Received April 22, 2021, accepted May 11, 2021, date of publication May 25, 2021, date of current version June 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3083523

Data Science Analysis and Profile Representation Applied to Secondary Prevention of Acute Coronary Syndrome

ANTONIO GARCÍA-GARCÍA[1,2], IGNACIO PRIETO-EGIDO[1], ALICIA GUERRERO-CURIESES[1], JUAN RAMÓN FEIJOO-MARTÍNEZ[1,3], SERGIO MUÑOZ-ROMERO[1,4], SERGIO MANZANO FERNÁNDEZ[5], PEDRO JOSÉ FLORES-BLANCO[5], JOSÉ LUIS ROJO-ÁLVAREZ[1,4], (Senior Member, IEEE), AND ANDRÉS MARTÍNEZ-FERNÁNDEZ[1]

[1] Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, 28943 Madrid, Spain
[2] Servicio de Sistemas de Información, Hospital Infantil Universitario Niño Jesús, 28009 Madrid, Spain
[3] Red Eléctrica de España, 28109 Madrid, Spain
[4] Persei Vivarium, 28013 Madrid, Spain
[5] Hospital Clínico Universitario Virgen de la Arrixaca, Unidad de Arritmias, 30120 Murcia, Spain

Corresponding author: Antonio García-García (antonio.ggarcia@urjc.es)

This work was supported in part by the Agencia Estatal de Investigación under Research Grant BigTheory PID2019-106623RB-C41, and in part by the Fondo Europeo de Desarrollo Regional (FEDER).

ABSTRACT The analysis of large amounts of data from electronic medical records (EMRs) and daily clinical practice data sources has received increasing attention in the last years. However, few systematic approaches have been proposed to support the extraction of the wealth and diversity of information from these data sources. Specifically, Acute Coronary Syndrome (ACS) data are available in many hospitals and health units because ACS shows elevated morbidity and mortality. This work proposes a method called Data Science Analysis and Representation (DSAR) to scrutinize and exploit, in a univariate way, scientific information content in limited ACS samples. DSAR uses Bootstrap Resampling to provide robust, cross-sectional, and non-parametric statistical tests on categorical and metric variables. It also constructs an informative graphical representation of the database variables, which helps to interpret the results and to identify the relevant variables. Our objectives were to validate DSAR by comparing it to conventional statistical methods when looking for the most relevant variables in the secondary prevention of ACS, and to determine the degree of correlation between them and the Exitus event (associated with patient death). To achieve this objective, we applied DSAR on an anonymized sample of 270 variables from 2377 patients diagnosed with ACS. The results showed that DSAR identified 44% significant variables while conventional methods offered weak correlation results. Then, the scientific literature was reviewed for a set of these variables, validating the agreement with clinical experience and previous ACS research. The conclusion is that DSAR is a valuable and a useful method for clinicians in the identification of potentially predictive variables and, overall, a good starting point for future multivariate secondary analyzes in the clinical field of ACS, or fields with similar information characteristics.

INDEX TERMS Acute coronary syndrome, data Science, health information systems, profile representation, bootstrap resampling, health statistics, secondary prevention, multivariate analysis, univariate analysis.

I. INTRODUCTION

We have a new scenario with information and communication technologies continually advancing toward better and faster solutions with limitless possibilities, the like of which we have not seen before. Statistical learning combined with data availability has the potential to change how medical information is processed and treatments are applied [1]. Electronic Health (eHealth) has been mainstreamed into the Healthcare sectors and it is transforming health systems by creating new market opportunities for different technology companies. Since 2005, the growth of eHealth solutions has been

The associate editor coordinating the review of this manuscript and approving it for publication was Kin Fong Lei.

mainly implemented as Electronic Medical Records (EMRs), composed by the union of Electronic Health Records (EHRs) and Health Information Systems (HIS), and it has increased its figures, thus becoming the main interaction model used by global Healthcare stakeholders. However, eHealth solutions are not adapted to the daily complexity of clinical practice and they still lack the integration required to truly exploit the use of available clinical data. As a result, more specialized departments are resorting to alternative *ad-hoc* software tools, so that today these complementary solutions coexist in many departments. Still, none of them represents a complete, easy to use, and integrated alternative to generate clinical insights to improve clinical practice from all the available hospital data [2].

Therefore, a critical knowledge gap can be observed from the translation of clinical practices into research, and from research into practices. Health Care Professionals (HCPs) want to understand how to treat their patients better so they can achieve the best health outcomes. The current tools available in the eHealth landscape do it by simple integration of the data available and they are often limited to operative analysis. However, they are unable to analyse large amounts of data strategically with easy-to-use tools provided to the clinicians. Hence, hospitals and healthcare professionals are currently missing essential opportunities for the advancement of improved treatments. Some limitations in the current tools are the following. First, the difficulty-of-use because the existing medical solutions are mostly static, in the sense that the information provided cannot be easily exploited. Added to this, we find the limited time of health professionals, making it difficult to improve care or to increase knowledge about the pathology [3]. Advanced tools from Machine Learning are not addressing the specific questions that HCPs want to answer. The potential of Artificial Intelligence (AI) technologies is excellent, but they are not yet useful at the user level and they take time to be fluently applicable in this setting. Furthermore, the incompleteness of data sources brings other limits, and despite the standardization efforts, current medical tools are still missing much essential information. Said information cannot be contained in EHRs or dedicated tools in some cases, whereas it could help to generate better knowledge about patients, therapies, and treatments. This is a consequence also derived from the extremely slow evolution of these solutions, which in many cases is associated with interoperability limitations. Overall, few principled approaches have been proposed for Medical Departments to generate domain-specific knowledge and information models capable of linking heterogeneous clinical information [4].

On the other hand, Cardiovascular Disease (CVD) (which include coronary heart disease and stroke) are the most common non-communicable diseases globally, responsible for an estimated 17.8 million deaths in 2017, of which more than three quarters were in low-income and middle-income countries [5]. Its most common expression is the Acute Coronary Syndrome (ACS), which represents a group of clinical entities with a common denominator, namely, the total or partial obstruction of an artery by a thrombus [6]. There is a significant variability among studies and official statistics in terminology, definition, and condition when evaluating the impact of coronary illness in a population. The reason for this is that it is hard to control the different cardiovascular risk factors, including arterial hypertension, diabetes mellitus, age, and many others [7]. CVD is one of the many medicine fields where a large number of EHRs are generated worldwide each day into routine clinical practice.

In this work, we propose a new system called Data Science Analysis and Representation (DSAR), consisting of two main elements to scrutinize and scientifically exploit complex healthcare databases information. First, DSAR uses Bootstrap Resampling in order to provide us with non-parametric statistical tests on categorical and continuous variables which are similar to handle. Second, DSAR offers a graphical and statistically principled representation of these variables to identify the regions where relevant features are present. The system is evaluated on an existing representative sample for ACS secondary prevention, consisting of patient recordings and follow-up information with features and fields of different and heterogeneous nature and origin.

The scheme of the paper is as follows. In Section II, the relevant background and related works summarize both the Data Science precedents and the ACS medical scenario. In Section III, the research hypotheses are formulated and the methods and material are presented. Then, in Section IV, the results for all the contributions are showed. Finally, in Sections V and VI, discussion is presented and conclusions are drawn, respectively.

II. BACKGROUND AND RELATED WORKS

A. DATA SCIENCE AND RELATED PRECEDENTS

Clinical risk prediction based on data analysis has been recognised as a useful tool for managing disease care and treatments. They constitute a promising technology nowadays to extract relevant information in medical therapeutics. In [8], the use of a large volume of heterogeneous data was proposed to build a risk prediction model for ACS. Similarly, major adverse cardiac events were also investigated in [9] using a large volume of clinical EHR with a multitask learning approach. Neural Networks are likewise employed [10], [11] to predict adverse outcomes for ACS, and they demonstrated to be successful in predicting potential ACS patients.

As in many other fields, Data Science (DS) and some related technologies offer a promising field to model clinical profiles. An increasing number of proposals scrutinize DS techniques for healthcare purposes, and their impact is just starting to provide the clinical professionals with effective diagnosis support. A glimpse of the DS potential for disease treatment has been reported [12]–[14]. Examples of new developments are going far beyond classical prediction approaches, and specific medical diagnosis applications are emerging [15].

B. ACUTE CORONARY SYNDROME IN CLINICAL

Within CVD discipline, and citing the Clinical Practice Guidelines of the European Society of Cardiology [16], [17], the clinical presentation of ACS is broad. It ranges from cardiac arrest, electrical or haemodynamic instability with cardiogenic shock (CS) due to ongoing ischaemia or mechanical complications such as severe mitral regurgitation, to patients who are already pain free again at the time of presentation. The leading symptom initiating the diagnostic and therapeutic cascade in patients with suspected ACS is acute chest discomfort described as pain, pressure, tightness, and burning. Chest pain-equivalent symptoms may include dyspnoea, epigastric pain, and pain in the left arm. Based on the electrocardiogram (ECG), two groups of patients can be differentiated:

- Patients with acute chest pain and persistent (>20 min) ST-segment elevation. This condition is termed ST-segment elevation ACS and generally reflects an acute total or partial coronary occlusion. Most patients will ultimately develop ST-segment elevation myocardial infarction (STEMI). The mainstay of treatment in these patients is immediate reperfusion by primary percutaneous coronary intervention (PCI) or, if not available in a timely manner, by fibrinolytic therapy [17].
- Patients with acute chest discomfort but no persistent ST-segment elevation [non-ST-segment elevation ACS (NSTE-ACS)] exhibit ECG changes that may include transient ST-segment elevation, persistent or transient ST-segment depression, T-wave inversion, flat T waves, or pseudo-normalization of T waves, or the ECG may be normal.

The pathological correlate at the myocardial level is cardiomyocyte necrosis (non-ST-segment elevation myocardial infarction (NSTEMI)) or, less frequently, myocardial ischaemia without cell damage (unstable angina). A small proportion of patients may present with ongoing myocardial ischaemia, characterized by one or more of the following: recurrent or ongoing chest pain, marked ST-segment depression on 12-lead ECG, heart failure, and haemodynamic or electrical instability [16]. Due to the amount of myocardium in jeopardy and the risk of developing CS and/or malignant ventricular arrhythmias, immediate coronary angiography and, if appropriate, revascularization are indicated.

The recommendations for acute (and long-term treatment in patients who have suffered an ACS are established in the aforementioned clinical guidelines of different cardiology and cardiovascular societies [16], [17]. Compliance with these recommendations at the time of patient discharge contributes to reducing mortality rates [18]. In Spain, there are known specific registries of patients with ACS, such as: DESCARTES, a Spanish registry of ACS descriptions made in 2002; MASCARA, which was carried out in 2005 with patients from different Spanish centers; Or GYSCA, published in 2010 [19].

III. HYPOTHESIS AND METHODS

Combining DS tools with existing ACS databases can provide new insights into ACS risk factors and it can also contribute to the continued reduction in mortality associated with this condition. However, as we have already mentioned, we need easy-to-use tools that offer easily interpretable information. DSAR meets those requirements, but before applying it in the clinical practice, it is necessary to validate the results provided by these methods. This is the objective of the present work, and based on it, the research hypotheses are defined below.

Therefore, we propose two research hypotheses regarding the DSAR system in its application to limited samples of patients diagnosed with ACS, which can be stated as follows:

- **Hypothesis 1.** The DSAR system can identify the variables, both metric and categorical, that are significant in relation to the variable Exitus (patient mortality) when the patient sample does not allow us to get results using conventional correlational statistical methods.
- **Hypothesis 2.** The DSAR system is sensitive enough to not discard too many significant variables and specific enough for limiting false positives, if we take as reference what the scientific evidence in the literature shows on patient mortality.

Hypothesis 1 will be verified by contrasting the DSAR results with those ones offered by conventional statistics, while Hypothesis 2 will be verified by comparing the results offered by DSAR with those ones that we have obtained by reviewing the scientific literature with larger samples.

The following subsection explains both the methods used to carry out the conventional statistical analysis and the methods carried out by our DSAR system. Finally, the sample is described in depth.

A. STATISTICAL METHODS

1) CONVENTIONAL-BASED CORRELATIONAL ANALYSIS

Feature selection is one of the main methods proposed in the literature to find the most important or useful variables from raw data in order to improve data mining tasks. In most medical applications of data mining, the first objective is to recognize less important variables that are unnecessary, irrelevant, or even distracting for the aforementioned goals. Deleting these variables reduces the high dimension of a dataset, and therefore it makes possible to build more easy-to–interpret models.

Our study describes two main types of variables, which we need to study and to relate, namely, metric and categorical variables. Up to date, the methods based on correlation analysis are among the most broadly used in the literature [20]. The conventional statistical methods applied in this work, as well as their graphic representation, are detailed in Appendix 1.

2) DSAR FUNDAMENTALS

Considering diagnosis problems as detection problems allows us to establish simple tests to analyze the statistical differences between two groups of patients, and in this way we can readily establish a vector representation supporting the detection of statistical differences in the features of a given sample. For this aim we use the Bootstrap Resampling, which is a powerful statistical tool, which was introduced by Efron [21] as a general method to quantify the uncertainty associated with a given estimator (e.g. means, standard deviations, or confidence intervals). The key idea is to perform iterative calculations of the statistics on the same dataset to estimate their resulting variation. Thus, the variability or estimators can be obtained without needing additional data, but from distinct resamples of the dataset. These bootstrap datasets are created by sampling with replacement, each one with the same size as the original dataset, so that some observations may appear more than once, and some others not at all, in a given bootstrap resample.

Suppose we have n data points, denoted by $x_1, x_2, \ldots, x_n$, and drawn from a distribution F. An empirical bootstrap resample is a set of observations of the same size n (denoted as $x^*{}_1, x^*{}_2, \ldots, x^*{}_n$). For any statistic v computed from the original sample data, we can define a replicated statistic v^* computed instead using the resampled data. Due to the variation of the statistic v^* will depend on the sample size, and in order to approximate this variation, resamples of the same size must be used. The bootstrap setup is as follows:

1) $x_1, x_2, \ldots, x_n$ is a data sample drawn from a distribution F.
2) v is a statistic computed from the sample.
3) $x^*{}_1, x^*{}_2, \ldots, x^*{}_n$ is a resample of the data of the same size as the original sample n.
4) F^* is the empirical distribution of the data (the resampling distribution).
5) v^* is the statistic computed from the resample.
6) For each bootstrap sample, we compute the bootstrap differences in the statistic: $\delta^* = v^* - v$.
7) Steps 3 to 6 are repeated a sufficient number of times (1000 is recommended).
8) According to the plug-in principle, F^* is equivalent to F and the variation of v is well-approximated by the variation of v^*. We will use this to estimate the size of confidence interval. E.g. we estimate the 90% bootstrap confidence interval for v as: $[v - \delta^*{}_{.05}, v - \delta^*{}_{.95}]$, where $\delta^*{}_{.05}$ and $\delta^*{}_{.95}$ denotes the 5% lowest and biggest values respectively.

Because the DSAR method can be used with both metric and categorical variables, we start by stating the statistical notation on metric variables. If M_j is a metric variable and its probability density function (*pdf*) is denoted as $f_{M_j}(M_j)$ [22], then we use some convenient criteria to establish two groups of the events, namely, G_1 and G_2 (in this study, the survival and death groups respectively). Then, the conditional distributions for this variable fulfill

$$f_{M_j}(M_j) = P(G_1) f_{M_j}(M_j|G_1) + P(G_2) f_{M_j}(M_j|G_2), \tag{1}$$

where $P(G_1)$ and $P(G_2)$ are the *a priori* probabilities of the patients in each group. Each conditional density has its own distribution parameters, and without loss of generality, we consider their conditional mean, deviation, and *pdf* shape denoted by:

$$m_j^{G_1}, \qquad m_j^{G_2}, \tag{2}$$

$$\sigma_j^{G_1}, \qquad \sigma_j^{G_2}, \tag{3}$$

$$f_{M_j}(M_j|G_1), \quad f_{M_j}(M_j|G_2). \tag{4}$$

We can define their differences in both groups so that they can be used as statistic measurements (v in the previous bootstrap setup), as follows:

$$\Delta m_j = m_j^{G_1} - m_j^{G_2}, \tag{5}$$

$$\Delta \sigma_j = \sigma_j^{G_1} - \sigma_j^{G_2}, \tag{6}$$

$$\Delta f_{M_j} = f_{M_j}(M_j|G_1) - f_{M_j}(M_j|G_2), \tag{7}$$

and now the bootstrap principle can be applied to estimate the distribution of the previous measurements to make statistical tests to detect significant differences in the two event groups.

As an example, figure 1 shows in the graph on the right the Δm and the $\Delta\sigma$ representation for the metric variable Urea (UA). The vertical yellow line indicates the zero, showing a significant difference in the mean in favor of the G2 group, which can be interpreted as that with increasing Urea levels there is a prevalence in Exitus. Furthermore, the analysis of the standard deviation, as a complementary tool, helps us to understand the differences in the distribution of the variable for both groups. In this case, it is observed that the standard deviation is negative, which implies a greater distribution of the variable in favor of group G2. On the other hand, in the two graphs on the left, the normalized *pdf* shape forms are represented for both groups (*npdf*) with their difference (Δ*npdf*), where we can see that it is established a prevalence in the proportion of Non Exitus in Urea levels below 50 and a prevalence in the proportion of Exitus beyond that level.

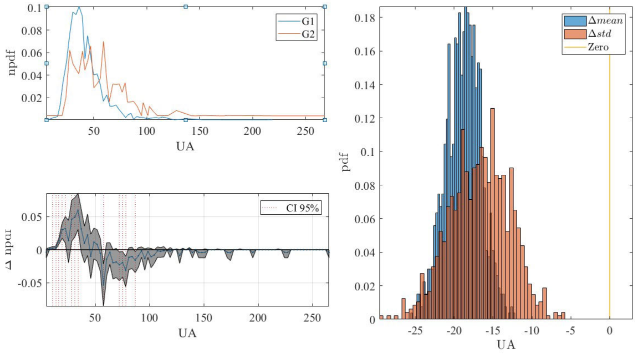


FIGURE 1. DSAR applied to metric variable urea.

We continue with categorical variables, which we denote by C_j. These variables can have a set of possible values or categories among a discrete set, which is $C_j.value = \{c_k^j, k = 1, \cdots, K_j\}$, where K_j is the number of possible categories for variable C_j. The probability mass density function (*pmf*) of that categorical variable is given by $P(c_k^j)$, which can be seen as the proportion of this category in a finite set of observed events. If we again consider two groups, the conditional probabilities for that variable are as follows,

$$P(c_k^j|G_1), \quad P(c_k^j|G_2), \tag{8}$$

and we can define a convenient statistic with their *pmf* differences, given by

$$\Delta P(c_k^j) = P(c_k^j|G_1) - P(c_k^j|G_2), \tag{9}$$

and apply bootstrap to achieve statistical test to detect significant differences. The differences in all the categories of this variable can be grouped in a feature vector, given by

$$\Delta \boldsymbol{p}_j = [\Delta P(c_1^j), \cdots, \Delta P(c_{K_j}^j)]^\top. \tag{10}$$

As an example, figure 2 shows the estimated distribution for Chronic Obstructive Pulmonary Disease (COPD) categorical variable, whose values are represented by COPD $= 0$ without the existence of COPD in the patient and COPD $= 1$ with existence of this. This plot is normalized to unit maximum amplitude, which is convenient for visualization purposes. All the categories are sorted with decreasing value of their Δp according to the estimated distribution in the representation. In accordance, we can have three possible situations for each distribution. First, if Δp overlaps zero, this means that the probability of (G_1) values is similar to the probability of G_2 ones, so these regions have no particular bias associated with the category. Second, a Δp with its *pdf* significantly located at negative values means that these values have a significant proportion in the Exitus group (in this example, the existence of COPD or COPD $= 1$). And third, Δp with its *pdf* significantly located at positive values means that these values have a significant proportion in the No Exitus group (in this example, the non-existence of COPD or COPD $= 0$).

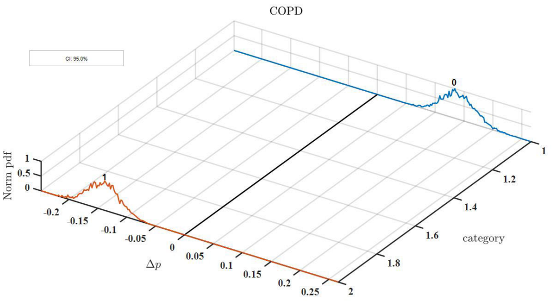


FIGURE 2. **DSAR applied to categorical variable COPD.**

Note that, again for intuitive representation purposes, each *pdf* is plotted in thicker line style, and also the category label is displayed at the top of the *pdf* for those significant categories, hence allowing us to straightforwardly scrutinize the most relevant categories for each feature.

Figures 1 and 2 give a detailed amount of statistical information. Nevertheless, its visualization is specific for clinicians (or other professionals without advanced knowledge of statistics) interested in making decisions or further analysis on relevant clinical data to be inferred from the patient. This is because we have a variety of data types in different forms.

Further advantage can be taken from the statistical notation and methods previously established using Chromosome Representation, which is a graphical tool putting together and at a glance the relevance of sets of features, as explained below. Note that despite of this name, it is no related with any biological representation. Let us assume that we have J_M metric variables, and the statistical tests for the j^{th} feature can be summarized in a vector,

$$\Delta \boldsymbol{r}_j = [\Delta m_j, \Delta\sigma_j], \tag{11}$$

where m_j is the mean and σ_j is the standard deviation. We also assume that we have J_C categorical variables, and all the occurrences of the j^{th} feature are represented in vector Δp_j. Then, we can define the chromosome representation of this form as the concatenation of vectors with the statistical difference markers through all the metric and categorical features, as follows:

$$\Delta \boldsymbol{u} = \Bigg[\bigcup_{j=1}^{J_M} \Delta \boldsymbol{r}_j, \bigcup_{j=1}^{J_C} \Delta \boldsymbol{p}_j\Bigg], \tag{12}$$

where $\bigcup$ denotes the row vector concatenation operator.

As an example, figure 3 shows the chromosome representation of the metric variable Urea (UA), represented by its vector $\Delta \boldsymbol{r}$ (composed by Δ*mean* and Δ*std*) and the categorical variable COPD represented by its vector $\Delta \boldsymbol{p}$ (composed of the differences of each of the occurrences of the variable: 0 and 1). The central horizontal line represents zero, so this graph allows us to appreciate in a grouped way, both for the categorical variables and for the metric ones, the significance in the distribution of each of these variables with respect to the aforementioned groups G1 and G2 of the threshold variable. The circle at the end of the vector identifies the significant values of the variable. Also note that the distribution of metric variables are represented in red and that of categorical variables in blue.

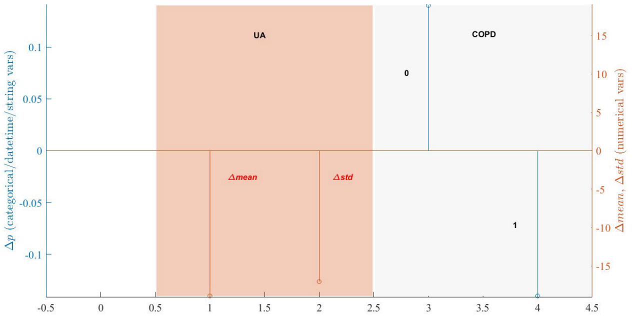


FIGURE 3. **Chromosome representation urea and COPD.**

Accordingly, this chromosome representation gives an overview of the relevance of each feature in the form, which can be reinforced by using the joint graphical representation of the bootstrap significance tests previously calculated.

The *pdf* shape in metric features is not included in the chromosome, as previous works show the convenience of analyzing them separately.

The DSAR method is developed and implemented with Matlab 2020b software under the license of Mathworks company.

B. DESCRIPTION OF THE DATA SAMPLE

Our initial sample consists of a set of 2585 anonymized patient episodes with a diagnosis of ACS. Information on these episodes was collected in a retrospective cohort study from the different and heterogeneous HIS belonging to Virgen de Arrixaca University Clinical Hospital, in the Murcia region in Spain. This sample includes patients who have suffered an ACS episode between 2011 and 2015, who were followed up for a year at discharge. The study was presented and approved by the Ethical Committee thus ensuring fulfilment with current legislation, all the necessary consent were assembled for the SCA recordings and data were anonymized for subsequent analysis by data processors.

In the original sample, each EHR contained 73 metric features and 275 categorical features (plus death/survival class). After discarding those features with more than 75% of not applicable or null values, due to the clinical context of the ACS follow-up itself, a final sample of 2377 patient samples with 270 features was obtained.

This sample is complex, very sparse, and with a lot of missing values. Moreover, the sample is imbalanced, with 246 $(10, 3\%)$ record from patients who died during their hospitalizations or after discharge, and the others 2131 $(89, 7\%)$ records from patients who survived after the ACS episode.

The features are structured in two datasets regarding the timeline of the ACS event:

- **Dataset 1, previous features:** This dataset includes the variables of both health antecedents and treatments prescribed in the patient before suffering the ACS event. This dataset is composed of a total of 79 variables.
- **Dataset 2, diagnostic features:** This dataset includes the variables with the diagnosis and pertinent interventions of the patient upon admission caused by the ACS event. This dataset is composed of a total of 191 variables.

And in order to facilitate a detailed and in-depth study of the results, the two datasets are disaggregated into Subsets determined by clinical criteria, as follows:

- **Dataset 1: Previous Features.**

Subset 1.1: Personal History. Variables related to the clinical history of the patient previous to the ACS event. This subset is composed of 40 variables.

Subset 1.2: Previous Treatment. Variables related to the treatment prescribed to the patient prior to the ACS event. This subset is composed of 39 variables.

- **Dataset 2: Diagnostic Features**

Subset 2.1: Exploration and Actual Disease. Variables related to the examination of the patient at the time of the ACS event. This subset is composed of 14 variables.

Subset 2.2: Complementary Tests. Variables related to the tests performed on the patient at the time of the ACS event. This subset is composed of 28 variables.

Subset 2.3: Laboratory. Variables related to patient laboratory determinations at the time of the ACS event. This subset is composed of 31 variables.

Subset 2.4: Interventionism. variables related to coronary revascularization treatment, that includes both percutaneous coronary intervention, myocardial revascularization surgery and fibrinolysis, performed on the patient during their hospitalization due to the ACS event. This subset is composed of 56 variables.

Subset 2.5: Hospital Evolution. Variables related to the evolution of the patient during his hospitalization due to the ACS event. This subset is composed of 62 variables.

C. RESEARCH METHODOLOGY

The following methodology was applied to evaluate both hypotheses:

- **Evaluation of Hypothesis 1:** This evaluation was carried out by applying both the DSAR method and the conventional statistical methods for all the variables of Datasets 1 and 2, taking the Exitus variable as a reference in the univariate analysis, and subsequently comparing the results.
- **Evaluation of Hypothesis 2:** This review was carried out by taking a group of variables of known diagnostic value in the field of ACS and comparing their scientific review in large samples with the results obtained by applying the DSAR method.

IV. RESULTS

A. IDENTIFICATION OF SIGNIFICANT VARIABLES

In this section, we first analyze if DSAR can obtain the significant variables in relation to the variable Exitus. When applied to the sample, DSAR identifies 117 (43%) variables as statistically significant, 37 corresponding to Dataset 1 and 80 to Dataset 2. The results for each Subset are detailed below:

- **Dataset 1: Previous Features.**

Subset 1.1: Personal History. For a total of 40 variables, DSAR identifies 17 as significant and 23 as non-significant. For this subset, examples of variables with significant results are age and sex, and examples of non-significant variables are use of cocaine or pre-existence of HIV.

Subset 1.2: Previous Treatment. For a total of 39 variables, DSAR identifies 20 as significant and 19 as non-significant. For this Subset, examples of variables with significant results are previous treatment with clopidogrel or betablockers, and examples of non-significant variables are previous treatment with B12 Vitamin or allergy treatments.

- **Dataset 2: Diagnostic Features**

Subset 2.1: Exploration and Actual Disease. For a total of 14 variables, DSAR identifies 7 as significant and 7 as non-significant. For this Subset, examples of variables with significant results are Body Surface Area (BSA) or the diagnosis upon admission, and an example of non-significant variables are weight and height treated independently.

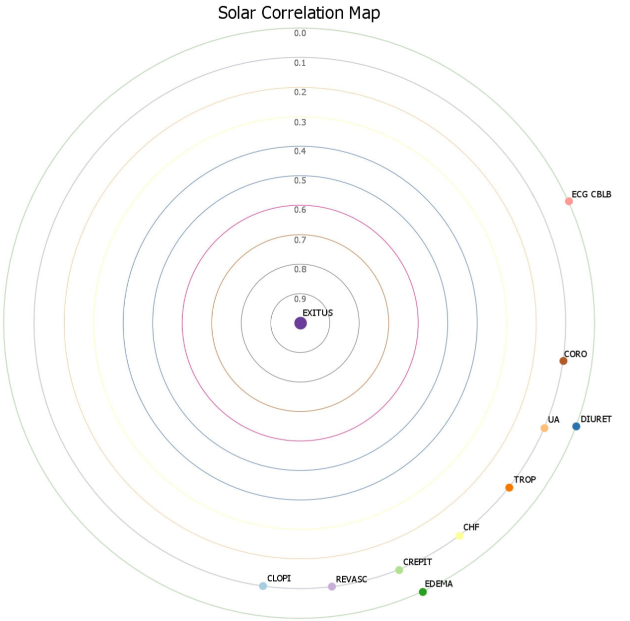


FIGURE 4. **Solar map of the 10 variables with the highest association coefficients.**

Subset 2.2: Complementary Tests. For a total of 28 variables, DSAR identifies 24 as significant and 4 as non-significant. For this Subset, examples of variables with significant results are Complete Block of left and right branch and Incomplete Block of left branch, and an example of a non-significant variable is a Incomplete Block of right branch.

Subset 2.3: Laboratory. For a total of 31 variables, DSAR identifies 6 as significant and 25 as non-significant. For this Subset, examples of variables with significant results are creatinine and urea and an example of a non-significant variable is cholesterol values.

Subset 2.4: Interventionism. For a total of 56 variables, 29 significant and 27 non-significant are obtained. For this Subset, examples of variables with significant results are the different locations of the coronography intervention or types of stents implanted in an intervention, and examples of non-significant variables are certain metric values on said stents, such as length and diameter.

Subset 2.5: Hospital Evolution. For a total of 62 variables, 16 significant and 46 non-significant are obtained. For this Subset, examples of variables with significant results may be the hospital events, such as inotropic use, and an example of a non-significant variable is local complications at vascular access.

The second step is to apply conventional statistical methods to the sample. figure 4 shows the solar map with the 10 variables that obtain the highest relationship measures using these methods. As can be seen, the obtained correlations are low, ranging between 0 and 0.1 in absolute value.

As can be seen, DSAR identifies almost half of the variables as significant, while the analysis by the used conventional methods does not attribute significance to any variable.

These results show the potential of DSAR to identify significance when conventional methods are inconclusive. The following section is devoted to validate DSAR results with other strategies.

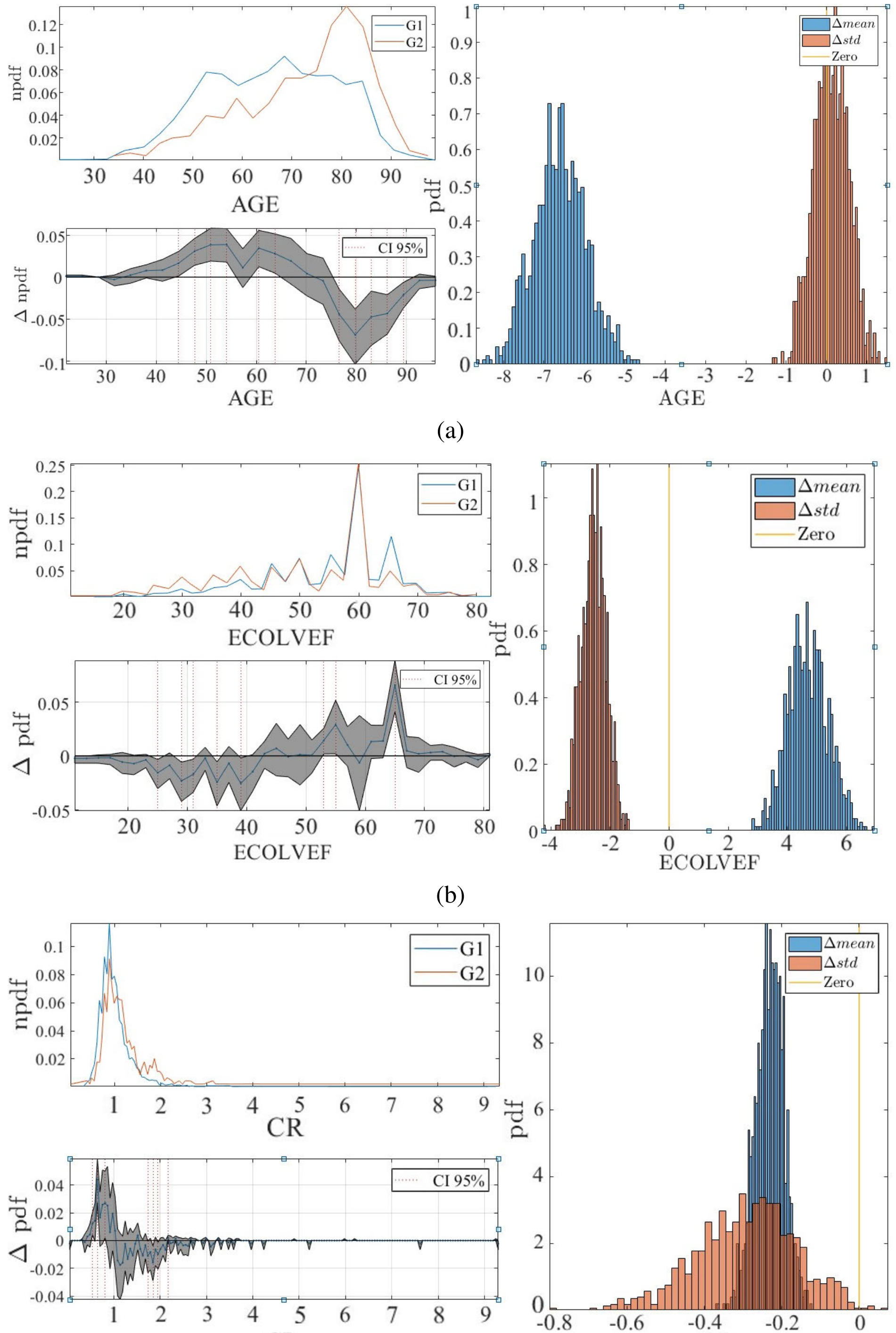
B. CONTRASTING DSAR RESULTS

The objective of this section is to verify that the results obtained by the DSAR system are consistent with the scientific evidence and literature related to ACS. Therefore, the scientific literature is reviewed, finding a subset of variables with marked and independent prognostic value with respect to the death of patients with ACS during one year after diagnosis.

For style reasons, all the variables studied in this section are presented in figures 5 and 6, which group the metric and categorical variables respectively.

When focusing on Dataset 1, the scientific literature establishes that ACS patients with *Arterial Hypertension (AH)* (AH $= 1$) represent a subset of higher risk, with AH having a high prognostic importance [23], [24]. Something similar occurs with patients with previous *Peripheral Vasculopathy (PV)* (PV $= 1$), where patients have a worse prognosis, according to observational studies [25], [26]. Pooled hazard ratios comparing cohorts with *Diabetes Mellitus (DM)* (DM $= 1$) versus without DM (DM $= 0$) were significantly higher in the hospital in DM for death from any cause [27]. It is also evidenced that ACS patients with DM are more likely to suffer from cardiogenic shock and / or kidney failure and die during hospitalization [28]. Regarding discharge, patients with diabetes also have a higher risk of dying, regardless of age and sex, than those without it. Younger patients with diabetes also have a notably higher risk of dying [28]. In the *Neoplasm (NEO)* study associated with ACS [29], prevalent and incident neoplasms (NEO $= 1$) increased the risk of all-cause mortality by 4 times. It is also contrasted that patients with *Chronic Heart Failure (CHF)* who develop ACS (CHF $= 1$), have markedly increased subsequent mortality [30]. This evidence is coherent with the results obtained using DSAR in our sample, whose graphical analysis for former variables is shown in figures 5 and 6. There, a significant proportion is observed concerning Exitus (with negative values in the Δ_p value) for patients who, as previous conditions, suffer from *AH*, *PV*, *DM*, *CHF*, or *NEO*, which are treated as boolean variables where value 1 represents the occurrence. Nevertheless, not having any of these preconditions shows a significant proportion in the non-Exitus population (with positive values in the Δ_p value and the value 0 for the variable).

Regarding *Age (AGE)*, elderly hospitalized with ACS present a higher risk of Exitus [31]–[33]. When analyzing DSAR result a turning point is observed near 78 years.



(a)

(b)

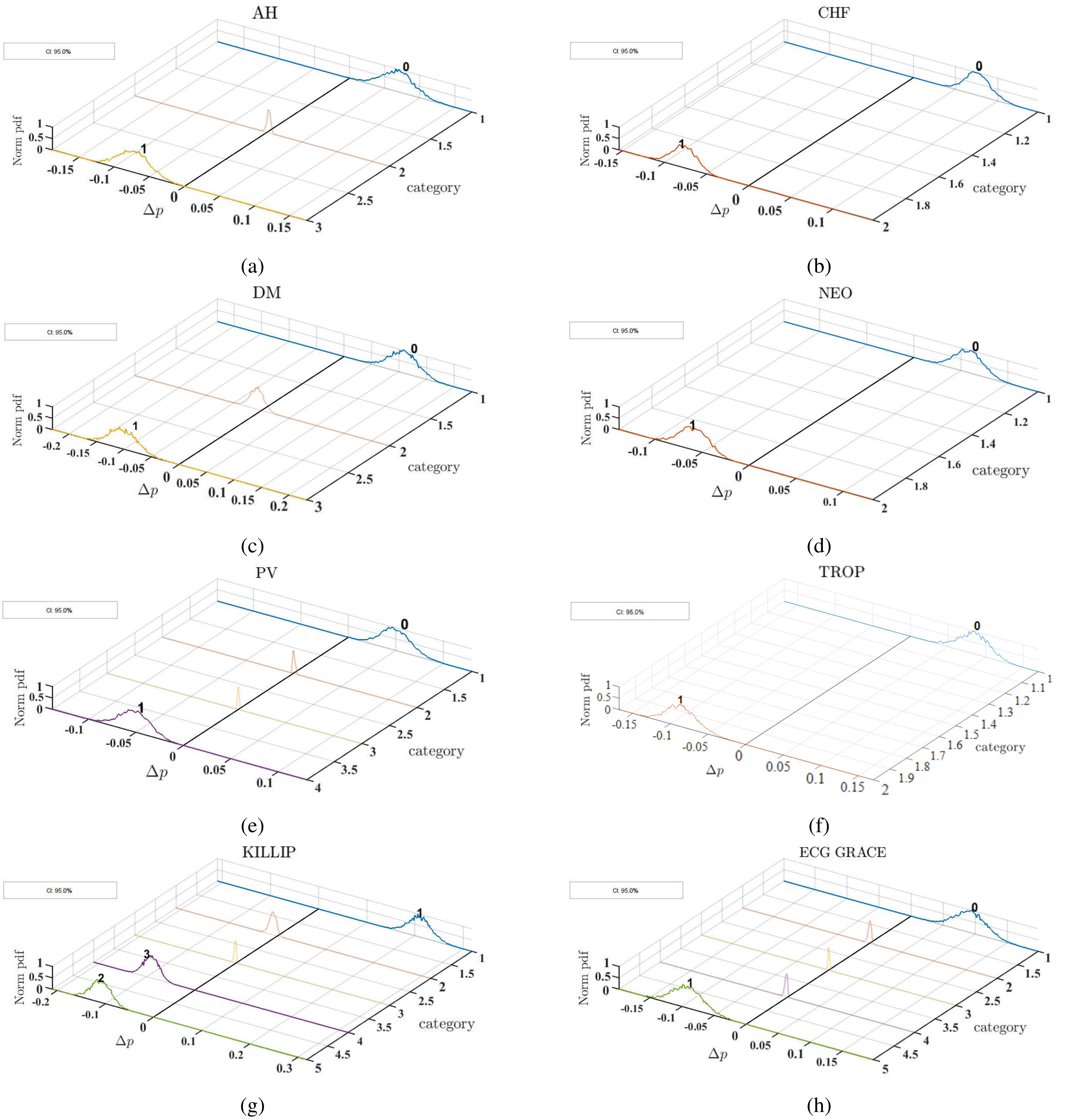
(c)

FIGURE 5. Selected examples of metric variables and their representation in DSAR (see text for details): (a) Age; (b) Left ventricular eyection fraction and (c) Creatinine.

The group under this age shows a significant proportion in the non-Exitus population, while the elderly have a significant proportion in the Exitus population.

A summary of the variables studied for this Dataset 1 is presented with the Chromosome Representation in the figure 7.

When analyzing Dataset 2, the literature shows a scientifically contrasted independent prognostic value for *Left ventricular ejection fraction (ECOLVEF)*, *Creatinine (CR)*, *Troponins (TROP)*, *Killip class (KILLIP)* and *ECG GRACE*. These variables are analyzed in more detail in the following paragraphs.



(a) (b) (c) (d) (e) (f) (g) (h)

FIGURE 6. Selected examples of categorical variables and their representation in DSAR (see text for details): (a) Arterial hypertension; (b) Chronic heart fairlure; (c) Diabetes mellitus; (d) Neoplasm; (e) Peripheral vasculopathy; (f) Troponins; (g) Killip class and (h) ECG GRACE.

ECOLVEF is the proportion of blood pumped out of the heart with each contraction of the left ventricle obtained in an Echocardiogram. Subjects with ECOLVEF <30% or between 30% and 49% had a higher mortality risk than those with ECOLVEF $\geq$ 50%. Observing DSAR results, they are consistent with the literature.

CR is an indicator of renal function and a relevant independent predictor of hospital death and major bleeding in ACS patients [34]. Besides, in-hospital worsening of renal function is a risk factor for 6-month mortality in these patients [35]. DSAR analysis shows an inflection point in the values close to 1 mg/dl, having those below this point a significant proportion in the non-Exitus population. At the same time, the other have a significant proportion in the Exitus population.

Among patients with ACS, even low Troponin T or I levels correlate with an increased risk of death and recurrent ischemic events (TROP $= 1$) compared to patients with Troponin levels below the decision limit (TROP $= 0$) [36].

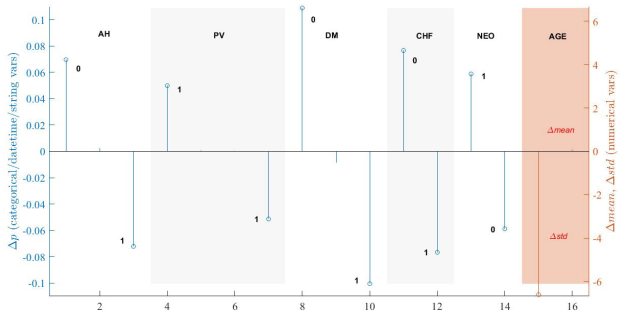


FIGURE 7. Chromosome representation of previous features selected.

Killip class (KILLIP) is a variable that registers a risk assessment classification called Killip. Killip was created in 1967 to apply to individuals with acute myocardial infarction (heart attack). It considers physical examination and the development of heart failure in order to predict and stratify the risk of mortality. Individuals with a low Killip class are less likely to die within the first 30 days after their myocardial infarction than individuals with a high Killip class [37]. Patients presenting ACS and higher Killip class also have poor 1-year survival. In summary, the Killip classification system is a reliable prognostic tool [38].

In our sample, patients were ranked by Killip class in the following way, taking into account that the higher the classification, the worse the prognosis of the ACS patient:

- Killip class I includes individuals with no clinical signs of heart failure.
- Killip class II includes individuals with rales or crackles in the lungs, an S3, and elevated jugular venous pressure.
- Killip class III describes individuals with frank acute pulmonary edema.
- Killip class IV describes individuals in cardiogenic shock or hypotension (measured as systolic blood pressure lower than 90 mmHg), and evidence of peripheral vasoconstriction (oliguria, cyanosis or sweating).

Our sample has the advantage of having a single categorical variable that is defined according to this classification. Analyzing the results of the application of the DSAR method it can be observed that the results are in accordance with the scientific literature, representing a greater risk of death in the Killip II (KILLIP = 2) and Killip III (KILLIP = 3) values and a lower risk in the Killip I value (KILLIP = 1). Note that the non-appearance of the Killip IV value (KILLIP = 4) is due to the fact that it has no significance in the sample (overlaps zero).

Finally, variable ECG GRACE is part of a different risk assessment classification. The *GRACE (Global Registry of Acute Coronary Events)* risk score is a prospectively studied scoring system used to stratify risk in patients with ACS in order to estimate in-hospital and 6-month to 3-year mortality [39]. It is composed of a series of variables to which a weight is assigned depending on its clinical severity. Although it is a multivariate tool, each of the variables is treated independently, and therefore, it can be used to assess the results obtained by DSAR in our sample. In fact, GRACE uses some of the already mentioned variables: *(Age, Killip Class, Arterial Hypertension, Troponin levels and Creatinine)* and confirms the results previously discussed. It also uses the categorical variable *ECG GRACE*, which reports a pathological deviation of the ST segment when performing an electrocardiogram (ECG GRACE = 1) or the absence of this deviation (ECG GRACE = 0). When applying DSAR the results coincide with GRACE.

A summary of the variables studied for this Dataset 2 is presented with the Chromosome Representation in the figure 8.

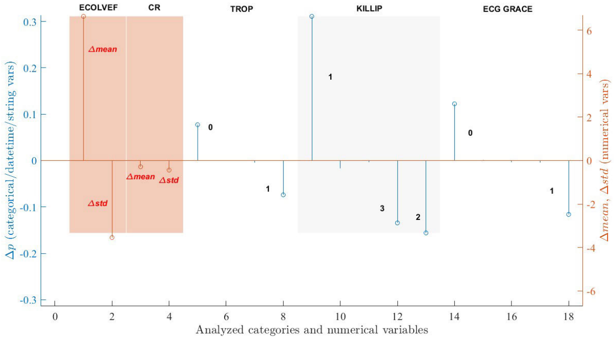


FIGURE 8. Chromosome representation of diagnostic features selected.

To recapitulate, in the previous section, we concluded that DSAR considers 117 variables of our sample as significant concerning the Exitus variable. In this section, we have contrasted with the literature 11 of these variables (6 from Dataset 1 and 5 from Dataset 2).

V. DISCUSSION

We have focused our study in the field of cardiovascular medicine and, more specifically, in the area of secondary prevention of ACS, due to its high prevalence and importance in the field of population health [5]. In this way, we presented, applied and validated the DSAR method in a sample of real anonymized clinical data.

Despite having an unbalanced sample size, this method, compared to other conventional analysis methods, demonstrated the ability to obtain relevant knowledge from univariate analysis. Although this does not necessarily imply causality, the review of some of the scientific literature results has agreed with some DSAR results. Other variables (106) in the SCA recording were not scrutinized and compared with the literature at this point, whereas their relevance should be tested and benchmarked. This is a relevant point to be covered preferably by performing subsequent multivariate analysis, which can be readily performed and interpreted as the univariate analysis proposed here and at the same time being able to give clear statistical relationships among features. Then, we could review the most relevant variable in each group. This interesting path is beyond the scope of the present work but it will deserve particular effort and attention in the near future.

Another contribution of this new method is that it focuses on the graphical presentation of results while providing us with a statistically principled analysis, enabling the review of results and facilitating the co-participation of diverse professionals with Data Science workflow.

Two limitations of our study are that the method is constrained to the application of univariate analysis and the aforementioned unbalance of the sample. Despite these difficulties, the obtained results represent quantitatively and qualitatively an approach with quality in the search for predictors for the ACS follow-up events.

VI. CONCLUSION

Taking into account that today we have a large amount of clinical information that is continuously being collected in the various health information systems, generally heterogeneous, and focused on the individual clinical process, we can conclude that DSAR is a useful statistical method that could help clinicians in better decision-making and that, based on limited data samples, it could provide us with potentially relevant information in order to improve the clinical process.

Likewise, this work opens the door to the need to formalize multivariate analysis methods that allow consolidating unique predictor variables and establishing contrasted dependencies between them, with the aim of improving the prognosis and treatment of ACS. Even further, it is also considered of high interest to transfer these to other areas of clinical knowledge with similar characteristics, that is, on samples based on structured variables with the ultimate goal of improving clinical practice and its shared knowledge.

APPENDIX 1 - CONVENTIONAL STATISTICAL METHODS

Pearson's coefficient [40] is often used in univariate statistics. This index is used to compute the statistical dependencies between two quantitative random variables X and Y for n observed cases in a sample, as follows,

$$\rho(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}, \tag{13}$$

where x_i and y_i are the individual sample points indexed with i, and $\bar{x}$ and $\bar{y}$ denote the sample means of X and Y respectively.

As it can be seen from the above formulation, Pearson's coefficient is a measure of association between two continuous features, but not necessarily between two categorical features or between continuous and categorical. Moreover, it can only identify linear relationships among dimensions and is very sensitive to disturbances in the dataset.

Most of the variables recorded in clinical databases are inherently categorical (i.e prescribed drug groups or hospital admission causes) or integer-valued numerical variables treated as categorical (mainly boolean variables associated to diseases). In both cases, quantitative association measures may be of interest, and the chi-squared test has shown to be reliable for categorical variables.

Although chi-squared is able to determine whether or not there is a statistical relationship between the variables, it can not provide a reliable guide to the strength of that relationship. Cramer's V [41] is a chi square based measure of association which is able to indicate the strength of the relationship between variables. Cramer's V varies from 0 (no association) to 1 (complete association: each variable is completely determined by the other).

Let a sample of size n of the linearly distributed categorical variables X and Y, defining the categories of X as i = 1, . . . r, and the categories of Y as j = 1,. . . ,k, and being the objective quantify the extent to which these variables are associated. Consider n_{ij} the frequencies which counts the number of times $X = i$ and $Y = j$.

Cramer's V is computed by taking the square root of the chi-squared statistic divided by the sample size and the minimum dimension minus 1, as follows,

$$V = \sqrt{\frac{\varphi^2}{min(k-1, r-1)}} = \sqrt{\frac{\chi^2/n}{min(k-1, r-1)}}, \tag{14}$$

where χ^2 is calculated as follows,

$$\chi^2 = \sum_{ij} \frac{(n_{ij} - \frac{n_i n_j}{n})^2}{\frac{n_i n_j}{n}}, \tag{15}$$

with n_{ij}, the number of times the values (X_i, Y_j) were observed.

Pearson's coefficient is useful to measure the correlation between two continuous variables and Cramer's V is a proper way to measure association between two categorical variables. Correlation Ratio is used in order to determine the correlation in a pair of a continuous variable and a categorical one, which has shown to be a good correlation measure [42]. Often marked with η^2, its value is in the range [0,1], where 0 means a category cannot be determined by a continuous measurement, and 1 means a category can be determined with absolute certainty. The Correlation Ratio works with a dependent variable, which would be the continuous one, and an independent variable, which would become the categorical one. Furthermore, an important difference with respect to the rest of the coefficients is that η^2 is useful not only to treat linear correlations but also for its application in non-linear correlation situations.

Suppose each observation is Y_{xi} where x indicates the category that observation is in and i is the label of the particular observation. Let n_x be the number of observations in category x, $\bar{y_x}$ is the mean of the category x and $\bar{y}$ is the mean of the whole population.

The correlation ratio η^2 is defined as to satisfy

$$\eta^2 = \frac{\sum_x n_x(\overline{y_x} - \overline{y})^2}{\sum_x n_x(y_{xi} - \overline{y})^2}. \tag{16}$$

After obtaining an association measure, the next step is to evaluate the quality of the results. The p-value [43] is a well-known method used to calculate the significance of the correlation measure.

In order to get an useful and easier view of the studied variables relationship, we use the Solar Correlation Map proposed in [44]. As shown in Figure 4, the solar correlation map is designed for the visual representation of the correlation of each input variable with the output variable. The output variable (Exitus) is represented by the Sun. Each circle around the sun is an orbit and each orbit represents a degree of correlation according to the results obtained by applying the stated methods (Pearson's coefficient for two continuous variables, Cramer's V for two categorical variables, and the Correlation Ratio between a categorical variable and another continuous one). Planets are input variables, so the more closed the orbit, the stronger the correlation. Planets on the first orbit have an absolute value of 0.9-1.0. The second orbit planets have a correlation coefficient of 0.8-0.9, and so forth.

This analysis by conventional statistical methods is performed with an algorithm developed with Python programming language, which chooses the most convenient method for each type of variable and represent the results using this Solar Maps.

AUTHOR CONTRIBUTIONS

AGG and IPE conducted the experiments and wrote the paper, together with JRFM, AGC and SMR. PJFB and SM contributed to the clinical analysis of experimental results and JLRA and AMF organized and reviewed the paper.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, and T. Kitai, "Artificial intelligence in precision cardiovascular medicine," *J. Amer. College Cardiology*, vol. 69, pp. 2655–2664, May 2017.

[2] D. R. Posircaru and L. Dan Serbanati, "Integrating legacy medical applications in a standardized electronic health record platform," in *Proc. E-Health Bioeng. Conf. (EHB)*, Nov. 2015, pp. 1–4.

[3] F. White, "Primary health care and public health: Foundations of universal health systems," *Med. Princ. Pract.*, vol. 24, no. 2, pp. 103–116, Feb. 2015.

[4] R. Goldberg, K. Currie, K. White, D. Brieger, P. Steg, S. G. Goodman, O. H. Dabbous, K. Fox, and J. Gore, "Six-month outcomes in a multinational registry of patients hospitalized with an acute coronary syndrome (the global registry of acute coronary events [GRACE])," *Amer. J. Cardiol.*, vol. 93, pp. 93–288, Feb. 2004.

[5] S. Kaptoge, L. Pennells, D. De Bacquer, M. T. Cooney, M. Kavousi, G. Stevens, L. M. Riley, G. Stevens, L. M. Riley, S. Savin, T. Khan, S. Altay, and P. Amouyel, "World health organization cardiovascular disease risk charts: Revised models to estimate risk in 21 global regions," *Lancet Global Health*, vol. 7, pp. 1332–1345, Oct. 2019.

[6] R. Virmani, A. P. Burke, A. Farb, and F. D. Kolodgie, "The pathology of vulnerable plaque," *J. Amer. College Cardiol.*, vol. 47, no. 8, pp. 8–13, 2006.

[7] J. VanHouten, J. M. Starmer, N. Lorenzi, D. J. Maron, and T. A. Lasko, "Machine learning for risk prediction of acute coronary syndrome," in *Proc. AMIA Symp.*, 2014, pp. 1940–1947.

[8] Z. Huang, W. Dong, H. Duan, and J. Liu, "A regularized deep learning approach for clinical risk prediction of acute coronary syndrome using electronic health records," *IEEE Trans. Biomed. Eng.*, vol. 65, pp. 956–968, 2018.

[9] Z. Huang and W. Dong, "Adversarial MACE prediction after acute coronary syndrome using electronic health records," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 5, pp. 2117–2126, Sep. 2019.

[10] C. L. McCullough, A. J. Novobilski, and F. M. Fesmire, "Use of neural networks to predict adverse outcomes from acute coronary syndrome for male and female patients," in *Proc. 6th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2007, pp. 512–517.

[11] L. S. Green and J. A. Abildskov, "Clinical applications of body surface potential mapping," *Clin. Cardiol.*, vol. 18, no. 5, pp. 245–249, May 1995.

[12] M. Viceconti, P. Hunter, and R. Hose, "Big data, big knowledge: Big data for personalized healthcare," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 4, pp. 1209–1215, Jul. 2015.

[13] S. Kumar and M. Singh, "Big data analytics for healthcare industry: Impact, applications, and tools," *Big Data Mining Analytics*, vol. 2, no. 1, pp. 48–57, Mar. 2019.

[14] L. Mertz, "What can big data tell us about health?: Finding gold through data mining," *IEEE Pulse*, vol. 7, no. 5, pp. 40–44, Sep. 2016.

[15] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "Health-CPS: Healthcare cyber-physical system assisted by cloud and big data," *IEEE Syst. J.*, vol. 11, no. 1, pp. 88–95, Mar. 2017.

[16] M. Roffi, C. Patrono, J. P. Collet, C. Mueller, M. Valgimigli, F. Andreotti, J. J. Bax, M. A. Borger, C. Brotons, D. P. Chew, B. Gencer, G. Hasenfuss, K. Kjeldsen, P. Lancellotti, U. Landmesser, J. Mehilli, D. Mukherjee, R. F. Storey, and S. Windecker, "2015 guidelines for the management of acute coronary syndromes in patients presenting without persistent st-segment elevation: Task force for the management of acute coronary syndromes in patients presenting without persistent st-segment elevation of the European society of cardiology (ESC)," *Eur. Heart J.*, vol. 37, pp. 267–315, Jun. 2016.

[17] B. Ibanez, S. James, S. Agewall, M. J. Antunes, C. Bucciarelli-Ducci, H. Bueno, A. Caforio, F. Crea, J. Goudevenos, S. Halvorsen, G. Hindricks, A. Kastrati, M. Lenzen, E. Prescott, M. Roffi, M. Valgimigli, C. Varenhorst, P. Vranckx, and P. Widimsky, "2017 ESC guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: The task force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European society of cardiology (ESC)," *Eur. Heart J.*, vol. 39, no. 2, pp. 119–177, 2018.

[18] M. Briel, I. Ferreira-Gonzalez, J. J. You, P. J. Karanicolas, E. A. Akl, P. Wu, P. Wu, B. Blechacz, D. Bassler, X. Wei, A. Sharman, and I. Whitt, "Association between change in high density lipoprotein cholesterol and cardiovascular disease morbidity and mortality: Systematic review and meta-regression analysis," *BMJ*, vol. 338, no. 1, p. b92, Feb. 2009.

[19] I. Ferreira-González, G. Permanyer-Miralda, J. Marrugat, M. Heras, J. Cuñat, E. Civeira, and F. Arós, "MASCARA (manejo del síndrome coronario agudo. Registro actualizado): Study, general findings," *Revista Española de Cardiología*, vol. 61, pp. 803–816, Jan. 2008.

[20] F. Bagherzadeh-Khiabani, A. Ramezankhani, F. Azizi, F. Hadaegh, E. W. Steyerberg, and D. Khalili, "A tutorial on variable selection for clinical prediction models: Feature selection methods in data mining could improve the results," *J. Clin. Epidemiol.*, vol. 71, pp. 76–85, Mar. 2016.

[21] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, Jan. 1979.

[22] C. M. Grinstead and J. L. Snell, *Introduction to Probability: Second Revised Edition*. American Mathematical Soc., 2012, pp. 1–4.

[23] S. Willich, J. Müller-Nordhorn, M. Kulig, S. Binting, H. Gohlke, H. Hahmann, K. Bestehorn, K. Krobot, and H. Völler, "Cardiac risk factors, medication, and recurrent clinical events after acute coronary disease. A prospective cohort study," *Eur. Heart J.*, vol. 22, no. 4, pp. 307–313, Feb. 2001.

[24] C. Picariello, C. Lazzeri, P. Attana, M. Chiostri, G. F. Gensini, and S. Valente, "The impact of hypertension on patients with acute coronary syndromes," *Int. J. Hypertension*, vol. 2011, pp. 101–108, Jun. 2011.

[25] G. Cotter, C. P. Cannon, C. H. McCabe, Y. Michowitz, E. Kaluski, A. Charlesworth, O. Milo, J. Bentley, A. Blatt, R. Krakover, R. Zimlichman, L. Reisin, A. Marmor, B. Lewis, Z. Vered, A. Caspi, and E. Braunwald, "Prior peripheral arterial disease and cerebrovascular disease are independent predictors of adverse outcome in patients with acute coronary syndromes: Are we doing enough? Results from the orbofiban in patients with unstable coronary syndromes-thrombolysis in myocardial infarction (OPUS-TIMI) 16 study," *Amer. Heart J.*, vol. 145, no. 4, pp. 622–627, Apr. 2003.

[26] P. Morillas, J. Quiles, A. Cordero, J. Guindo, F. Soria, P. Mazón, J. R. Gonzalez-Juanatey, and V. Bertomeu, "Impact of clinical and subclinical peripheral arterial disease in mid-term prognosis of patients with acute coronary syndrome," *Amer. J. Cardiol.*, vol. 104, no. 11, pp. 1494–1498, Dec. 2009.

[27] M. Lettino, P. Andell, U. Zeymer, P. Widimsky, N. Danchin, A. Bardaji, J. A. Barrabes, A. Cequier, M. J. Claeys, L. De Luca, and J. Dörler, "Diabetic patients with acute coronary syndromes in contemporary European registries: Characteristics and outcomes," *Eur. Heart J.-Cardiovascular Pharmacotherapy*, vol. 3, pp. 198–213, Oct. 2017.

[28] K. Franklin, R. J. Goldberg, F. Spencer, W. Klein, A. Budaj, D. Brieger, M. Marre, P. G. Steg, N. Gowda, and J. M. Gore, "Implications of diabetes in patients with acute coronary syndromes: The global registry of acute coronary events," *Arch. Internal Med.*, vol. 164, pp. 1457–1463, Jul. 2004.

[29] A. Cordero, R. López-Palop, P. Carrillo, J. Núñez, A. Frutos, V. Bertomeu-González, F. Yépez, N. Alcantara, F. Ribes, M. Juskova, and V. Bertomeu-Martínez, "Prevalence and postdischarge incidence of malignancies in patients with acute coronary syndrome," *Revista Española de Cardiología*, vol. 71, no. 4, pp. 267–273, Apr. 2018.

[30] P. Abrahamsson, J. Dobson, C. B. Granger, J. J. V. Mcmurray, E. L. Michelson, M. Pfeffer, S. Pocock, S. D. Solomon, S. Yusuf, and K. Swedberg, "Impact of hospitalization for acute coronary events on subsequent mortality in patients with chronic heart failure," *Eur. Heart J.*, vol. 30, no. 3, pp. 338–345, Sep. 2008.

[31] E. Ahmed, K. F. Al-Habib, A. El-Menyar, N. Asaad, K. Sulaiman, A. Hersi, W. Almahmeed, A. A. Alsheikh-Ali, H. Amin, A. Al-Motarreb, S. A. Saif, R. Singh, J. Al-Lawati, and J. A. Suwaidi, "Age and clinical outcomes in patients presenting with acute coronary syndromes," *J. Cardiovascular Disease Res.*, vol. 47, pp. 177–188, Jun. 2013.

[32] A. Avezum, M. Makdisse, F. Spencer, J. M. Gore, K. A. A. Fox, G. Montalescot, K. A. Eagle, K. White, R. H. Mehta, E. Knobel, and J. P. Collet, "Impact of age on management and outcome of acute coronary syndrome: Observations from the global registry of acute coronary events (GRACE)," *Amer. Heart J.*, vol. 149, no. 1, pp. 67–73, 2005.

[33] X. Dai, J. Busby-Whitehead, and K. P. Alexander, "Acute coronary syndrome in the older adults," *J. Geriatric Cardiol.*, vol. 13, pp. 101–108, Feb. 2013.

[34] J. J. Santopinto, K. A. A. Fox, R. J. Goldberg, A. Budaj, G. Pinero, A. Avezum, D. Gulba, J. Esteban, J. M. Gore, J. Johnson, and E. P. Gurfinkel, "Creatinine clearance and adverse hospital outcomes in patients with acute coronary syndromes: Findings from the global registry of acute coronary events (GRACE)," *Heart*, vol. 89, no. 9, pp. 1003–1008, Sep. 2003.

[35] R. Latchamsetty, J. Fang, E. Kline-Rogers, D. Mukherjee, R. F. Otten, T. M. LaBounty, M. S. Emery, K. A. Eagle, and J. B. Froehlich, "Prognostic value of transient and sustained increase in in-hospital creatinine on outcomes of patients admitted with acute coronary syndrome," *Amer. J. Cardiol.*, vol. 99, no. 7, pp. 939–942, Apr. 2007.

[36] B. M. Scirica and D. A. Morrow, "Troponins in acute coronary syndromes," *Prog. Cardiovascular Diseases*, vol. 47, no. 3, pp. 177–188, Nov. 2004.

[37] T. Killip and J. T. Kimball, "Treatment of myocardial infarction in a coronary care unit," *Amer. J. Cardiol.*, vol. 20, no. 4, pp. 457–464, Oct. 1967.

[38] O. Itzahki Ben Zadok, T. Ben-Gal, A. Abelow, A. Shechter, O. Zusman, Z. Iakobishvili, T. Cohen, N. Shlomo, R. Kornowski, and A. Eisen, "Temporal trends in the characteristics, management and outcomes of patients with acute coronary syndrome according to their killip class," *Amer. J. Cardiol.*, vol. 124, no. 12, pp. 1862–1868, Dec. 2019.

[39] K. A. A. Fox, O. H. Dabbous, R. J. Goldberg, K. S. Pieper, K. A. Eagle, F. Van de Werf, Á. Avezum, S. G. Goodman, M. D. Flather, F. A. Anderson, and C. B. Granger, "Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: Prospective multinational observational study (GRACE)," *BMJ*, vol. 333, no. 7578, p. 1091, Nov. 2006.

[40] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Pearson Correlation Coefficient*. Berlin, Germany: Springer, 2009, pp. 1–4.

[41] H. Cramer, *Mathematical Methods of Statistics*. Princeton Univ. Press, 1946, p. 286.

[42] A. H. Shaldehi, "Using Eta (η) correlation ratio in analyzing strongly nonlinear relationship between two variables in practical researches," *J. Math. Comput. Sci.*, vol. 07, no. 03, pp. 213–220, Oct. 2013.

[43] R. L. Wasserstein and N. A. Lazar, *The ASA Statement on P-Values: Context, Process, and Purpose*. American Statistician, 2016, pp. 129–133.

[44] S. Zapf and C. Kraushaar, *A New Visualization to Beautifully Explore Correlations*. Newton, MA, USA: O'Really, 2017.

ANTONIO GARCÍA-GARCÍA received the degree in computer engineering from the Pontifical University of Salamanca and the master's degree in direction and management of ICT for health from the Carlos III Health Institute. He has more than 18 years of experience working as the Head of Information Systems and Data Protection Delegate, Niño Jesús University Children's Hospital. He is currently an Associate Professor at the Department of Biomedical Engineering, Universidad Rey Juan Carlos. His research interests include the improvement of clinical decision support systems, the methodology for managing eHealth projects, and the standardization of health knowledge.

IGNACIO PRIETO-EGIDO received the degree in telecommunication engineering from Carlos III University, Madrid, and the M.Sc. degree in telecommunication networks for developing countries and Ph.D. degree from Rey Juan Carlos University (URJC). He has more than ten years of experience working on projects in Ecuador, Cambodia, Peru, Guatemala or Bolivia, with funding from international institutions, such as the European Union, the Development Bank of Latin America, or USAID. He is currently an Associate Professor at Rey Juan Carlos University and collaborates with the EHAS Foundation, focusing his work on designing, deploying, and evaluating innovative solutions to reduce maternal and infant mortality in rural areas of low-resource countries. He has published results in journals, such as *IEEE Communications Magazine*, *Journal of Telemedicine and Telecare or Telemedicine*, and *e-Health*. His research interest includes improving health care in rural areas of low-resource settings through eHealth.

ALICIA GUERRERO-CURIESES received the B.Sc. degree in telecommunication engineer from Universidad de Valladolid, in 1998, and the Ph.D. degree in telecommunication from University Carlos III de Madrid, in 2003. She is currently an Associate Professor with the Department of Signal Theory and Communications, Universidad Rey Juan Carlos, Spain. Her research interests include statistical learning theory, pattern recognition and their applications to biomedical signal processing, and remote sensing.

JUAN RAMÓN FEIJOO-MARTÍNEZ received the degree in telecommunication engineering from Universidad de Vigo, Spain, in 1991, and the Ph.D. degree from Universidad Carlos III de Madrid, Madrid, Spain, in 2007. He was with France Telecom in telecommunication transmission network planning. Since 2003, he has been with the Telecommunication Division, Red Eléctrica de España, mainly in network design and maintenance. He has also been an Adjunct Professor with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid. He is currently with University Rey Juan Carlos, Spain. He has participated in several research and development projects involving reliability and resiliency of large critical infrastructures. His main research interests include network reliability improvement, estimation and prediction methods based on statistical learning theory, support vector machines, Bayesian methods, cybersecurity, and big data analytics.

SERGIO MUÑOZ-ROMERO received the degree in engineering and Ph.D. degree in machine learning from Universidad Carlos III, Spain, in 2009 and 2015, respectively. His current research interests include interpretable machine learning algorithms and statistical learning theory, mainly dimensionality reduction and feature selection methods for real-world biomedical problems, especially for aging, oncology and cardiology.

SERGIO MANZANO FERNÁNDEZ was a Resident Intern at the Cardiology Service of the Virgen de la Arrixaca University Clinical Hospital, Murcia, from 2005 to 2010, where he became an Optional Specialist. He has been a Professor with the Department of Internal Medicine, University of Murcia, since 2013. He is the author of more than 80 publications indexed in journals of recognized international impact and has actively participated in more than 200 oral communications and posters in both national and international conferences of the specialty. His research interests include prolonged electrocardiographic monitoring in atrial fibrillation and heart failure or arrhythmic risk stratification.

PEDRO JOSÉ FLORES-BLANCO received the degree in medicine from the University of Murcia, in 2010. He was a Cardiology Specialist. He completed his residence at the Virgen de la Arrixaca University Clinical Hospital in Murcia, from 2011 to 2016. He was a Doctor from the University of Murcia in 2016. He is a collaborating researcher in various research projects. He has published more than 30 scientific articles in indexed national and international journals and more than 100 communications to national and international congresses.

JOSÉ LUIS ROJO-ÁLVAREZ (Senior Member, IEEE) received the degree in telecommunication engineering from the University of Vigo, Spain, in 1996, and the Ph.D. degree in telecommunication from the Polytechnic University of Madrid, Spain, in 2000. Since 2016, he has been a Full Professor with the Department of Signal Theory and Communications, University Rey Juan Carlos, Madrid. He has published more than 110 articles in JCR journals and more than 160 international conference communications. He has participated in more than 55 research projects (with public and private funding), and directed more than ten of them, including several actions in the National Plan for Research and Fundamental Science. He co-started a Pioneer Degree Program on biomedical engineering, involving hospitals and companies in the electromedic and eHealth field. His research interests include statistical learning theory, digital signal processing, and complex system modeling, with applications to digital communications and to cardiac signals and image processing.

ANDRÉS MARTÍNEZ-FERNÁNDEZ received the degree in telecommunications engineering and the Ph.D. degree from the Technical University of Madrid (UPM), in 1994 and 2003, respectively. He is currently an Associate Professor with the Department of Signal Theory and Communications and Telematic Systems and Computing, Rey Juan Carlos University (URJC). He has directed many research projects financed by national and international entities, such as the European Commission, the World Bank, the Inter-American Development Bank, or USAID. He has published numerous articles in high-impact indexed journals and edited and written several books on the subject. His research interests include application of information and communication technologies to improve the quality of life of isolated and dispersed populations in developing countries, working mainly in rural communications and telemedicine.