Potential Conflicts of Interest: None to disclose.

Word Count: 4715 (excluding references)

Abstract

Introduction. This study aims to obtain data on the prevalence and incidence of structural heart disease in a population setting and, to analyse and present those data on the application of spatial and machine learning methods that, although known to geography and statistics, need to become used for healthcare research and for political commitment to obtain resources and support effective public health program implementation.

Methods and analysis. We will perform a cross-sectional survey of randomly selected residents of Salamanca (Spain). 2400 individuals, stratified by age and sex and by place of residence (rural and urban) will be studied. The variables to analyse will be obtained from the clinical history, different surveys including social status, Mediterranean diet, functional capacity, electrocardiogram, echocardiogram, VASERA and biochemical as well as genetic analysis.

Ethics and dissemination. The study has been approved by the ethical committee of the health care community. All study participants will sign an informed consent for participation in the study. The results of this study will allow the understanding of the relationship between the different influencing factors and their relative importance weights in the development of structural heart disease. For the first time, a detailed cardiovascular map showing the spatial distribution and a predictive machine learning system of different structural heart diseases and associated risk factors will be created and will be used as a regional policy to establish effective public health programs to fight heart disease. At least ten publications in the first-quartile scientific journals are planned.

Trial registration number. NCT03429452.

Abstract word count: 250

Strengths and limitations

- To obtain data on the prevalence and incidence of structural heart disease in the setting of a population-based study enrolling a total of 2400 individuals, stratified by age, sex and by place of residence (rural and urban), in a Spanish community.
- To create a population-based established control group providing availability of normative reference values quantification for echocardiographic, electrocardiographic, VASERA, biochemical and genetic parameters.
- To show the spatial distribution of the different patterns of structural heart disease through the spectrum of age and sex and between urban and rural residences.
- To develop a predictive model of structural heart disease using cardiovascular heterogeneous data (including images and machine learning techniques)
- To establish the study as the global observatory on cardiovascular health research and development of the regional healthcare government to support effective public health program implementation.

Keywords (MeSH terms)

Structural heart disease · population · rural · urban · spatial analysis · Multiple factor analysis · Principal component analysis · multivariate statistics · Cokriging · geostatistics · machine learning

Abbreviations

ABI	ankle-brachial index
ACE	angiotensin-converting enzyme
ba-PWV	brachial ankle pulse wave velocity
CA	correspondence analysis
CAVI	cardio-ankle vascular index
CEIC	clinical research ethics committee
ECG	electrocardiogram
GP	Gaussian process
MCA	multiple correspondence analysis
MFA	multiple factor analysis
ML	machine learning
NSAIDs	nonsteroidal anti-inflammatory drugs
PACS	picture archiving and communication system
PCA	principal component analysis
RAAS	renin-angiotensin-aldosterone system
VNP	virtual private network
2D	two dimensional

Introduction

Each year heart diseases cause almost 4 million deaths in Europe and the United States; that is one out of four deaths.[1 2] Although number of deaths from heart disease has decreased, the burden of heart disease is increasing. In 2015, more than 85 million people in Europe were living with cardiovascular disease.[2] The increase in the prevalence of classical cardiovascular risk factors, dietary factors, physical activity and probably other social factors make the largest contribution to the risk of heart disease. Overall cardiovascular disease health care costs in the European Union and the United States have increased rapidly over the last ten years; currently surpassing 200 billion euro a year.[2 3]

In this sense, public health delivery planning requires reliable information about contemporary population-level disease prevalence and incidence. Furthermore, community healthcare systems should obtain and provide their own data before implementing any effective health program as these regional systems are highly influenced by geographic diversity, the availability of resources and infrastructure, and the characteristics of healthcare systems and patterns of reimbursement.[4] This is well illustrated by the attention of myocardial infarction where the exchange of accurate and timely information between the health care community, decision makers, and the public program effects, has been essential.[5-8]

Policies need to consider both standardized rates, which describe disease prevalence and incidence independently of changes in population, and absolute numbers of patients affected, which describe the impact of the disease on the population, political commitment, resources and services of interest.[4,9] Limited data exist on estimation of heart disease prevalence in a population setting. Previous studies have frequently been

based on selected cohorts, which may not represent the general population.[10-13] Other studies have restricted case identification to those made in general practice consultations or hospital admissions.[14-16] However, it is only by considering presentations across the whole spectrum of structural heart disease that the full burden of the disease can be captured and an accurate distinction can be made between incident and prevalent cases. Thus, contemporary population-based studies of heart disease prevalence and incidence are needed to inform resource planning and research priorisation but current evidence is scarce.

Spatial analysis is a great tool to investigate population behaviour, relations and consequently determine future action plans or policies. Spatial methods are varied, ranging from descriptive spatial analysis to complex interpolation algorithms. Gaussian Process (GP) procedures, such as cokriging, have distinct advantages over conventional spatial prediction techniques.[17] They allow researchers to include measured spatial variability in the geostatistical estimation process and they smooth predicted values based on the proportion of total sample variability accounted by random noise. Furthermore, GP helps mitigate the effect of variable sample density caused by hot spots (some zones are usually oversampled). Hence, geostatistic techniques are suitable methods to apply on population studies.

Furthermore, the volume of quantitative and imaging data, generated by population studies, will also be a key driver in the future for research and how we provide care. In this sense, machine learning (ML) to train algorithms to recognize cardiac damage on a better level, avoiding diagnostic errors and improving the early identification of the disease offers new approaches to leveraging the increasing volume of data available for analyses[18-21]. Thus, we are convinced that ML can play a key role in population-based epidemiological studies when trying to recognise patients-disease vulnerability earlier.

The objectives of this study are: to obtain data on the prevalence and incidence of structural heart disease in a population setting; to show the spatial distribution of the different patterns of structural heart disease through the spectrum of age and sex and between urban and rural; to develop a predictive model of structural heart disease using cardiovascular heterogeneous data (including images and ML techniques); to generate new hypotheses which might contribute to healthcare research and to political commitment to obtain resources and support effective public health program implementation.

In this article we describe the design, data and imaging acquisition, analysis methods and quality assurance metrics for the SALMANTICOR study.

Methods

Study Design and Participants

The SALMANTICOR study is a cross-sectional descriptive population-based study of the prevalence of structural heart disease and their risk factors that will enrol a total of 2400 individuals, stratified by age, sex and by place of residence (rural and urban), in a Spanish community: Salamanca. Structural heart disease refers to any of the following heart abnormalities including congenital heart disease, cardiomyopathies, valvar heart disease, ischemic heart disease, pericardial diseases and rhythm or conduction disorders.

The province of Salamanca is located on the western Spain, bordered in the west by Portugal. It has an area of 12.349 km2 and had a population of 342,857 people in 2014; 167,459 (49%) male and 175.398 (51%) female citizens. It is divided into 362 municipalities; more than half are villages with fewer than 300 people. In fact, 227,878 (67%) people live in 10 municipalities of more than 5,000 individuals that will be

considered for future analysis as urban areas and 114,581 (33%) people live in the rest of municipalities and consequently will be considered as rural areas.

Spain´s and consequently Salamanca's healthcare system is public, guaranteeing universal coverage. In total, 98.7% of the population are insured for this public Spanish healthcare system. In Salamanca, a total of 35 primary health centres throughout the province provide healthcare services to the overall population: 18 to the urban-considered municipalities and 17 to the rural-considered municipalities (**Figure 1**).

Individuals aged ≥18 years included in the lists of all primary healthcare facilities of the province of Salamanca represented the reference population of 295,975 subjects: mean age 52.9±19.8 years; 52.4% females; 61.3% residing in urban areas. A sample size of 2400 subjects is calculated based on an expected prevalence of structural heart disease of 6% with a confidence interval of 95% and a 1% precision. In order to obtain the necessary sample size, 35% more requests for participation will be made, estimating errors of location from the healthcare database or refuses to participate in the study. Thus, 3564 people will be randomly selected from the primary care lists.

Cohort participants will undergo a basal examination visit, in these primary healthcare centres, between 2015 and 2018. Surviving participants are expected to return for a 5 and 10-year follow-up visit. Institutional review committee approval was obtained and all participants will provide informed consent. The SALMANTICOR study is designed to provide echocardiographic parameters characterizing cardiac structure and function in all individuals. SALMANTICOR participants will undergo surveillance for cardiovascular events, including heart failure, incident coronary heart disease, and all-cause mortality.

Medical investigation process

Medical history, surveys completion, and examinations will be obtained at the subject´s primary care referral centre and will be analysed and interpreted centrally at the University Hospital of Salamanca. A complete medical history, physical examination and the surveys completion checkout will be performed by a cardiologist in a separate office, where examinations and blood sample extraction will be performed. Echocardiographic measures will be initially performed. Participant's blood pressure and VASERA measures will be taken within 30 minutes after starting the echocardiographic exam and after the subject will be resting for 10 minutes. ECG will be performed after VASERA to finalize with the blood sample extraction.

Questionnaires

After obtaining written informed consent, trained interviewers will use a structured questionnaire to collect baseline data in face-to-face interviews at the time of physical examination. Self-reported diseases will be verified by individuals' primary care doctors according to recognized international standards. The questionnaire will collect information on demographics and cardiovascular risk factors, cardiovascular and non-cardiovascular medical history, physical examination, medication, socio-economic status, dietary habits as well as life-style and physical activity. (**Table 1**)

Echocardiographic Assessment

A standardized echocardiography ultrasound examination, including M-mode, 2D, spectral, colour flow and tissue Doppler will be performed by a certified technical professional using Philips CX-50 scanner with a standard 2.5-3.5-MHz phased-array probe. Image acquisition will be performed using a preprogramed acquisition protocol (**Table 2)**; following American and European Society of Echocardiography recommendations.[22-24] All studies will be acquired and stored digitally on a local PACS and transferred from field primary care centres to a secure server at the Salamanca

University Hospital on the same day via a dedicated VPN connection. Development of the imaging and analysis protocol, field centre echocardiography manual of operations, reading centre manual of operations, field centre sonographer, training of sonographer occurred from July 2015 to October 2015, followed by the initiation of the SALMANTICOR visit in November 2015, which was continued until May 2018.

For patients in sinus rhythm, >3 full cardiac cycles will be recorded for each view, with recording beginning once the view is optimized. For subjects in atrial fibrillation, >5-second acquisitions per view will be recorded. Sonographers are instructed to continuously optimize both imaging depth and sector width to maintain a frame rate of 50 to 80 frames per second. Sonographers are also instructed to adjust 2D gain and compression, when necessary, to optimally demonstrate left ventricle endocardial borders. The colour Doppler Nyquist limit will be set at 64 cm/s. Colour Doppler gain will be set just below the level at which random background noise will be seen. Sonographers will optimally align spectral Doppler parallel to the direction of the blood flow of interest. Sonographers will optimize the baseline shift and velocity range so that the spectral envelope will occupy approximately three fourths of the display. All spectral Doppler acquisitions will be performed with a sweep speed between 75 to 100 cm/s, and a sample volume length of 3 mm for pulsed-wave Doppler. The tissue Doppler sample volume will be placed at the level of annulus (mitral and tricuspid) and the baseline shift and velocity range will be optimized. All tissue Doppler acquisitions will be performed with similar acquisitions of spectral Doppler with a filter setting of 100 Hz.

Echocardiograms will be obtained at the subject´s primary care referral centre and sonographers will not perform any measurements on the images obtained because all measurements will be analysed and interpreted centrally at the University Hospital of

Salamanca. All SALMANTICOR echocardiograms will be read by a certified cardiologist and over-read by a board-certified cardiologist with expertise in echocardiography variables assessment (**Table 3)**. Over-reads of echocardiograms will be performed to confirm the accuracy of key quantitative measurements and to identify clinically important findings. Inter and intra-reader reproducibility was assessed before initiating the trial. For inter-reader reproducibility, intra-class correlation values ranged from 0.85 to 0.99 with left atrial volume and LV end-diastolic volumes having the highest intra-class correlation values (0.97-0.99). Intra-class correlation values were slightly better from intra-reader assessments for all measures.

Vascular Function Assessment

Cardio-ankle vascular index (CAVI), brachial ankle pulse wave velocity (baPWV) and ankle-brachial index (ABI) will be estimated using the VaSera VS-1500® device (Fukuda Denshi) as described by our group.[25] The baPWV will be calculated, as well as CAVI, which provides a more accurate estimation of the atherosclerosis degree. CAVI integrates cardiovascular elasticity derived from the aorta to the ankle pulse velocity through an oscillometric method; it is used as a good measure of vascular stiffness and does not depend on blood pressure.[26] CAVI values will be automatically calculated by substituting the stiffness parameters in the following equation to detect the vascular elasticity and the ba-PWV; where p is the blood density, Ps and Pd are systolic blood pressure and diastolic blood pressure in mm Hg, respectively; and baPWV is measured between the aortic valve and ankle.

$$stiffnes\ parameter\ \beta = 2p\ x \frac{1}{(Ps - Pd)} \times\ In\ \left(\frac{Ps}{Pd}\right) \times\ baPWV^2$$

The average coefficient of the variation of CAVI is <5%, which is small enough for clinical use and confirms that CAVI has favourable reproducibility.[27 28] CAVI and ABI

will be measured in the resting position. baPWV is estimated using the following equation; where tba is the time the same waves were transmitted to the ankle.

$$baPWV = \frac{(0.5934\ x\ height\ [cm] + 14.4724)}{tba}$$

For the study, the lowest ABI and the highest CAVI and baPWV obtained will be considered. CAVI is classified as normal (CAVI<8), borderline (8≤CAVI<9) and abnormal (CAVI≥9). Abnormal CAVI represents subclinical atherosclerosis, and baPWV ≥17.5 is considered abnormal.[29 30] ABI ≤ 0.9 was considered abnormal.

Electrocardiographic examination

Electrocardiographic examination will be performed using a General Electric MAC 3500 ECG System (Niskayuna, New York, USA), which automatically measures wave voltage and duration. ECG will be performed by the same nurse trained to carefully standardized procedures for ECG acquisition. The standard 12-lead ECGs will be obtained at a paper speed of 25 mm/sec, amplitude of 10 mm/1mV, and a filter range 0.04 to 40 Hz from all patients. ECG tracing will be interpreted in a similar way to the echocardiographic protocol by independent cardiologist and over-read by a board-certified cardiologist with expertise in electrocardiography (Dr. Jesús Hernández) at the University Hospital of Salamanca. ECG measurements and interpretations will be done following standard methods,[31 32] (**Table 4**).

Laboratory test

Venous blood sampling will be performed at the end of the examination after participants have fasted and abstained from smoking, consumption of alcohol and caffeinated beverages for 12 hours, following the protocol used in our hospital for other multidisciplinary projects.[25] A total of 20 mL of venous blood will be drawn for research testing. Blood will be drawn as follows: EDTA 10 mL and serum 10 mL.

Aliquots of plasma (3 x 2 mL), serum (4 x 2 mL) and white cell pellet (3 x 2 mL) will be stored in freezers (-80ºC) until the analysis. All biomaterial (serum, plasma and white blood cells) will be stored in the IBSAL biobank. Referral for biobanking is carried out through a specific electronic database. Biochemical tests include NT-proBNP, troponin, haemoglobin, blood cell count, thrombocytes, ferritin and iron, transferrin and iron saturation, potassium, sodium and creatinine, glycated haemoglobin, plasma glucose, aspartate aminotransferase, alanine aminotransferase, total cholesterol, triglycerides, HDL and LDL, uric acid, high-sensitive C-reactive protein, thyroid-stimulating hormone. Further, biomarkers indicative of different pathophysiological mechanisms relevant to heart disease will be analysed. A white cell pellet will be used for genotyping.

Results and Outcomes

After the clinical history is performed and the echocardiogram and electrocardiogram are interpreted, a clinical report is sent to the patient and to the primary care medical doctor. Individuals needing a further evaluation will be sent to the Cardiology Department through a preference standardized protocol.

Individuals will be contacted at 5-years intervals to ascertain the clinical status and to repeat the described basal evaluations. Clinical outcomes will include cardiovascular MACE, commencing dialysis and first hospitalization.

Statistical Analysis

Casual and multivariate inference

Data input will be stored in a database designed for the project. Normal distribution of variables will be verified using the Kolmogorov-Smirnov test. Quantitative variables will be displayed as mean ± standard deviation if normally distributed or as the median (interquartile range) if asymmetrically distributed and qualitative variables will be

expressed as frequencies. Analysis of the difference of means between variables of two categories will be carried out using a Student´s t test or a Mann-Whitney U test, as appropriate, while qualitative variables will be analysed using a χ^2 test. To analyse the relationship between qualitative variables of more than two categories and quantitative variables, an analysis of variance and the least significant difference test will be used in the post-hoc tests. The relationship of quantitative variables to each other will be tested using Pearsons or Spearmans correlation as appropriate. ANCOVA (covariance analysis) will be performed to adjust the variables that can affect the results as confounders. A multivariate analysis of variance (MANOVA) will be performed in cases with more than one dependent variable to identify whether changes in the independent variables have significant effects on the dependent variables. The association between the variables studied will be performed by multiple linear regression. Data will be analysed using the SPSS version 23.0 statistical package (SPSS Inc., Chicago, Illinois, USA). A value of $p < 0.05$ will be considered as statistically significant.

Spatial analysis

Additionally, this research aims having a spatial understanding of the structural heart disease abnormalities in the province of Salamanca. Such demanding task will be carried out by applying different statistic procedures as Multiple Factor Analysis (MFA) and Cokriging.

MFA is an extension of Principal Component Analysis (PCA) tailored to handle distinct variables (quantitative, categorical or frequency) and different data tables collected on the same observations.[33] MFA is put into practice depending on the data tables and the variables types: in the case of quantitative variables a PCA is applied; Multiple Correspondence Analysis (MCA) is applied in case of categorical variables [34];

and Correspondence Analysis (CA) for frequency variables.[35] Cokriging is a multivariate geostatistical procedure used for interpolation purposes.[36] This method is a generalization of a multivariate linear-weighted regression model, where weights depend on distance, direction and orientation of the neighbouring data to the unsampled location.

In the SALMANTICOR study, we will further combine MFA and Cokriging. In our case, we have two different levels of observations, participants and municipalities. As a mathematical comparison, municipalities contain participants, therefor if we want to extend our investigation to a spatial analysis we need to use the resulting MFA projections on their corresponding municipality areas and then apply a Cokriging analysis on the unsampled municipalities (**Figure 2**) (**supplementary data**). This combination will provide a spatial understanding of the Salamanca population and will cover the whole analysis, however if we want to focus on a specific questionnaire we could skip the MFA and look at the results obtained from the MCA, PCA or CA and then apply a Cokriging analysis. In addition, if we require analysing a particular item from a questionnaire we could also perform the analysis. To summarize, we have a versatile methodology that permit to study as concrete aspects as wider analysis of our study.

The R packages FactoMineR and Gstat will be used in order to apply MFA and Cokriging respectively.[37] [38] An additional code will be shared in a public Github repository.

Machine learning

The SALMANTICOR study will also be analysed following the ML pipeline represented in **Figure 3**. Our first step will consist in the development of scalable methods for ML optimization with the aim to develop a first approach to the predictive

structural heart disease model. Our ML model will start from ingesting raw data, leveraging data processing techniques to wrangle, process and engineer meaningful features and attributes from this data (feature engineering). The derived features are attributes or properties shared by all the independent units on which analysis or prediction is to be done. In our case, clinical variables and variables quantified from imaging data will be chosen. Features will be combined with scalable ML algorithms, including deep learning process and automatic extraction of data functionalities, in order to develop the model (fit model). The model´s basic behaviour and functionalities will be tested to develop a robust and reliable model (training model). We will validate, train and improve the ML model in a trial an error process until satisfactory model performance (validation). The SALMANTICOR study sample will be randomly divided into a train dataset (70% of the sample) and a validation dataset (30% of the sample), following previous published ML models.[39] We will use our train dataset to fit our ML model and the validation dataset to evaluate our results. This process will be repeated multiple times to guarantee a robust fit without overfitting. We will build our predictor models using: random forest, gradient boosting, logistic regression, K-nearest neighbours, support vector machine, linear discriminant analysis and naive Bayesian network models (**supplementary data**). Our ML pipeline setup will compare the performance of each algorithm on the dataset using a set of carefully selected evaluation criteria (i.e., classification accuracy, logarithmic loss, confusion matrix, area under curve, F1 score, mean absolute error, mean squared error) and the categorization of the specific cardiac problem.

For the realization of this ML models we will use free software (Python) and free open-source unified workbench such as Scikit-learn.[40]

Quality control

Different processes will be carried out to guarantee study data quality and thus maximize the validity and reliability of measurements of the results. To this effect, field work operation manuals have been prepared. These documents specify the adequate procedure for performing each test. All of these actions will confirm adequate performance of each procedure. Monthly meetings will be held with the principal investigator of the study to analyse the entire process, and an annual report on study progress will be prepared.

Ethical Review Board and dissemination plan

The study has been approved by the clinical research ethics committee (CEIC) of the health area of Salamanca ('CEIC of Salamanca Health Area, 9/29/2014). Participants will be required to sign an informed consent form prior to participation in the study, in accordance with the declaration of Helsinki and the WHO standards for observational studies. The study has been registered in ClinicalTrials.gov with identifier NCT03429452. Participants will be informed of the objectives of the project and of the risks and benefits of the examinations made. None of the examinations pose life-threatening risks for the type of participants to be included in the study. The study includes the obtaining of biological samples (including genetics analysis); the study participants therefore will be informed in detail. The confidentiality of the recruited participants will be ensured at all times in accordance with the provisions of current legislation on personal data protection (15/1999 of December 13, LOPD), and the conditions contemplated by Act 14/2007 on biomedical research.

We will use a variety of methods to ensure that our work will achieve maximum visibility. Publication of our study protocol provides an important first step towards this direction. In this paper, we have sought to offer a comprehensive overview of relevant

literature, while underlining current research gaps that necessitated the design and implementation of the SALMANTICOR study. Similarly, the study results, given their applicability and implications for the general population, will be disseminated in research meetings and in at least ten articles published in scientific journals. Finally, population-based control groups are difficult to obtain, specifically in case-control cardiovascular studies where structural heart disease has to be rolled out. The SALMANTICOR study will provide availability of normative reference values quantification for echocardiographic, electrocardiographic, biochemical, genetics, VASERA and other parameters. Thus, international cooperation sharing data and participating in Horizon 2020 programs with the SALMANTICOR population are contemplated.

Patient and public involvement

Patients´ representatives will have an increasingly present voice in the SALMANTICOR study. There is currently an only patient organization for heart disease in the province of Salamanca, "El Paciente Experto". This organization has provided counselling in the design of the study, will jointly interpret the results of the study with the investigators of SALMANTICOR, will help to disseminate them to society, and will be involved when establishing new policies for health improvement and education empowerment with the Administration to halt the epidemic of cardiovascular disease.

A clinical report will be sent to all participants and their primary care medical doctors immediately after the clinical history is performed and the echocardiogram and electrocardiogram interpreted. Finally, the global and most important observations from the SALMANTICOR study will be sent by letter to all participants and to all doctors,

primary care and specialists, of the province of Salamanca through the Medical College of Salamanca and our health Administration.

Data statement

Our data will be accessed at the Institute of Research of the University Hospital of Salamanca. Furthermore, our dataset will be published in a public repository. Additional code for our spatial analysis will be shared in a public Github repository.

Discussion

A major strength of the SALMANTICOR study is the selection of a representative population-based cohort across primary care, with a probable significate number of structural heart disease cases of each age, sex and place of residence category to allow overall and subpopulation analyses. This population-based approach increases the generalizability of the findings compared to surveys that addressed cardiovascular risk factors but have never included an echocardiographic assessment.[11 14 41-44] Moreover, in view of the similarity of trends in cardiovascular disease and population ageing from Spain with other developed countries,[45] our findings are likely to be broadly applicable to them.

Echocardiography in the SALMANTICOR study is designed to address three specific aims. The first one is to characterize the abnormalities of cardiac structure and function in a community-based sample and to assess how these abnormalities vary by place of residence (rural or urban), by age and, by sex. The study uses standard and novel echocardiographic techniques to characterize five specific domains of cardiac structure. These data will be used to define the population distribution of these measurements and to determine their relationship with the cardiovascular risk factors,

including hypertension, diabetes mellitus, coronary disease, renal insufficiency, and prognostically relevant biomarkers such as N-terminal pro-brain natriuretic peptide and high-sensitivity troponin.

The second aim is to investigate ventricular-arterial coupling in addition to the association of cardiac structure and function with arterial stiffness assessed by CAVI, baPWV and ABI.

The third aim is to prospectively examine the extent to which these non-invasive measures associate with incidences of adverse cardiovascular outcomes and to determine the degree to which these associations also vary by age, sex and by place of residence (rural or urban). By accomplishing these objectives, this study is developing an echocardiographic imaging database that will facilitate future investigations to compare these echocardiographic measures both with studies previously performed in other Countries,[12 13] and to be used as a very well established control group. Furthermore, our study will provide availability of normative reference values quantification for electrocardiographic, biochemical, genetics, VASERA and other parameters.

Adequate public health and service delivery planning requires reliable information about contemporary population-level disease incidence. SACYL is the regional health-care government authority of Castilla y Leon providing universal access to health services for 2,5 million people. SACYL is closely integrated with other public services and policies as part of a holistic approach to improving population health. In this sense, our study data will be used to understand the cardiovascular health needs of our community and to improve people´s health and wellbeing, and how they can be developed. SALMANTICOR will be established as the global observatory on cardiovascular health research and development of SACYL, since we will include real-

time data about the burden of cardiovascular disease, people´s social circumstances and living conditions, lifestyles and diet, economic factors, access to healthcare and other services, as well as our genes, age and sex. In addition to understating the overall picture of our population´s health, the data will be disaggregated to identify inequalities for example by gender, sex, and urban or rural place of residence. This will support the prioritization of interventions depending on the needs of different groups and will require effective actions for the prediction and prevention of cardiovascular disease; from macro-policies down to individuals and families, empowering people to take control of their health. In this sense, two new medical technology research lines have been identified by the SALMANTICOR investigators: exploring the use of spatial methods and exploring modern computational methods developed in the field of ML.

The use of spatial methods in healthcare research enables disease distribution patterns to be identified and has become popular in the field of public health,[46-48] Cancer and other disease mortality atlases have shown us that many risk factors of a territorial nature, influence geographical patterns, making it possible to select disease indicators and so reveal their geographical structure.[49 50] However, the number of spatial analyses published in major epidemiology journals is still very low.[51] One of the reasons is that the application of spatial methods requires specific training and has resulted in their substitution with less optimal methods from healthcare research. Therefore, it is important to promote spatial methods, especially those which are simple to interpret in the field of population-based studies and which could be potentially used in combination with other computational methods to facilitate interpretation, prediction and healthcare policies. Cardiology spatial analysis has been developed mainly in optimization problems and prevalence prediction. As an example of optimization, travel time isochrones analysis has been deployed in different facilities in order to identify

exposed areas and act accordingly.[52] Nevertheless, prevalence predictions are the most common geostatistical techniques in healthcare and it is not an exception in cardiology.[53 54]

The incorporation of ML in medicine holds promise for substantially improved health-care delivery[18-21]. ML provides methods, techniques, and tools that can help solving diagnostic and prognostic problems in a variety of cardiac medical domains[55-63]. Furthermore, ML offers new approaches to leveraging the growing volume of heterogeneous data, including imaging data, available for analyses. To date, ML has been used in two broad and highly interconnected areas: automatization of tasks that might otherwise be performed by a human and generation of clinically important knowledge. However, it is argued that the successful implementation of ML methods can help the integration of computer-based systems in the healthcare environment providing opportunities to really improve the efficiency of medical care and to be used as a regional policy to establish effective public health programs. In this sense, The SALMANTICOR study represents an excellent opportunity to explore ML algorithms for estimating and ranking the impact of environmental and classical risk factors in the development of structural heart disease in a population-based setting.

References

1. CDC, NCHS. Underlying Cause of Death 1999-2015 on CDC WONDER Online Database, released 2017. Data are from the Multiple Cause of Death Files, 1999-2015, as compiled from data provided by the 56 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed Dec. 6, 2017.
2. European Cardiovascular Disease Statistics 2017 on www.ehnheart.org/cvd:statistics.html, released 2017. Data are from the European Heart Network (EHN), a Brussels-based Alliance of heart foundations and likeminded non-governmental organisations throughout Europe, with member organisations in 25 countries. Accessed Dec. 6, 2017.
3. Mozaffarian D, Benjamin EJ, Go AS, et al. Heart disease and stroke statistics--2015 update: a report from the American Heart Association. *Circulation* 2015;131(4):e29-322. doi: 10.1161/CIR.0000000000000152
4. Pearson TA, Palaniappan LP, Artinian NT, et al. American Heart Association Guide for Improving Cardiovascular Health at the Community Level, 2013 update: a scientific statement for public health practitioners, healthcare providers, and health policy makers. *Circulation* 2013;127(16):1730-53. doi: 10.1161/CIR.0b013e31828f8a94
5. Gerber Y, Weston SA, Enriquez-Sarano M, et al. Contemporary Risk Stratification After Myocardial Infarction in the Community: Performance of Scores and Incremental Value of Soluble Suppression of Tumorigenicity-2. *J Am Heart Assoc* 2017;6(10) doi: 10.1161/JAHA.117.005958
6. Dondo TB, Hall M, Timmis AD, et al. Geographic variation in the treatment of non-ST-segment myocardial infarction in the English National Health Service: a cohort study. *BMJ open* 2016;6(7):e011600. doi: 10.1136/bmjopen-2016-011600
7. Zhang L, Desai NR, Li J, et al. National Quality Assessment of Early Clopidogrel Therapy in Chinese Patients With Acute Myocardial Infarction (AMI) in 2006 and 2011: Insights From the China Patient-Centered Evaluative Assessment of Cardiac Events (PEACE)-Retrospective AMI Study. *J Am Heart Assoc* 2015;4(7) doi: 10.1161/JAHA.115.001906
8. Regueiro A, Bosch J, Martin-Yuste V, et al. Cost-effectiveness of a European ST-segment elevation myocardial infarction network: results from the Catalan Codi Infart network. *BMJ open* 2015;5(12):e009148. doi: 10.1136/bmjopen-2015-009148
9. Conrad N, Judge A, Tran J, et al. Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals. *Lancet* 2017 doi: 10.1016/S0140-6736(17)32520-5
10. Dawber TR, Meadors GF, Moore FE, Jr. Epidemiological approaches to heart disease: the Framingham Study. *Am J Public Health Nations Health* 1951;41(3):279-81.
11. Teo K, Chow CK, Vaz M, et al. The Prospective Urban Rural Epidemiology (PURE) study: examining the impact of societal influences on chronic noncommunicable diseases in low-, middle-, and high-income countries. *Am Heart J* 2009;158(1):1-7 e1. doi: 10.1016/j.ahj.2009.04.019

12. Shah AM, Cheng S, Skali H, et al. Rationale and design of a multicenter echocardiographic study to assess the relationship between cardiac structure and function and heart failure risk in a biracial cohort of community-dwelling elderly persons: the Atherosclerosis Risk in Communities study. *Circulation Cardiovascular imaging* 2014;7(1):173-81. doi: 10.1161/CIRCIMAGING.113.000736 [published Online First: 2013/11/12]
13. Vasan RS, Xanthakis V, Lyass A, et al. Epidemiology of Left Ventricular Systolic Dysfunction and Heart Failure in the Framingham Study: An Echocardiographic Study Over 3 Decades. *JACC Cardiovasc Imaging* 2017 doi: 10.1016/j.jcmg.2017.08.007
14. Yusuf S, Hawken S, Ounpuu S, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet* 2004;364(9438):937-52. doi: 10.1016/S0140-6736(04)17018-9
15. O'Donnell MJ, Xavier D, Liu L, et al. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. *Lancet* 2010;376(9735):112-23. doi: 10.1016/S0140-6736(10)60834-3
16. Chambers J, Kabir S, Cajeat E. Detection of heart disease by open access echocardiography: a retrospective analysis of general practice referrals. *Br J Gen Pract* 2014;64(619):e105-11. doi: 10.3399/bjgp14X677167
17. englund EJ. A variance of statisticians. *Math Geol* 1990;22(4):417-55.
18. Deo RC. Machine Learning in Medicine. *Circulation* 2015;132(20):1920-30. doi: 10.1161/CIRCULATIONAHA.115.001593
19. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* 2016;375(13):1216-9. doi: 10.1056/NEJMp1606181
20. Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. *N Engl J Med* 2017;376(26):2507-09. doi: 10.1056/NEJMp1702071
21. Shameer K, Johnson KW, Glicksberg BS, et al. Machine learning in cardiovascular medicine: are we there yet? *Heart* 2018 doi: 10.1136/heartjnl-2017-311198
22. Lang RM, Badano LP, Mor-Avi V, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *Eur Heart J Cardiovasc Imaging* 2015;16(3):233-70. doi: 10.1093/ehjci/jev014
23. Marwick TH, Gillebert TC, Aurigemma G, et al. Recommendations on the use of echocardiography in adult hypertension: a report from the European Association of Cardiovascular Imaging (EACVI) and the American Society of Echocardiography (ASE)dagger. *Eur Heart J Cardiovasc Imaging* 2015;16(6):577-605. doi: 10.1093/ehjci/jev076
24. American College of Cardiology Foundation Appropriate Use Criteria Task F, American Society of E, American Heart A, et al. ACCF/ASE/AHA/ASNC/HFSA/HRS/SCAI/SCCM/SCCT/SCMR 2011 Appropriate Use Criteria for Echocardiography. A Report of the American College of Cardiology Foundation Appropriate Use Criteria Task Force, American Society of Echocardiography, American Heart Association, American Society of Nuclear Cardiology, Heart Failure Society of America, Heart Rhythm Society, Society for Cardiovascular Angiography and Interventions, Society of

Critical Care Medicine, Society of Cardiovascular Computed Tomography, and Society for Cardiovascular Magnetic Resonance Endorsed by the American College of Chest Physicians. *J Am Coll Cardiol* 2011;57(9):1126-66. doi: 10.1016/j.jacc.2010.11.002

25. Gomez-Marcos MA, Martinez-Salgado C, Gonzalez-Sarmiento R, et al. Association between different risk factors and vascular accelerated ageing (EVA study): study protocol for a cross-sectional, descriptive observational study. *BMJ open* 2016;6(6):e011031. doi: 10.1136/bmjopen-2016-011031 [published Online First: 2016/06/09]
26. Takaki A, Ogawa H, Wakeyama T, et al. Cardio-ankle vascular index is a new noninvasive parameter of arterial stiffness. *Circulation journal : official journal of the Japanese Circulation Society* 2007;71(11):1710-4. [published Online First: 2007/10/30]
27. Shirai K, Hiruta N, Song M, et al. Cardio-ankle vascular index (CAVI) as a novel indicator of arterial stiffness: theory, evidence and perspectives. *Journal of atherosclerosis and thrombosis* 2011;18(11):924-38. [published Online First: 2011/06/02]
28. Shirai K. Analysis of vascular function using the cardio-ankle vascular index (CAVI). *Hypertension research : official journal of the Japanese Society of Hypertension* 2011;34(6):684-5. doi: 10.1038/hr.2011.40 [published Online First: 2011/06/07]
29. Hu H, Cui H, Han W, et al. A cutoff point for arterial stiffness using the cardio-ankle vascular index based on carotid arteriosclerosis. *Hypertension research : official journal of the Japanese Society of Hypertension* 2013;36(4):334-41. doi: 10.1038/hr.2012.192 [published Online First: 2013/01/18]
30. Kawai T, Ohishi M, Onishi M, et al. Cut-off value of brachial-ankle pulse wave velocity to predict cardiovascular disease in hypertensive patients: a cohort study. *Journal of atherosclerosis and thrombosis* 2013;20(4):391-400. [published Online First: 2012/12/28]
31. Macfarlane PW, Katibi IA, Hamde ST, et al. Racial differences in the ECG--selected aspects. *J Electrocardiol* 2014;47(6):809-14. doi: 10.1016/j.jelectrocard.2014.08.003
32. Rijnbeek PR, van Herpen G, Bots ML, et al. Normal values of the electrocardiogram for ages 16-90 years. *J Electrocardiol* 2014;47(6):914-21. doi: 10.1016/j.jelectrocard.2014.07.022
33. Escofier B, Pages J. Multiple factor for analysis (ALMULT package). *Comput Stat Data Anal* 1994;18:121-40.
34. Guisado-Clavero M, Roso-Llorach A, Lopez-Jimenez T, et al. Multimorbidity patterns in the elderly: a prospective cohort study with cluster analysis. *BMC Geriatr* 2018;18(1):16. doi: 10.1186/s12877-018-0705-7
35. Benzecri JP. L´Analyse des Données. Volume II. L´Analyse des correspondances. *Paris Dunod* 1973
36. Wackermagel H. Multivariate Geostatistics: An Introduction with Applications. *New York, NY: Springer-Verlag* 2003
37. Le S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software* 1990;25(1):1-18.
38. Pebesma EJ. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences* 2004;30:683-91.

39. Frizzell JD, Liang L, Schulte PJ, et al. Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure: Comparison of Machine Learning and Other Statistical Approaches. *JAMA Cardiol* 2017;2(2):204-09. doi: 10.1001/jamacardio.2016.3956
40. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;12:2825-30.
41. Grau M, Elosua R, Cabrera de Leon A, et al. [Cardiovascular risk factors in Spain in the first decade of the 21st Century, a pooled analysis with individual data from 11 population-based studies: the DARIOS study]. *Rev Esp Cardiol* 2011;64(4):295-304. doi: 10.1016/j.recesp.2010.11.005
42. Masia R, Pena A, Marrugat J, et al. High prevalence of cardiovascular risk factors in Gerona, Spain, a province with low myocardial infarction incidence. REGICOR Investigators. *J Epidemiol Community Health* 1998;52(11):707-15.
43. Rigo Carratala F, Frontera Juan G, Llobera Canaves J, et al. [Prevalence of cardiovascular risk factors in the Balearic Islands (CORSAIB Study)]. *Rev Esp Cardiol* 2005;58(12):1411-9.
44. Felix-Redondo FJ, Fernandez-Berges D, Fernando Perez J, et al. [Prevalence, awareness, treatment and control of cardiovascular risk factors in the Extremadura population (Spain). HERMEX study]. *Aten Primaria* 2011;43(8):426-34. doi: 10.1016/j.aprim.2010.07.008
45. Roth GA, Johnson C, Abajobir A, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J Am Coll Cardiol* 2017;70(1):1-25. doi: 10.1016/j.jacc.2017.04.052
46. Elliott P, Wartenberg D. Spatial epidemiology: current approaches and future challenges. *Environ Health Perspect* 2004;112(9):998-1006.
47. Abellan JJ, Richardson S, Best N. Use of space-time models to investigate the stability of patterns of disease. *Environ Health Perspect* 2008;116(8):1111-9. doi: 10.1289/ehp.10814
48. Kontopantelis E, Stevens RJ, Helms PJ, et al. Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study. *BMJ open* 2018;8(2):e020738. doi: 10.1136/bmjopen-2017-020738
49. Ho NT, Thompson C, Nhan LNT, et al. Retrospective analysis assessing the spatial and temporal distribution of paediatric acute respiratory tract infections in Ho Chi Minh City, Vietnam. *BMJ open* 2018;8(1):e016349. doi: 10.1136/bmjopen-2017-016349
50. Lopez-Abente G, Aragones N, Perez-Gomez B, et al. Time trends in municipal distribution patterns of cancer mortality in Spain. *BMC Cancer* 2014;14:535. doi: 10.1186/1471-2407-14-535
51. Auchincloss AH, Gebreab SY, Mair C, et al. A review of spatial methods in epidemiology, 2000-2010. *Annu Rev Public Health* 2012;33:107-22. doi: 10.1146/annurev-publhealth-031811-124655
52. Collaborators GBDRF, Forouzanfar MH, Alexander L, et al. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2015;386(10010):2287-323. doi: 10.1016/S0140-6736(15)00128-2

53. Przybysz R, Bunch M. Exploring spatial patterns of sudden cardiac arrests in the city of Toronto using Poisson kriging and Hot Spot analyses. *PLoS One* 2017;12(7):e0180721. doi: 10.1371/journal.pone.0180721
54. Ogunniyi MO, Holt JB, Croft JB, et al. Geographic variations in heart failure hospitalizations among medicare beneficiaries in the Tennessee catchment area. *Am J Med Sci* 2012;343(1):71-7. doi: 10.1097/MAJ.0b013e318223bbd4
55. Garcia EV, Cooke CD, Folks RD, et al. Diagnostic performance of an expert system for the interpretation of myocardial perfusion SPECT studies. *J Nucl Med* 2001;42(8):1185-91.
56. Paul AK, Shill PC, Rabin MRI, et al. Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease. *Applied Intelligence* 2018;48:1739-56. doi: 10.1007/s10489-017-1037-6
57. Raghavendra U, Fujita H, Gudigar A, et al. Automated technique for coronary artery disease characterization and classification using DD-DTDWT in ultrasound images. *Biomedical Signal Processing and Control* 2018;40:324-34. doi: 10.1016/j.bspc.2017.09.030
58. Alizadehsani R, Zangooei Mh, Hosseini MJ, et al. Coronary artery disease detection using computational intelligence methods. *Knowledge-Based Systems* 2016;109:187-89. doi: 10.1016/j.knosys.2016.07.004
59. Tan JH, Hagiwara Y, Pang W, et al. Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. *Comput Biol Med* 2018;94:19-26. doi: 10.1016/j.compbiomed.2017.12.023
60. Alizadehsani R, Hosseini MJ, Khosravi A, et al. Non-invasive detection of coronary artery disease in high-risk patients based on the stenosis prediction of separate coronary arteries. *Comput Methods Programs Biomed* 2018;162:119-27. doi: 10.1016/j.cmpb.2018.05.009
61. Arabasadi Z, Alizadehsani R, Roshanzamir M, et al. Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. *Comput Methods Programs Biomed* 2017;141:19-26. doi: 10.1016/j.cmpb.2017.01.004
62. Acharya UR, Fujita H, Lih OS, et al. Automated detection of coronary artery disease using different durations of ECG segments with convolutional neural network. *Knowledge-Based Systems* 2017;132:62-71. doi: 10.1016/j.knosys.2017.06.003
63. Acharya UR, Fujita H, Adam M, et al. Automated characterization and classification of coronary artery disease and myocardial infarction by decomposition of ECG signals: A comparative study. *Information Sciences* 2017;377:17-29. doi: 10.1016/j.ins.2016.10.013

Author statement

Jose Ignacio Melero-Alegría: data acquisition, surveys completion, physical, electrocardiographic and VASERA examinations, design of the work, drafting the work and revising it critically, final approval of the version to be published; Manuel Cascón: data acquisition, surveys completion, conception and design of the work, drafting the work and revising it critically, final approval of the version to be published; Alfonso Romero; conception and design of the work, interpretation of data, drafting the work of revising it critically, primary care coordination, final approval of the version to be published; Pedro Pablo Vara: echocardiographic data acquisition, interpretation of data, final approval of the version to be published; Manuel Barreiro-Pérez: conception and design of the echocardiographic protocol, analysis and interpretation of echocardiographic data, drafting the work and revising it critically for important intellectual content, final approval of the version to be published; Victor Vicente-Palacios: conception and design of the spatial and machine learning analysis, analysis and interpretation of data, drafting the work and revising it critically for important intellectual content, final approval of the version to be published; Fernando Pérez-Escanilla: conception and design of the work, interpretation of data, primary care coordination, final approval of the version to be published; Jesús Hernández-Hernández: conception and design of the electrocardiographic protocol, analysis and interpretation of ECG data, drafting the work and revising it critically for important intellectual content, final approval of the version to be published; Beatriz Garde: conception and design of the lifestyle, Mediterranean and exercise surveys, analysis and interpretation of data, final approval of the version to be published; Sara Cascón: conception and design of the work, coordinator of 5 out of 35 primary care centres, acquisition of data, final approval of the version to be published; Ana Martín-García: analysis and interpretation of echocardiographic data, final approval of the version to be published; Elena Díaz- Peláez: analysis and interpretation of echocardiographic data, final approval of the version to be published; José María de Dios: conception and design of the work, coordinator of 5 out of 35 primary care centres, acquisition of data, final approval of the version to be published; Aitor Uribarri: conception and design of the work (surveys), analysis and interpretation of data, final approval of the version to be published; Javier Jiménez-Candil: conception and design of the work, analysis and interpretation of ECG data, final approval of the version to be published; Ignacio Cruz-González: conception and design of the work (surveys), analysis and interpretation of data, final approval of the version to be published; Baltasara Blazquez: conception and design of the work, coordinator of 5 out of 35 primary care centres, acquisition of data, final approval of the version to be published; José Manuel Hernández: conception and design of the work, coordinator of 5 out of 35 primary care centres, acquisition of data, final approval of the version to be published; Clara Sánchez Pablos: data acquisition, surveys completion, physical, electrocardiographic and VASERA examinations, final approval of the version to be published; Inmaculada Santolino: conception and design of the work, coordinator of 5 out of 35 primary care centres, acquisition of data, final approval of the version to be published; M. Concepción Ledesma: conception and design of the work, coordinator of 5 out of 35 primary care centres, acquisition of data, final approval of the version to be published; Paz Muriel: conception and design of the work, coordinator of 5 out of 35 primary care centres, acquisition of data, final approval of the version to be published;

P. Ignacio Dorado-Díaz: conception and design of the spatial and machine learning analysis, analysis and interpretation of data, drafting the work and revising it critically for important intellectual content, final approval of the version to be published; Pedro L Sánchez: conception and design of the study, interpretation of data, drafting the work, 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.

Tables

Table 1. Questionnaires.

Name of the questionnaire	Number of variables	Principal variables	Time of completion
Demographics & Cardiovascular risk factors	12	Sex, age, residence, smoking, alcohol consumption, hypertension, hypercholesterolemia, diabetes, previous heart disease, family history	5 minutes
Cardiovascular & non-cardiovascular history	23	Coronary heart disease, arrhythmias, valvulopathies, heart failure, cardiac healthcare visits in the past and where (public or private attention), stroke, vascular peripheral disease, bleeding history, chronic kidney disease, chronic lung disease, asthma, rheumatic disease, depressive disorder, dementia, anxiety, dependency	12 minutes
Physical examination	8	Body mass index, abdominal perimeter, heart rate, oxygen saturation, blood pressure, heart murmurs & sounds	8 minutes
Medication	24	Aspirin, clopidogrel, ticagrelor, prasugrel, warfarin, acenocumarol, dabigatran, ribaroxaban, apixaban, edoxaban, betabloquers, ACE inhibitors, RAAS antagonists, calcium channel blocker, diuretics, aldosterone inhibitors, statin, ezetimibe, fibrate, ivabradine, ranolazine, proton-pump inhibitor, NSAIDs, corticoids	10 minutes
Socio-economic status	13	Marital status, education, employment, annual income, homeownership, housing quality, medical coverage	8 minutes
Dietary habits & life-style	39	Number of meals, diet, beverage, salt, bread, olive-oil, coffee, chocolate and potatoes dietary counselling, Mediterranean diet adherence, number of sleeping hours, siesta practice, pet ownership	12 minutes
Physical activity	7	Number of days, number of hours, intensity	5 minutes
Total	**126**		**60 minutes**

Table 2. Echocardiographic imaging protocol required views.

Parasternal position	
Parasternal long axis	2D imaging (at deep depth) 2D imaging (at shallow depth) Colour Doppler of the mitral and aortic valves
Parasternal short axis, aortic valve level	2D imaging of AV Colour Doppler of AV 2D imaging of RVOT Colour Doppler of RVOT PW and CW Doppler of RVOT
Parasternal short axis, mitral valve level	2D imaging
Parasternal short axis, left ventricle apex	2D imaging
Apical position	
Apical 4-chamber view	2D imaging 2D imaging, focused/zoomed of left ventricle 2D imaging, focused on left atrium Colour Doppler of mitral valve/left atrium PW Doppler of mitral flow CW Doppler of mitral flow TDI of septal and lateral mitral annulus
Apical 4-chamber view, focused on the RV	2D imaging Colour Doppler of tricuspid valve/right atrium CW Doppler of tricuspid regurgitation TDI of lateral tricuspid annulus
Apical 5-chamber view	2D imaging Colour Doppler of LVOT PW of LVOT flow CW of transaortic flow
Apical 2-chamber view	2D imaging 2D imaging focused/zoomed on LV 2D imaging focused on left atrium Colour Doppler mitral valve/left atrium
Apical 3-chamber view	2D imaging 2D imaging focused/zoomed on LV 2D imaging focused on left atrium Colour Doppler mitral valve/left atrium Colour Doppler of aortic valve PW of LVOT flow CW of transaortic flow
Subcostal view	
Inferior vena cava	2D imaging (5-s acquisition)

Table 3. Echocardiographic parameters.

Structure and function assessment	Number of variables	Principal variables Time of completion
Aorta & Atrias & ventricles	39	Ascending aorta (mm), LV diastolic dimension (mm), LV systolic dimension (mm), left ventricular mass index (g/m^2), left atrial volume index by biplanar Simpson method (mL/m^2), right ventricular diastolic dimension (mm), right atrial volume index (mL/m^2), biplanar Simpson left ventricular ejection fraction (%), mitral E-wave (cm/s), mitral A-wave (cm/s), mitral E/A, mitral deceleration time (cm/s), pulmonary artery systolic pressure (mm Hg), mitral E/e´septal annulus, mitral E/e´lateral annulus, mitral E/e´average of annulus
Valves	41	Aortic valve jet peak velocity (m/s), aortic mean gradient (mm Hg), aortic cups number, aortic valve calcification, aortic regurgitation presence and grade, mitral valve calcification, mitral mean gradient (mm Hg), mitral pressure half time (msec), mitral prolapse, mitral regurgitation presence and grade, tricuspid regurgitation presence and grade, pulmonary regurgitation presence and grade
Pericardium	3	Pericardial effusion presence and grade

Table 4. 12-lead ECG parameters.

Rhythm	Sinus rhythm Auricular tachycardia Atrial fibrillation Common atrial flutter Uncommon atrial flutter Nodal rhythm Atrial ectopies Ventricular ectopies Atrial paced rhythm Ventricular paced rhythm with sinusal activity Ventricular paced rhythm with atrial fibrillation Atrial and ventricular paced rhythm
Heart rate	
P wave	P duration Sinus P morphology Pulmonary P morphology Interatrial block
PQ time	
AV block	Not present First degree AV block Second degree AV block, Mobitz I Second degree AV block, Mobitz II 2:1 AV block Third degree or complete AV block
QRS duration	
QRS axis	
RR time	
QT time	
QT corrected time	
Brugada pattern	Not present Type I Type II Type III
AV block	Not present First degree AV block Second degree AV block, Mobitz I Second degree AV block, Mobitz II 2:1 AV block Third degree or complete AV block
Early repolarization pattern	Not present Inferior Lateral Inferior & lateral
Bundle branch configuration	Not present Complete left bundle branch block Complete right bundle branch block Incomplete left bundle branch block Incomplete right bundle branch block
Intraventricular conduction disturbances	
Fascicular block configuration	Not present

	Left anterior fascicular block Left posterior fascicular block
Notch QRS presence	
Left ventricular hypertrophy	
Delta waves presence	
Repolarization changes of digitalis	
Pathological Q-waves presence and position	
Significant ST elevation	
Significant ST depression	
Negative T-waves presence and position	

Figure legends

Figure 1. Province of Salamanca map and distribution of the total of 35 primary health centres: 18 in urban-considered municipalities (blue) and 17 in rural-considered municipalities (red). Municipalities of more than 5,000 individuals are considered as urban areas in the SALMANTICOR study.

Figure 2. The left panel represents the spatial analysis pipeline that SALMANTICOR will use for map plotting purposes. We will combine multiple factor analysis (MFA) and Cokriging. We will inquire and analyse participants from municipalities and questionnaires. Initially, for quantitative variables principal component analysis (PCA) is applied; for categorical variables, multiple correspondence analysis (MCA); and for frequency variables, correspondence analysis (CA). We will then assemble the normalized data in a single table that is analysed via PCA to describe the spatial behaviours of our samples within crossvariograms (crossvariog). We then will apply a linear model coregionalization (LMC) to finally interpolate the results over the different municipalities of the province of Salamanca using Cokriging. Maps in the right panel represent municipal spatial patterns examples of how we will represent municipal (Salamanca is divided into 362 municipalities) distribution of structural heart disease and dyslipidaemia prevalence.

Figure 3. Machine learning (ML) pipeline for the SALMANTICOR study. The learning algorithm will take heterogeneous data that will be pre-processed to create input data for the ML algorithm. Furthermore, raw images will also be used in the ML algorithm using neural network modelling. The output of the ML algorithm will also need to be processed and improved until a satisfactory model is developed.

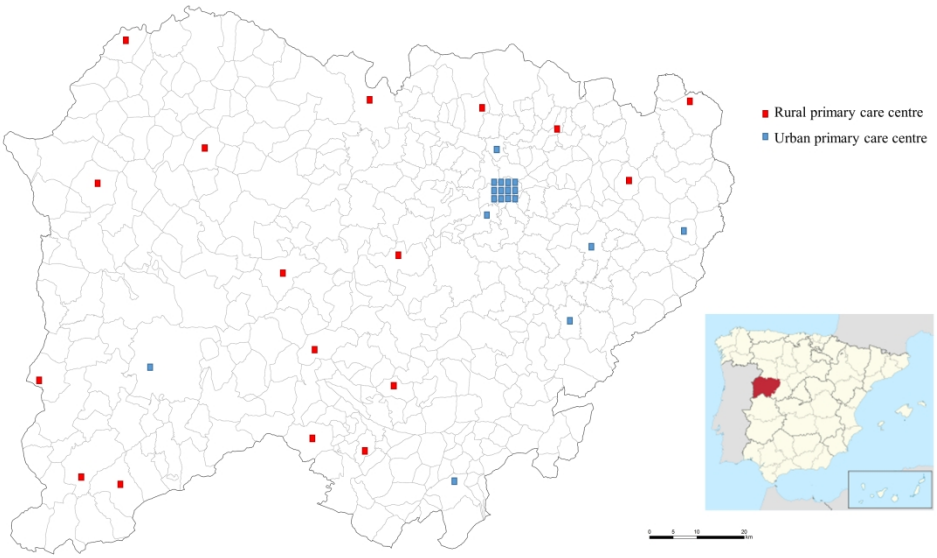

Figure 1

338x190mm (300 x 300 DPI)

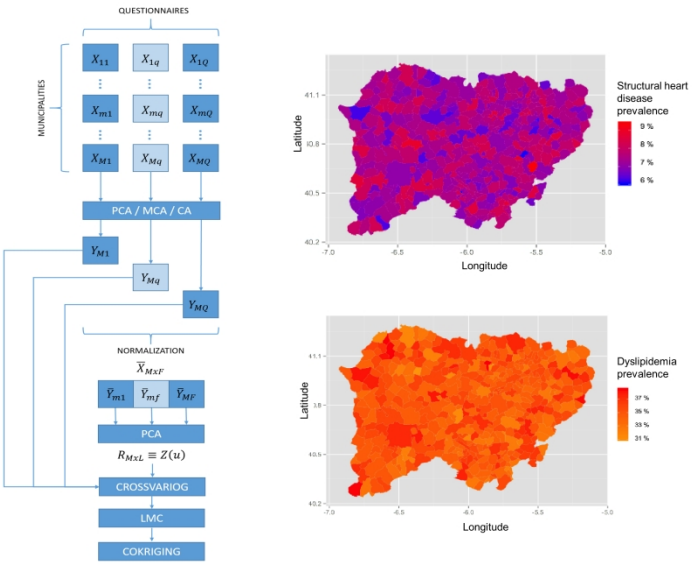

Figure 2

338x190mm (300 x 300 DPI)

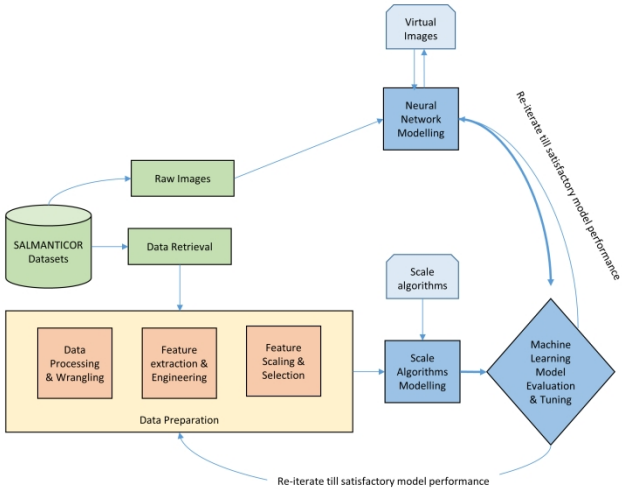

Figure 3

338x190mm (300 x 300 DPI)

Supplementary data of the SALMANTICOR study

Spatial analysis

We will combine multiple factor analysis (MFA) and Cokriging statistics procedures to provide a spatial analysis of the SALMANTICOR population.

Our study will inquire and analyzed N individuals from M municipalities. Q questionnaires were handed to all the participants. Let X_{nmq} be a matrix block where n is the number of participant of a m municipality and k is the correspondent questionnaire of our departing matrix D_{MxQ}.

Therefore, depending on the type of k questionnaire, we will employ a PCA, MCA or CA, to each block X_{nmq} obtaining $\overline{Y}_{mq} = \frac{1}{\lambda_{mq}} Y_{mq}$ where λ_{mq} is its first singular value.

Hence, we join all the resulting $\overline{Y}_{mq}$ forming a $\overline{X}_{MxF}$ matrix where M are the municipalities and F the resulting factors.

$$\overline{X}_{mf} = [\overline{Y}_{m1} | \overline{Y}_{m2} | \dots | \overline{Y}_{mf} | \dots | \overline{Y}_{mF}]$$

Finally, a generalized PCA is applied on $\overline{X}_{MxF}$

After performing MFA we will proceed to project the resulting coordinates that represents our municipalities over the resulting L latent variables obtaining R_{MxL}.

Adding the spatial coordinates u to each municipality we attain $Z(u) = [u|R]$. Once we get the $Z(u)$ matrix, we will apply a spatial interpolator such as Cokriging.

We will then describe the spatial behavior of our samples using variograms. Variograms are illustrations of how the semivariance acts in function of the distance. Semivariance is defined as half the expectation between two different values at two

locations (u and u + h), and is used in univariate analyses. To transfer our analysis to a multivariate problem we will need to build crossvariograms.

A crossvariogram γ_{ij} describes the degree of spatial dependence of our projected variables measuring the variation between two samples depending on the distance h (also known as lag) between them.

After this step, we will define

$$\Gamma(h) = \frac{1}{2}\left[\left(Z_i(u) - Z_i(u+h)\right) \cdot \left(Z_j(u) - Z_j(u+h)\right)\right]$$

with $i, j = 1 \ldots M$ and hence, the crossvariogram

Using a more practical approach, we will need to build a set of experimental crossvariograms based on our matrix $Z(u)$.

Therefore, we will obtaine $\frac{L(L+1)}{2}$ experimental semivariograms, and subsequently these direct and crossvariograms will need to be fitted. The different parts of a theoretical semivariogram are:

Nugget: It represents variability at small distances ($h \approx 0$).

Sill: The semivariance b value at which the semivariogram levels off.

Range: The a distance at which the semivariogram reaches the sill value.

The Linear Model of Coregionalization (LMC) permits all the $\frac{L(L+1)}{2}$ semivariograms to be fitted as linear combinations of S basic semivariogram functions (Gaussian, Exponential, Spherical, etc). The LMC can be expressed as a multivariate nested semivariogram model.

$$\Gamma(h) = \sum_{s=1}^{S} B_s g_s(h)$$

where $\Gamma(h)$ is the S×S matrix of semivariogram values at lag h, and B_s is the S×S matrix of sills of the basic semivariogram function $g_s(h)$. B_s has to be positive semidefinite, to assure that the variance-covariance matrix is also positive.

Once $\Gamma(h)$ is set, we will need to interpolate over the different polygons that represents the municipalities and shape the province of Salamanca. For fulfilling this task, we will apply Cokriging.

Cokriging is the multivariate extension of kriging, whose main purpose is to compute a weighted average of the sample values in close proximity to a grid point, polygon or volume. It searches for the best linear unbiased estimator, based on assumptions on covariances. There are different procedures such as ordinary, universal, or simple Cokriging.

As an example, we present simple Cokriging.

$$\bar{Z}_{i_0}(u_0) = m_{i_0} + \sum_{i=1}^{L}\sum_{\alpha=1}^{M} w_\alpha^i (Z_i(u_\alpha) - m_i)$$

where u_0 is an unsampled municipality and u_α a sample location, w_α^i is the weight and m corresponds to the means of our variables. We can associate a simple cokriging system to this estimator as $C_{ij}w_i = c_{ii_0}$, where C_{ij} is the M×M covariance matrix, and c_{ii_0} is the M_0×M covariance matrix between the unsampled and sample locations.

Machine learning

The following table describes the selected machine learning (ML) algorithms to be used in the SALMANTICOR study.

Algorithm	Type	Description
Random Forest	Combine methods	Classification ensemble through a combination set of non-correlated independently decision trees
Gradient Boosting	Combine methods	Ensemble technique in which decision trees are not independently, but sequentially

Algorithm	Type	Description
Logistic regression	Regression	The go-to method for categorical or binary classification
K-nearest Neighbors	Supervised classification	Classifies each unlabeled example by the majority label among its k-nearest neighbors in the training set
Support Vector Machine	Supervised classification	Classification and regression technique through construction of separating hyperplanes to maximize the margin and to produce a generalization ability
Linear discriminant analysis	Linear discriminant	Searches for directions in the data that have the largest variance and subsequently project the data onto it combining Fisher vectors
Naive Bayes classifier	Probabilistic supervised classification	The Bayesian classification is used as a probabilistic learning method

STROBE statement SALMANTICOR

STROBE Statement—checklist of items that should be included in reports of observational studies

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract: Population-based study
		(b) Provide in the abstract an informative and balanced summary of what was done and what was found: A cross-sectional survey of randomly selected residents of Salamanca (Spain). 2400 individuals, stratifies by age and sex and by place of residence (rural and urban) will be studied. The variables to analyze will be obtained from the clinical history, different surveys including social status, Mediterranean diet, functional capacity, electrocardiogram, echocardiogram, VASERA and biochemical and genetic analysis.
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported: pages 8-9
Objectives	3	State specific objectives, including any prespecified hypotheses: page 10
Methods		
Study design	4	Present key elements of study design early in the paper: The SALMANTICOR study is a cross-sectional descriptive population-based study of the prevalence of structural heart disease and their risk factors that will enroll a total of 2400 individuals, stratifies by age, sex and by place of residence (rural and urban), in a Spanish community: Salamanca
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection: pages 11-17

Participants	6	(*a*) *Cohort study*—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study*—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants: Individuals aged ≥18 years included in the lists of all primary healthcare facilities of the province of Salamanca represented the reference population
		(*b*) *Cohort study*—For matched studies, give matching criteria and number of exposed and unexposed *Case-control study*—For matched studies, give matching criteria and the number of controls per case
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable: The SALMANTICOR study is designed to provide echocardiographic parameters characterizing cardiac structure and function in all individuals. SALMANTICOR participants will undergo surveillance for cardiovascular events, including heart failure, incident coronary heart disease, and all-cause mortality.
Data sources/ measurement	8 *	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group: pages 11-16 and tables
Bias	9	Describe any efforts to address potential sources of bias: Spain´s and consequently Salamanca healthcare system is public, guaranteeing universal coverage. In total, 98.7 percent of the population are insured for this public Spanish healthcare system. In Salamanca, a total of 35 primary health centers throughout the province provide healthcare services to the overall population: 18 to the urban-considered municipalities and 17 to

		the rural-considered municipalities (Figure 1). Individuals aged ≥18 years included in the lists of all primary healthcare facilities of the province of Salamanca represented the reference population of 295,975 subjects: mean age 52.9±19.8 years; 52.4% females; 61.3% residing in urban areas
Study size	10	Explain how the study size was arrived at: A sample size of 2400 subjects is calculated based on an expected prevalence of structural heart disease of 6% with a confidence interval of 95% and a 1% precision. In order to obtain the necessary sample size, 35% more requests for participation will be made, estimating errors of location from the healthcare database or refuses to participate in the study. Thus, 3564 people will be randomly selected from the primary care lists.
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why: pages 16-17
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding: pages 16-19
		(*b*) Describe any methods used to examine subgroups and interactions: pages 16-19
		(*c*) Explain how missing data were addressed: pages 16-19
		(*d*) *Cohort study*—If applicable, explain how loss to follow-up was addressed *Case-control study*—If applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy: pages 16-19
		(*e*) Describe any sensitivity analyses

Continued on next page

Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	
		(b) Give reasons for non-participation at each stage	
		(c) Consider use of a flow diagram	
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	
		(b) Indicate number of participants with missing data for each variable of interest	
		(c) *Cohort study*—Summarise follow-up time (eg, average and total amount)	
Outcome data	15*	*Cohort study*—Report numbers of outcome events or summary measures over time	
		Case-control study—Report numbers in each exposure category, or summary measures of exposure	
		Cross-sectional study—Report numbers of outcome events or summary measures	
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	
		(*b*) Report category boundaries when continuous variables were categorized	
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	
Discussion			
Key results	18	Summarise key results with reference to study objectives: pages 20-24	
Limitations	19	Discuss limitations of the study, taking into account sources of potential	

	9	bias or imprecision. Discuss both direction and magnitude of any potential bias. pages 16-19
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence. pages 16-19
Generalisability	21	Discuss the generalisability (external validity) of the study results. pages 16-19
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based.

This study was supported by ~~by~~ national (PI14/00695, Institute of Health Carlos III, Spanish Ministry of Economy and Competitiveness) and community (GRS1030/A/14, SACYL, Junta Castilla y León) competitive grants and by the Spanish Cardiovascular Network (RIC and CIBERCV) from the Institute of Health Carlos III, Spanish Ministry of Economy and Competitiveness, Obra Social "la Caixa" and Philips Ibérica Healthcare division.

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.