IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 14, 2020, accepted January 31, 2021, date of publication February 8, 2021, date of current version February 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3057770

A Resampling Univariate Analysis Approach to Ovarian Cancer From Clinical and Genetic Data

LUIS BOTE-CURIEL[1], SERGIO RUIZ-LLORENTE[2], SERGIO MUÑOZ-ROMERO[1,3], MÓNICA YAGÜE-FERNÁNDEZ[2], ARANTZAZU BARQUÍN[2], JESÚS GARCÍA-DONAS[2], AND JOSÉ LUIS ROJO-ÁLVAREZ[4], (Senior Member, IEEE)

[1]Department of Signal Theory and Communications, Universidad Rey Juan Carlos, 28942 Fuenlabrada, Spain
[2]Unit of Gynecological, Genitourinary and Skin Tumors, Hospital HM Sanchinarro, Fundación Investigación HM Hospitales, 28050 Madrid, Spain
[3]Center for Computational Simulation, Universidad Politécnica de Madrid, 28223 Pozuelo de Alarcón, Spain
[4]Department of Signal Theory and Communications and Telematic Systems and Computation, Universidad Rey Juan Carlos, 28942 Fuenlabrada, Spain

Corresponding author: José Luis Rojo-Álvarez (joseluis.rojo@urjc.es)

This work was supported in part by the Science and Innovation Ministry Grants meHeart-RisBi and Beyond under Grant PID2019-104356RB-C42 and Grant PID2019-106623RB-C41, in part by FINALE Project under Grant TEC2016-75161-C2-1-R and Grant TEC2016-75161-C2-2-R, in part by KERMES Project under Grant TEC2016-81900-REDT, and in part by FEDER Fundings under project PID2019-106623RB-C41.

ABSTRACT Ovarian cancer (OC) is the second most common gynecological malignancy and the gynecological tumor with the worst prognosis. To try to improve this situation, Data Science technologies could be a useful tool to help clinicians to know more about the disease. In our case, we are interested in exploring OC data to discover relationships between clinical and genetic factors and the disease progression. For it, we propose an analysis framework for simple and univariate statistical descriptions of features of different types, based on bootstrap resampling. Foremost, we define the framework for metric, categorical, and dates variables and determine what are the advantages and disadvantages of using different bootstrap resampling strategies, based on their statistical basis. Then, we use it to perform a univariate analysis over an OC dataset that allows to explore how is the disease progression, having platinum-free interval as indicator, in relation to clinical and genetic features of different types. Also, it provides a first set of variables possibly relevant for survival prediction. Results obtained show that some features have led to individual differences between both platinum resistant (<6 months) and platinum sensitive(>6 months) groups. It can be concluded that this could be an indicator that the database could be discriminatory for the hypotheses studied, though it is convenient to make multivariate analyses to check how relationships among features are influenced.

INDEX TERMS Bootstrap resampling, data science analytics, genetic data, hypothesis test, ovarian cancer, univariate analysis.

I. INTRODUCTION

Ovarian cancer (OC) is the second most common gynecological malignancy, with an estimated annual incidence of 225 000 women worldwide and 5-year survival rate of approximately 45% [1], being the gynecological tumor with the worst prognosis (140 000 exitus per year) [2]. Regarding histology, epithelial tumors account for >90% of all the ovarian carcinomas. These tumors include high grade serous OC (HGSOC), which is the most common subtype, as well as endometrioid, clear cell, and mucinous histologies. Standard treatment includes cytoreductive debulking surgery and platinum based chemotherapy [2]. Despite optimal therapy, outcomes remain quite poor as the vast majority of the patients are diagnosed in advanced stages of disease. Moreover, despite approximately 85% diagnosed OC initially respond to chemotherapies, appearance of relapses is common and responses to subsequent therapies are generally short-lived [3].

Data Science technologies refer to a group of tools used to extract relevant information from datasets in a wide range of applications, while Big Data refers to the application of Data Science to large volumes of data [4]. These technologies have been strongly developed in the last years, and almost all sectors over the world are currently turning to use them extensively. Several applications and solutions in academia and industry fall into areas such as finance, remote sensing, transportation, education, marketing and advertising,

The associate editor coordinating the review of this manuscript and approving it for publication was Eunil Park.

or tourism [5]. In addition, Data Science along with Big Data are leading healthcare towards a new era because in the past decades there has been a massive growth in biomedical data, such as genomic sequences, electronic health records (EHR), and biomedical signals and images [6]. These applications include areas like individual disease diagnosis, disease prognosis, disease prevention and prediction, or design of tailored health treatments based on lifestyle [7].

Overall, Data Science and Big Data can be viewed from two different approaches. On the one hand, predictive analyses and pattern recognition can be performed using machine learning and deep learning algorithms. On the other hand, analysis of databases from descriptive statistics can be extremely useful, given the relevant information they can contain [8]. These descriptive statistics are the simplest analysis in Data Science, and they involve the summarisation and description through the use of basic statistical methods. Normally, when we want to use Data Science and Big Data technologies in a new application field, it is necessary to perform an analysis of these basic statistics before going into more advanced or sophisticated analysis such as machine learning data models [9].

In healthcare and other many environments, databases contain a number of features (variables) of different types, namely, metric, categorical, date, and text. However, many data processing and Data Science analyses are designed to work with the same type of features in their input space. For instance, linear discriminant analysis, regression analysis, or analysis of variance, as well as principal component analysis, are purposed to use metric variables [10]. On the other hand, multiple correspondence analysis provides us with instruments to scrutinize categorical variables [11], whereas Bag of Words [12] or Latent Dirichlet Allocation [13] yield different ways to extract knowledge from text analysis. It turns out in practical terms that, while necessary, dealing with variables with different types in datasets can have some undesired effects, including both the difficulty to use the analysis methods with mixed variable types, and the difficulty to provide unified exploration overviews of datasets, specially when the number of features grows.

The focal point of this article is to perform an exploration of OC data from a univariate descriptive statistics approach, using a unified analysis framework for different type of features, to try to discover relationships between clinical and genetic factors and the disease progression. To accomplish this purpose, we need to complete and expand the work initiated in previous papers [8], [9], [14], where the use of bootstrap resampling has been proposed as a framework to unify the analysis of variables of different types. However, these previous works have been limited primarily to a single type of bootstrap resampling. In this article, we first determine what are the advantages and disadvantages of the use of different bootstrap resampling strategies, based on their statistical basis. Secondly, using the best resulting bootstrap resampling strategy, we perform an initial univariate analysis of an OC dataset to find relationships between clinical and genetic features and the disease progression. At the same time, this analysis provides a first set of features possibly relevant for survival prediction.

The scheme of the article is as follows. In Section II, we describe the OC database used in this work, which is composed of two parts, namely, clinical data and genetic data. Then, in Section III, we present an analysis framework employed to unify univariate statistical descriptions of different types of features using bootstrap resampling, and we expose and compare different bootstrap resampling methods. After that, experiments with synthetic data and results with OC data are provided in Section IV. Finally, discussion and several conclusions are established in Section V.

II. DATABASE DESCRIPTION

The database used in this work was created as part of the *BRCAness* initiative from the Innovation Oncology Laboratory of the Gynecologycal, Genitourinary and Skin Cancer Department, at Clara Campal Comprehensive Cancer Center (Madrid, Spain). The department has conducted a multicenter observational study [1] since 2013 focused on the identification of biomarkers with a potential impact in clinical practice. Approximately 300 OC patients have already been included so far in this study, which is supported by 12 national Health Care Institutions. Inclusion criteria referred both to age (>18 years old) and disease status (Stage IC or superior). Of all these patients, 54 cases were molecularly characterized by means of Next Generation Sequencing (NGS), either whole-exome sequencing (WES) or predesigned targeted gene panels (Onco80).

Extensive clinical information of each patient has been collected including age at diagnosis, personal or familial antecedents, *BRCA* and *TP53* status, histological subtype, grade and stage of the disease at diagnosis, anatomical location, presence of perineural or vascular invasion, CA-125 biomarker evaluation, surgical procedures, information related to the different treatment lines prescribed for each patient (number of cycles, doses, associated toxicities, grade of response and relapses), and date of the last clinical follow up or exitus. Other important clinical features such as overall survival (OS), progression-free survival (PFS), or platinum-free interval (PFI) were also included. In total, the clinical database consists of 54 entries (one for each patient) and 106 clinical features.

WES profiling (SureSelect Human All Exon V6) was performed in 20 patients categorized according to extreme degree of response to platinum based therapies either in long-term (PFS>36 months, 11 patients) or short-term (PFS<8 months, 9 patients) responders. Sequencing was performed using genomic DNA extracted from either formalin-fixed paraffin embedded tumoral tissue or peripheral blood to ultimately define somatic and germinal variants. Subsequently, 34 patients showing intermediate degree of response

[1] An informed consent was obtained from all the study participants, and the study was approved by the Institutional Review Board of HM Hospitals Ethics committee.

to platinum agents were screened using a Onco80 predesigned mutational panel which allows to identify genetic alterations in 80 *locus* widely associated to cancer development. Subsequent filtering of sequencing data revealed 7 077 genetic alterations for the studied patients (WES, 3 somatic mutations per patient on average; Onco80 panel, 201 alterations per patient on average). Most relevant genetic variables included genetically altered *loci*, amino acids substitution, pathogenicity scores (Grantham's distance, conservation scores according to several *in silico* programs), or the resultant genetic changes. A preliminary analysis of this database has been previously presented in [15].

III. DATA ANALYSIS

In this work, we use an analysis framework for simple and univariate statistical descriptions of features in databases. Different data types in a database demand analysis of categorical, metric, date, and text features. Therefore, we state a similar analysis framework for all of them, making results more interpretable by the users (managers, clinicians, and others). Specifically, this framework is established in terms of hypothesis tests for differences in proportions, means, standard deviations, and probability density distributions between two interest groups, using bootstrap resampling. There are precedent applications of this analysis framework in customer relationship management in the Spanish hospitality industry [9], telecontrol event severity in maintenance forms from Spanish high-power distribution grids [8], and organization management in Spanish Red Cross [14].

From a database point of view, we work with simple two-dimensional tables (electronic sheets). These tables have a set of N observations, expressed as $\{E^n, n = 1, \ldots, N\}$, with a data structure given by a concatenation of J features, denoted as $\{F_j, j = 1, \ldots, J\}$. Hence, the value of the j^{th} feature of the n^{th} measure is written as E_j^n.

Each feature can belong to one type in a set of different possible data types. In our case, we study in this work categorical, metric, and date types. So, for each j^{th} feature, we define a type denoted by

$$F_j.type \in \{\mathfrak{C}, \mathfrak{M}, \mathfrak{D}\},$$

where special letters denote the three mentioned data types considered in this work, respectively.

A. METRIC VARIABLES

By M_j we denote a feature F_j such that $F_j.type = \mathfrak{M}$. Its probability density function (*pdf*) is denoted as $f_{M_j}(M_j)$. If we establish two groups of observations, namely, G_1 and G_2, this *pdf* distribution can be seen as the marginal distribution

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

where $P(G_1)$ and $P(G_2)$ are the a priori probabilities in each group, and $f_{M_j}(M_j|G_1)$ and $f_{M_j}(M_j|G_2)$ are the conditional *pdf* to each group for this feature.

For each group, it is also important to consider the mean and standard deviation, denoted by

$$m_j^{G_1}, \quad m_j^{G_2},$$
$$\sigma_j^{G_1}, \quad \sigma_j^{G_2}.$$

To detect significant differences between both groups, we can define several statistics as the differences in means, standard deviations, and conditional *pdf*s, this is,

$$\Delta m_j = m_j^{G_1} - m_j^{G_2},$$
$$\Delta \sigma_j = \sigma_j^{G_1} - \sigma_j^{G_2},$$
$$\Delta f_{M_j} = f_{M_j}(M_j|G_1) - f_{M_j}(M_j|G_2).$$

These statistics could be used to make hypothesis tests to detect significant differences between the two groups. For example, with the difference in means we could use Student's t-test, one of the classical hypothesis tests which is used for comparing the means of two independent or paired samples [16]. In this test, under the null hypothesis of zero mean difference (same means in both groups), the t-statistic is calculated applying the proper expression of the standard error, which is distinct when the two groups have the same variance or different variance. With this t-statistic and the degrees of freedom, the p-value is obtained from the corresponding t-distribution. After this, the p-value is compared with the selected significant level and the null hypothesis is rejected or not.

B. CATEGORICAL VARIABLES AND DATES

We denote by C_j a feature F_j with $F_j.type = \mathfrak{C}$. This feature can have several possible categories among a discrete set identified as $C_j.value = \{v_j^k, k = 1, \ldots, K_j\}$, where K_j is the number of possible categories of C_j. The probability mass function (*pmf*) of that categorical variable is given by $P(v_j^k)$, which can be seen as the proportion of presence of categories in a finite observed set. If we consider two groups, G_1 and G_2, conditional *pmf*s for that variable are as follows,

$$P(v_j^k|G_1), \quad P(v_j^k|G_2),$$

and we can define a statistic as the difference in conditional *pmf*s according to

$$\Delta P(v_j^k) = P(v_j^k|G_1) - P(v_j^k|G_2).$$

To analyse variables with date types, this same approach can be applied. To be precise, the day of the week, the day of the month, the number of month, and the year can be expressed as categorical variables.

The difference in conditional *pmf*s statistic could be used to perform hypothesis tests. In this case, as difference of conditional *pmf*s is basically a difference in proportions of each category, a z-test, for example, could be used. This test is very used in classical statistics for difference in proportions to compare two samples [16]. To this effect, the z-statistic is calculated under the null hypothesis of zero proportion difference (same proportions in the two groups). Using this

z-statistic, the p-value is extracted from the z-distribution. Finally, the null hypothesis is rejected or not depending on whether the p-value is higher or lower of the selected significant level.

C. HYPOTHESIS TEST

Difference-based statistics previously defined could be used with many hypothesis tests. However, we are interested in a hypothesis test with a similar form for any type of feature, stating a equivalent framework that makes results more interpretable. This is the case of the following hypothesis test that we define. Although it is indicated for difference in proportions, it is valid for any of the difference-based statistics without loss of generality:

- Null hypothesis, $H_0 : \Delta P(v_j^k) = 0$, there is no difference between groups for this category.
- Alternative hypothesis, $H_A : \Delta P(v_j^k) \neq 0$, there is difference between groups for this category. If $\Delta P(v_j^k) > 0$ ($\Delta P(v_j^k) < 0$), then the proportion of this category is larger in G_1 (G_2).

However, when there is difference between both groups, we need to establish whether this difference is large enough to support statistical significance. To deal with this, we calculate an estimation of the *pdf* of the difference in proportions, $\Delta P(v_k^j)$, employing bootstrap resampling methods. We denote this *pdf* as $f_{\Delta P}(\Delta P)$. If the confidence interval (CI) over the estimation of $f_{\Delta P}(\Delta P)$ overlaps 0, we do not reject the null hypothesis, H_0. Nevertheless, if the CI over the estimation of $f_{\Delta P}(\Delta P)$ does not overlap 0, we reject the null hypothesis, H_0, and accept the alternative hypothesis, H_A. With this, if the CI over the estimation of $f_{\Delta P}(\Delta P)$ is located at positive (negative) values, this category is a relevant property in G_1 (G_2).

Whenever we reject or not a hypothesis, we could be wrong. We call type I error or false positive as the rejection of a true null hypothesis, i.e., when the null hypothesis is true, but is rejected by our decision. Type II error or false negative is defined, however, as the non-rejection of a false null hypothesis, i.e., when the null hypothesis is false, but it is not rejected by our decision.

D. BOOTSTRAP ESTIMATORS

Bootstrap resampling is a very standard statistical technique [17]. The idea is that if we want to make some inference of a population in terms of some statistic whose calculation is known, but its actual distribution is not easy to obtain analytically, we can resample with replacement the sample data and make inferences on the resamples.

In our case, given a sample dataset, we resample it with replacement B times. We obtain B replications of the difference in proportions, $\Delta P^*(v_k^j)$, that are used to build a histogram, which represents an estimation of $f_{\Delta P}(\Delta P)$ when is conveniently scaled. From now on, this estimation of $f_{\Delta P}(\Delta P)$ is written as $f^*_{\Delta P}(\Delta P)$. We use $f^*_{\Delta P}(\Delta P)$ to determine the hypothesis test. All the steps of this method, which we call *Method 1*, are synthesized in Algorithm 1. The asterisk symbol * is used to identify the quantities that have been estimated throughout the bootstrap process.

1 Steps to Calculate Method 1

Input: Dataset divided in two groups, G_1 and G_2, and amount of resamples, B.
Output: Rejection or not of the null hypothesis, H_0.

for $b \leftarrow 1$ **to** B **do**
  Resample G_1 and G_2, separately.
  Obtain the proportions per category conditional to G_1 and G_2, $P^*(v_j^k|G_1)$ and $P^*(v_j^k|G_2)$.
  Obtain the difference of proportions between G_1 and G_2, $\Delta P^*(v_j^k) = P^*(v_j^k|G_1) - P^*(v_j^k|G_2)$.
end for
Obtain an estimation of $f_{\Delta P}(\Delta P)$, $f^*_{\Delta P}(\Delta P)$, using the B replications of $\Delta P^*(v_j^k)$ calculated previously.
Obtain the CI over $f^*_{\Delta P}(\Delta P)$.
if CI over $f^*_{\Delta P}(\Delta P)$ overlaps 0 **then**
  Not reject the null hypothesis, H_0, since the proportion is the same between G_1 and G_2.
else if CI over $f^*_{\Delta P}(\Delta P)$ doesn't overlap 0 **then**
  Reject the null hypothesis, H_0, and accept the alternative hypothesis, H_A, since the proportion is different between G_1 and G_2.
  if CI over $f^*_{\Delta P}(\Delta P) > 0$ **then**
    $P^*(v_j^k|G_1) > P^*(v_j^k|G_2)$, which means that this category is a relevant property in G_1.
  else if CI over $f^*_{\Delta P}(\Delta P) < 0$ **then**
    $P^*(v_j^k|G_1) < P^*(v_j^k|G_2)$, which means that this category is a relevant property in G_2.
  end if
end if

Apart from this Method 1, we find other two relevant methods in the literature, which we denominate here *Method 2* and *Method 3*. Method 2 [17] is quite similar to Method 1, with the difference that a z-score normalization is carried out for each of the $\Delta P^*(v_j^k)$ of the B resamples. Hence, if $m_{\Delta P^*}$ is the mean and $\sigma_{\Delta P^*}$ is the standard deviation of the set $\{\Delta P^{*1}(v_j^k), \ldots, \Delta P^{*B}(v_j^k)\}$, we perform the following normalization:

$$\widetilde{\Delta P}^*(v_j^k) = \frac{\Delta P^*(v_j^k) - m_{\Delta P^*}}{\sigma_{\Delta P^*}},$$
$$\widetilde{0} = \frac{0 - m_{\Delta P^*}}{\sigma_{\Delta P^*}}.$$

The steps to calculate this Method 2 are detailed in Algorithm 2. Method 3 [18] is also similar to Method 1 proposed previously, but it has a significant variation. In this case, instead of resampling group G_1 and group G_2 independently, we resample both groups G_1 and G_2 jointly, and subsequently, we split randomly into new G_1 and G_2 groups. In addition, it is taken into account the initial difference of proportions between groups $\Delta P(v_j^k) = P(v_j^k|G_1) - P(v_j^k|G_2)$. The steps of this Method 3 are presented in the Algorithm 3.

2 Steps to Calculate Method 2

Input: Dataset divided in two groups, G_1 and G_2, and amount of resamples, B.

Output: Rejection or not of the null hypothesis, H_0

for $b \leftarrow 1$ **to** B **do**
  Resample G_1 and G_2, separately.
  Obtain the proportions per category conditional to G_1 and G_2, $P^*(v_j^k|G_1)$ and $P^*(v_j^k|G_2)$.
  Obtain the difference of proportions between G_1 and G_2, $\Delta P^*(v_j^k) = P^*(v_j^k|G_1) - P^*(v_j^k|G_2)$.
end for
Obtain $m_{\Delta P^*}$ and $\sigma_{\Delta P^*}$ of the set of the B replications of $\Delta P^*(v_j^k)$ calculated previously.
Obtain $\widetilde{\Delta P^*}(v_j^k)$ and $\widetilde{0}$ using $m_{\Delta P^*}$ and $\sigma_{\Delta P^*}$.
Obtain a normalised estimation of $f_{\Delta P}(\Delta P)$, $\widetilde{f}^*_{\Delta P}(\Delta P)$, using the B replications of $\widetilde{\Delta P^*}(v_j^k)$.
Obtain the CI over $\widetilde{f}^*_{\Delta P}(\Delta P)$.
if CI over $\widetilde{f}^*_{\Delta P}(\Delta P)$ overlaps $\widetilde{0}$ **then**
  Not reject the null hypothesis, H_0, since the proportion is the same between G_1 and G_2.
else if CI over $\widetilde{f}^*_{\Delta P}(\Delta P)$ doesn't overlap $\widetilde{0}$ **then**
  Reject the null hypothesis, H_0, and accept the alternative hypothesis, H_A, since the proportion is different between G_1 and G_2.
  if CI over $\widetilde{f}^*_{\Delta P}(\Delta P) > \widetilde{0}$ **then**
    $P^*(v_j^k|G_1) > P^*(v_j^k|G_2)$, which means that this category is a relevant property in G_1.
  else if CI over $\widetilde{f}^*_{\Delta P}(\Delta P) < \widetilde{0}$ **then**
    $P^*(v_j^k|G_1) < P^*(v_j^k|G_2)$, which means that this category is a relevant property in G_2.
  end if
end if

3 Steps to Calculate Method 3

Input: Dataset divided in two groups, G_1 and G_2, and amount of resamples, B.

Output: Rejection or not of the null hypothesis, H_0.

Obtain the difference of proportions between G_1 and G_2, $\Delta P(v_k^j) = P(v_j^k|G_1) - P(v_j^k|G_2)$.
for $b \leftarrow 1$ **to** B **do**
  Resample G_1 and G_2, jointly.
  Split randomly into 2 new G_1 and G_2 groups.
  Obtain the proportions per category conditional to G_1 and G_2, $P^*(v_j^k|G_1)$ and $P^*(v_j^k|G_2)$.
  Obtain the difference of proportions between G_1 and G_2, $\Delta P^*(v_j^k) = P^*(v_j^k|G_1) - P^*(v_j^k|G_2)$.
end for
Obtain an estimation of $f_{\Delta P}(\Delta P)$, $f^*_{\Delta P}(\Delta P)$, using the B replications of $\Delta P^*(v_j^k)$ calculated previously.
Obtain the CI over $f^*_{\Delta P}(\Delta P)$.
if CI over $f^*_{\Delta P}(\Delta P)$ overlaps $\Delta P(v_j^k)$ **then**
  Not reject the null hypothesis, H_0, since the proportion is the same between G_1 and G_2.
else if CI over $f^*_{\Delta P}(\Delta P)$ doesn't overlap $\Delta P(v_j^k)$ **then**
  Reject the null hypothesis, H_0, and accept the alternative hypothesis, H_A, since the proportion is different between G_1 and G_2.
  if CI over $f^*_{\Delta P}(\Delta P) < \Delta P(v_j^k)$ **then**
    $P^*(v_j^k|G_1) > P^*(v_j^k|G_2)$, which means that this category is a relevant property in G_1.
  else if CI over $f^*_{\Delta P}(\Delta P) > \Delta P(v_j^k)$ **then**
    $P^*(v_j^k|G_1) < P^*(v_j^k|G_2)$, which means that this category is a relevant property in G_2.
  end if
end if

In order to illustrate the functioning of these three methods and their differences in interpretation, we present a simple example. We generate a sample dataset composed of two groups, G_1 and G_2, using a Bernoulli distribution with only one feature, F_0, with $F_0.type = \mathfrak{C}$, which we denoted C_0. This feature can have two possible categories, $C_0.value \equiv \{v_0^0 = 0, v_0^1 = 1\}$. In this example, we focus only on category $v_0^1 = 1$. G_1 is generated with a Bernoulli parameter $p = 0.5$, while G_2 is generated with a set of parameters $p = \{0.0, 0.1, 0, 2, \ldots, 0.9, 1.0\}$. In Fig. 1 (a), we show Method 1, where $f^*_{\Delta P}(\Delta P)$ is calculated. The red line indicates the 0 and the grey area represents the CI. The only situation in which null hypothesis, H_0, is not rejected is when G_2 is generated with $p = 0.5$. In this case, the CI over $f^*_{\Delta P}(\Delta P)$ overlaps 0. This is because, if G_1 is generated with $p = 0.5$, then $P(v_1^0|G_1) \approx 0.5$, and if G_2 is also generated with $p = 0.5$, then $P(v_0^1|G_2) \approx 0.5$. Therefore, we would have that $\Delta P^*(v_0^1) \approx 0$.

An example of Method 2 and Method 3 is presented in the Fig. 1 (b) and Fig. 1 (c), respectively. In Method 2, we calculate $\widetilde{f^*}_{\Delta P}(\Delta P)$, where red points indicate the $\widetilde{0}$. In Method 3, we compute $f^*_{\Delta P}(\Delta P)$, and red points indicate the initial difference of proportions between groups $\Delta P(v_0^1)$. In the same way as in Method 1, in Method 2 and Method 3 only when G_2 is generated with $p = 0.5$, null hypothesis H_0 is not rejected.

Bootstrap resampling methods provide robust and easy-to-obtain estimates. However, the rationalization of their computational intensity and the revision of their theoretical foundations in large data scenarios deserve more attention. Some authors have raised in emerging concepts like the bag of little bootstraps [19], which are receiving special attention for Big Data analysis scenarios.

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENTS WITH SYNTHETIC DATA

In the previous section, a hypothesis test method has been presented (Method 1) along with two different variants (Method 2 and Method 3). However, we would like to determine the performance of each method in relation to the number of false positives and false negatives, and how they are distributed for each of the methods.

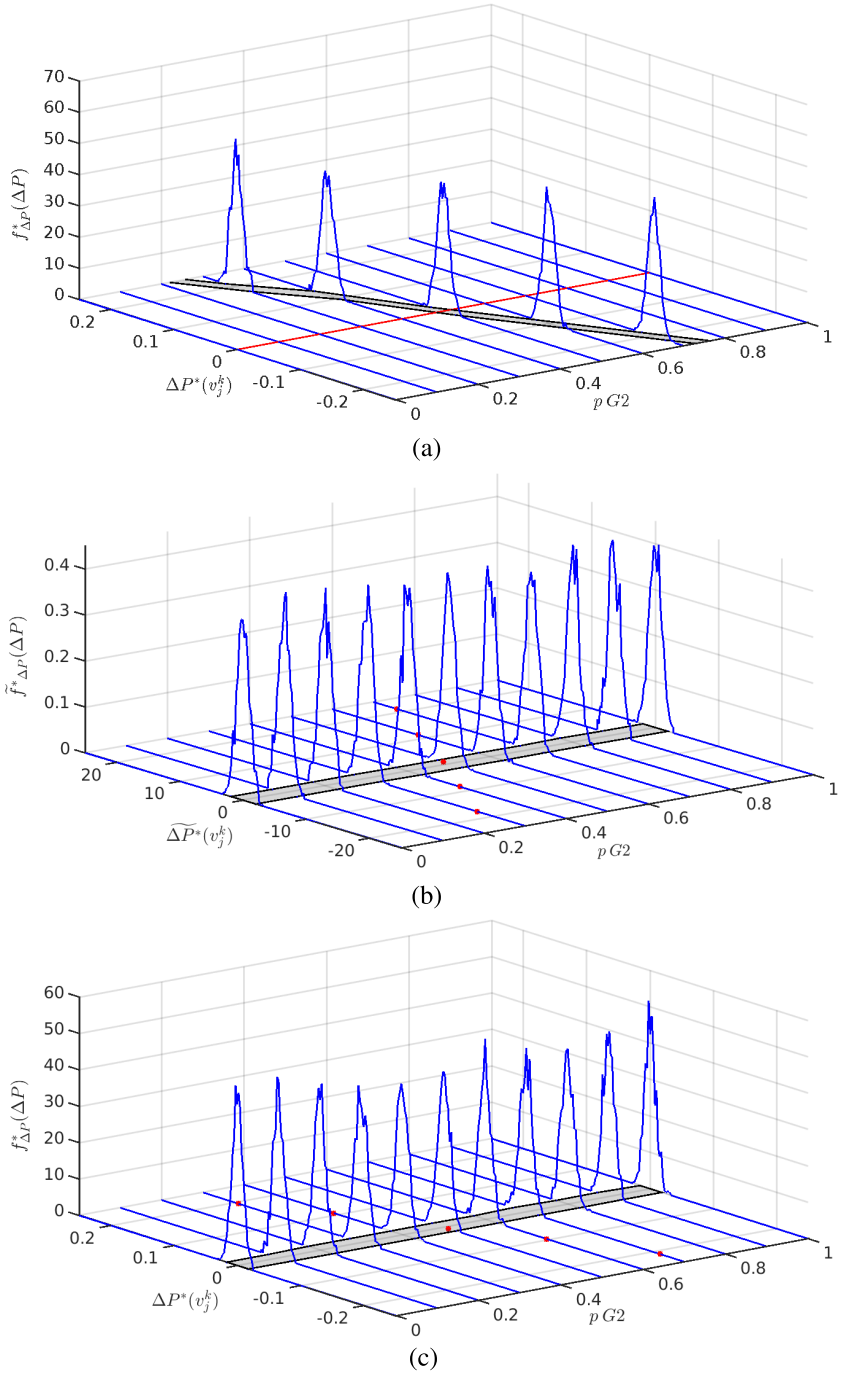

FIGURE 1. Simple example performed with Methods 1, Method 2, and Method 3. (a) In Method 1, $f^*_{\Delta P}(\Delta P)$ is calculated for several proportion parameters, p, of group G_2, where grey area represents the CI and red line marks the 0. (b) In similar way, in Method 2, $\widetilde{f}^*_{\Delta P}(\Delta P)$ is calculated, where red points mark the $\widetilde{0}$. (c) Similarly, in Method 3, $f^*_{\Delta P}(\Delta P)$ is also calculated, where red points mark the initial $\Delta P(v_j^k)$.

A false positive occurs, in Method 1, when the CI over $f^*_{\Delta P}(\Delta P)$ does not overlap 0 while having to overlap it; In Method 2, when the CI over $\widetilde{f}^*_{\Delta P}(\Delta P)$ does not overlap $\widetilde{0}$ while having to overlap it; And, in Method 3, when the CI over $f^*_{\Delta P}(\Delta P)$ does not overlap $\Delta P(v_j^k)$ while having to overlap it. Similarly, a false negative can be observed, in Method 1, when the CI over $f^*_{\Delta P}(\Delta P)$ overlaps the 0 while not having to overlap it; In Method 2, when the CI over $\widetilde{f}^*_{\Delta P}(\Delta P)$ overlaps $\widetilde{0}$ while not having to overlap it; And, in Method 3, when the CI over $f^*_{\Delta P}(\Delta P)$ overlaps $\Delta P(v_j^k)$ while not having to overlap it.

For scrutinizing the behavior of the three methods in terms of false positives, we generate several one-feature sample

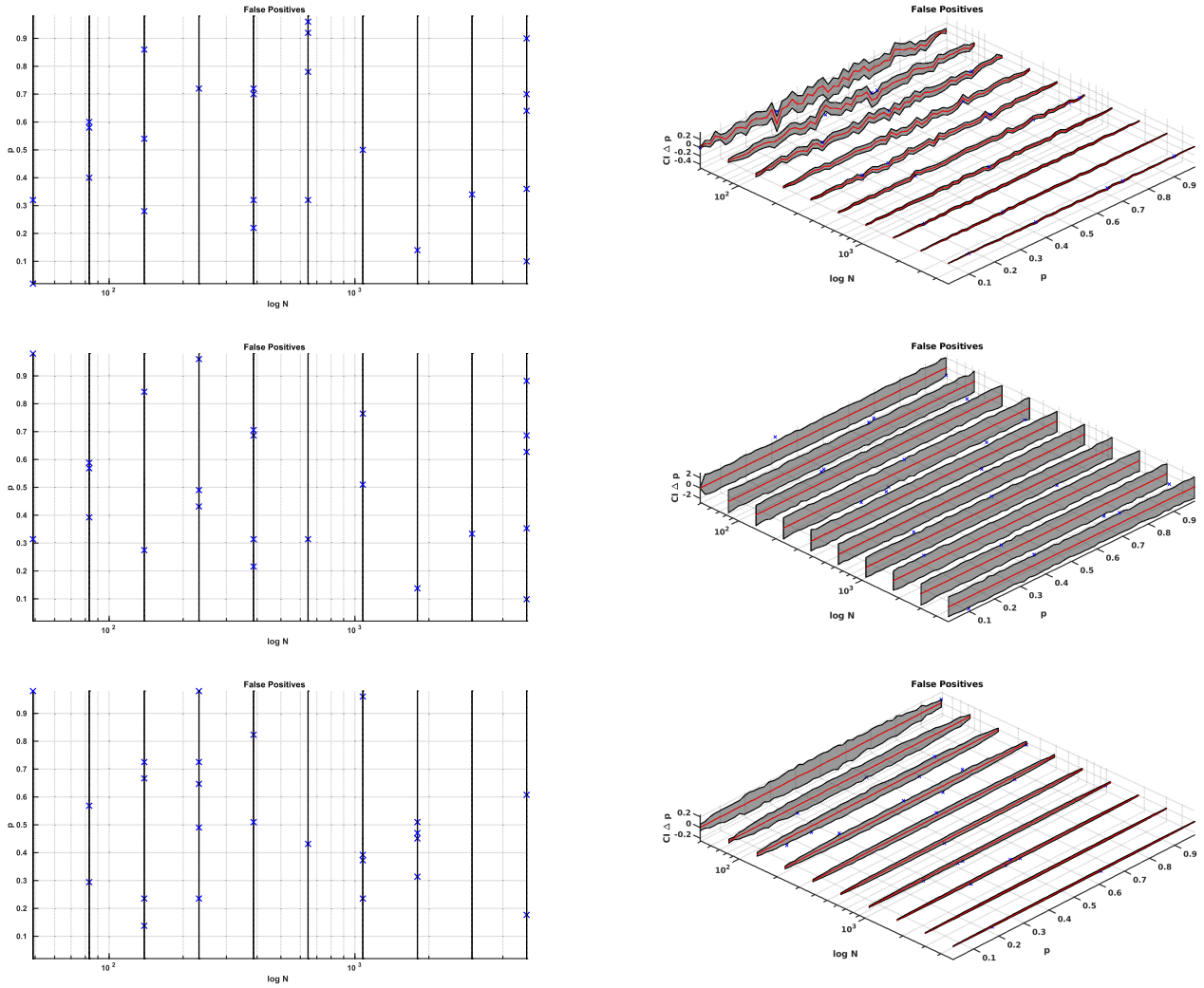

FIGURE 2. **Experiments with synthetic data that show the performance of Method 1 (first row), Method 2 (second row), and Method 3 (third row) according to false positives. In the left column, false positives represented with blue crosses for each of the three methods, where *N* is the sample size and *p* is the Bernoulli distribution proportion parameter for category $v_0^1 = 1$ both in G_1 and in G_2. In the right column, a 3D view of previous plots, with CI on the vertical axis depicted in grey and mean values depicted with a red line.**

dataset composed of two groups, G_1 and G_2, using Bernoulli distributions with two categories, $v_0^0 = 0$ and $v_0^1 = 1$. In particular, we increase, for several sample size, N, the Bernoulli distribution proportion parameter, p, for the category $v_0^1 = 1$ in both groups, G_1 and G_2, so that the difference in proportions between groups, Δp, is always 0. Thus, for each method, we calculate the CI over the estimation of $f_{\Delta P}(\Delta P)$, i.e., $f^*_{\Delta P}(\Delta P)$ for Method 1 and Method 3, and $\widetilde{f}^*_{\Delta P}(\Delta P)$ for Method 2, and check if overlaps 0, $\widetilde{0}$, and $\Delta P(v_j^k)$ for each method, respectively. If it does not overlap, we have a false positive.

In a similar way, for false negatives, we also generate several one-feature sample dataset composed of two groups, G_1 and G_2, using Bernoulli distributions with two categories, $v_0^0 = 0$ and $v_0^1 = 1$. In this case, however, we increase, for several sample size, N, the difference in proportion, Δp, for the category $v_0^1 = 1$ in both groups, G_1 and G_2. As for false positives, we calculate the CI over $f^*_{\Delta P}(\Delta P)$ for Method 1 and Method 3, and $\widetilde{f}^*_{\Delta P}(\Delta P)$ for Method 2, and we check if overlaps 0, $\widetilde{0}$, and $\Delta P(v_j^k)$, respectively. If it overlaps, we have a false negative.

Results of the experiments are presented in Fig. 2 and Fig. 3. In detail, Fig. 2 shows the false positives of each of the three methods (Method 1 in the first row, and Method 2 and 3 in the others, respectively). The left column of this figure has three 2D plots in which the logarithm of the sample size, N, is represented on one axis, the Bernoulli distribution proportion parameter p of category $v_0^1 = 1$ in each group (G_1 and G_2) is represented in the other axis, and false positives are represented with blue crosses. The right column of Fig. 2 has three 3D plots in which the logarithm of the sample size, N is represented on one axis, the Bernoulli distribution proportion parameter p of category $v_0^1 = 1$ in each group (G_1 and G_2) is represented on other, and the CI over the

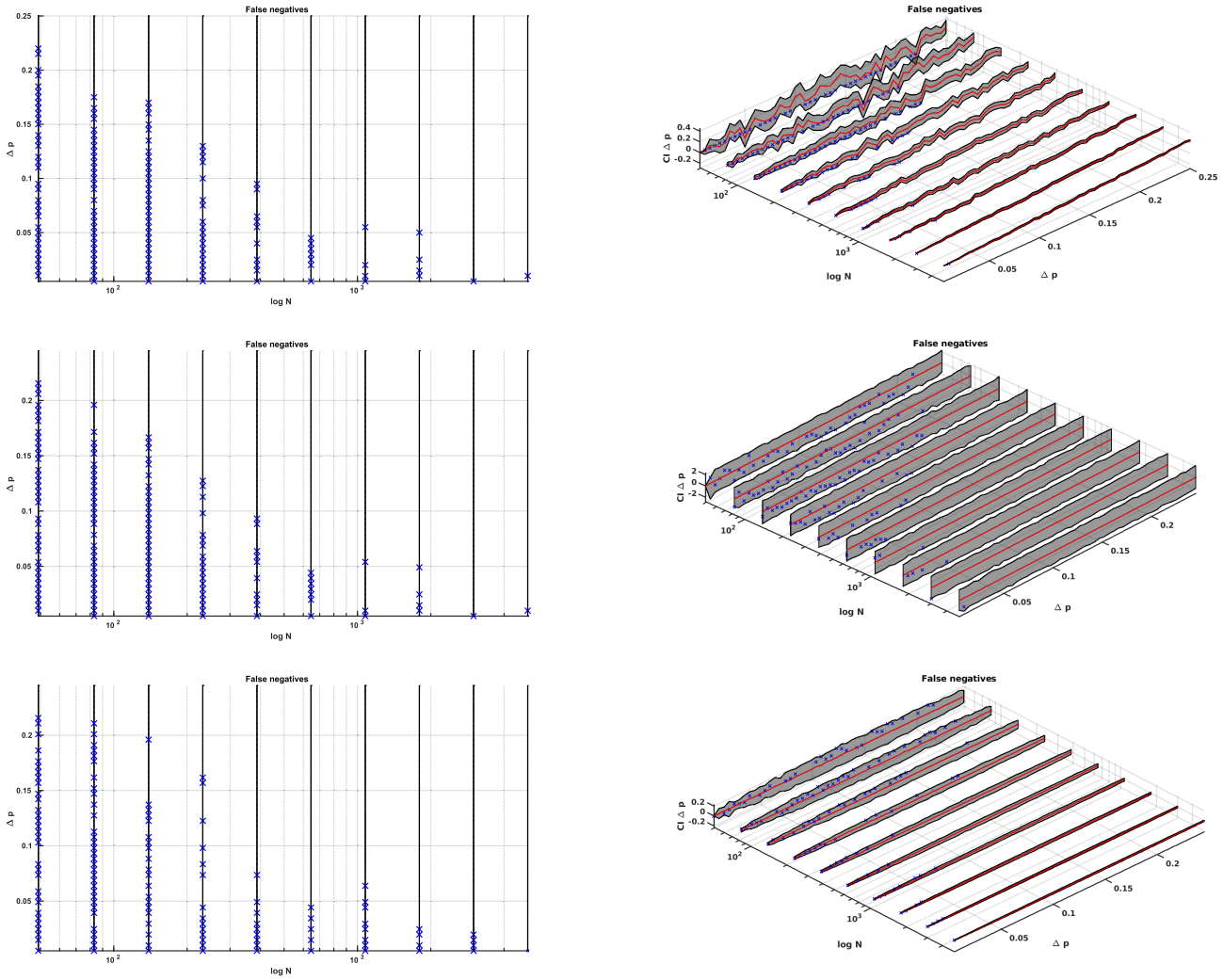

FIGURE 3. Experiments with synthetic data that show the performance of Method 1 (first row), Method 2 (second row), and Method 3 (third row) according to false negatives. In the left column, false negatives are represented with blue crosses for each of the three methods, where N is the sample size and Δp is the the difference in proportions between groups G_1 and G_2. In the right column, a 3D view of previous plots, with CI on the vertical axis depicted in grey and mean values depicted with a red line.

estimation of $f_{\Delta P}(\Delta P)$ is the vertical axis. The area formed by the CI is depicted in gray, the mean value is depicted with a red line, and the false positives are represented with blue crosses. We can observe that false positives are not distributed according to a specific pattern in any of the three methods, and it exhibits a similar profile for both small or large N. This is naturally consistent with the intuition that we establish a confidence level in our CI (here 95%), so that the well-known result of a number of experiment repetitions allows us to visualise it in these plots.

Likewise, Fig. 3 shows the false negatives produced by Method 1, Method 2, and Method 3. The three 2D plots of the left column of this figure have, on an axis, the logarithm of the sample size, N, and on the other, the difference in proportions Δp between groups G_1 and G_2, along with the false positives represented with blue crosses. The three 3D plots of the right column of Fig. 2 have, on a horizontal axis, the logarithm of the sample size, N, on the other horizontal, the difference in proportions Δp between groups G_1 and G_2, and on vertical axis, the CI over the estimation of $f_{\Delta P}(\Delta P)$. In a similar way, the area formed by the CI is showed in grey, the mean value is depicted with a red line, and the false negatives are represented with blue crosses. In this case, however, we can observe that false negatives are clearly concentrated in small sample sizes N, which is also consistent with the well-known effect of the power of the statistical-test being lower with small-sized datasets.

Shapes of the CI in the 3D plots for each method are different. However, for both false positives and false negatives, each method has a very similar shape. In Method 1 and Method 3, the CI width decreases with increasing sample sizes, N, and in the ends it is clearly seen that the CI is smaller. Though, CI of Method 1 is much noisier than Method 3. A different behavior is that of Method 2. In this case, shapes of the

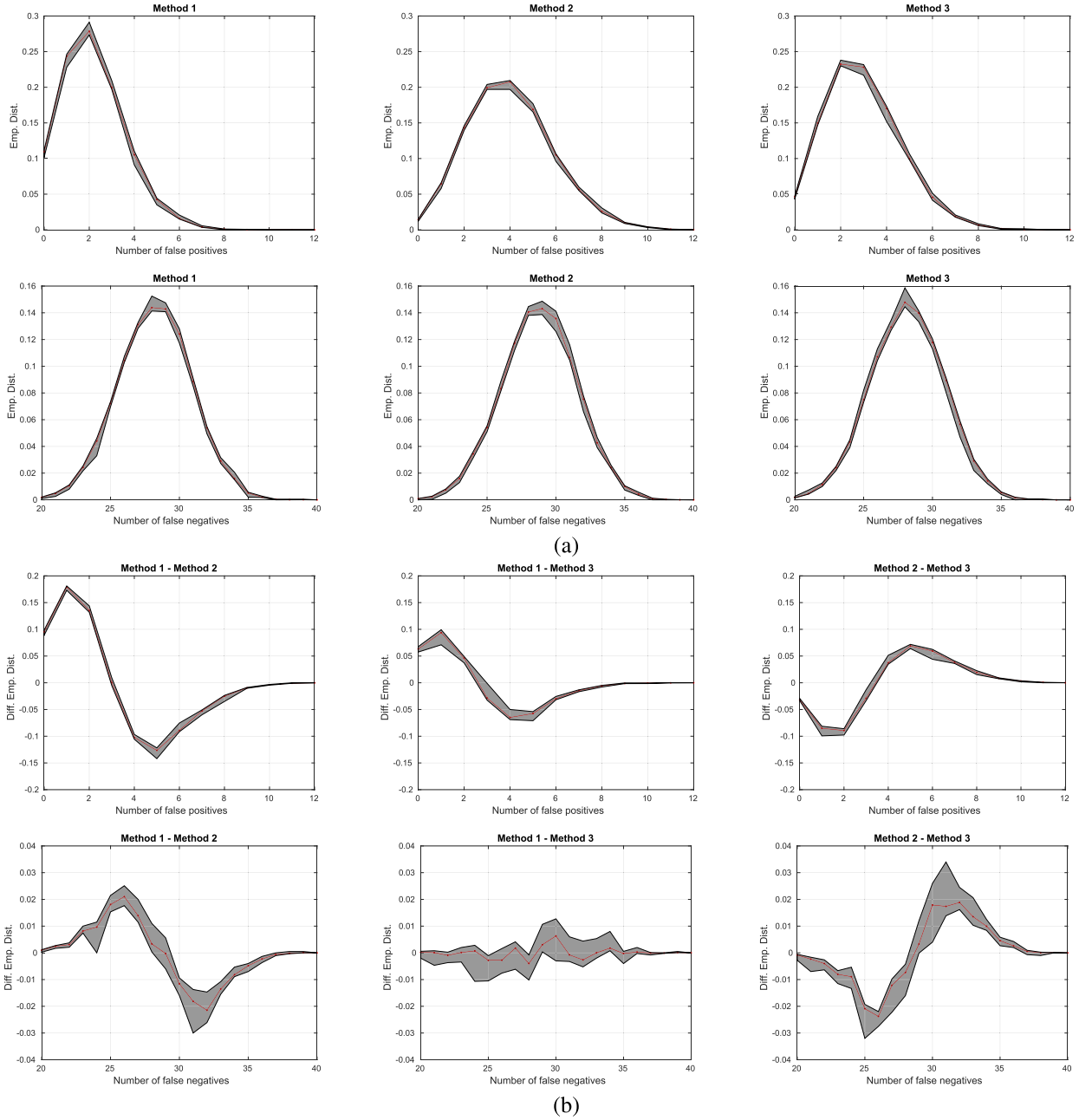

FIGURE 4. Experiments with synthetic data that quantify the number of false positives and false negatives introduced by each of the three methods. (a) In the first row, CI of empirical distribution of the number of false positives for Method 1, Method 2, and Method 3, for a sample of $N = 83$. In the second row, CI of empirical distribution of the number of false negatives for Method 1, Method 2, and Method 3, for a sample of $N = 83$. (b) In the first row, CI of the difference of empirical distributions of the number of false positives for Method 1, Method 2, and Method 3, for a sample of $N = 83$. In the second row, CI of the difference of empirical distributions of the number of false negatives for Method 1, Method 2, and Method 3, for a sample of $N = 83$.

CI are more uniform and do not change when the sample size N increases. This is due to the z-score normalization carried out.

Regarding the quantity of false positives and false negatives introduced by each of the three methods, it is not straightforward to determine whether there is some difference by visual inspection of these graphs. For this reason, we repeat the previous experiments several times and we count the number of false positives and false negatives for a sample size of $N = 83$. In the first row of Fig. 4 in Panel (a), CI of empirical distribution of the number of false positives are presented for Method 1, Method 2 and Method 3. We can observe that the distribution of Method 1 is slightly more skewed towards low values of the number of false positives than the others, i.e., Method 1 introduces a lower number of false positives. Similarly, in the second row of Fig. 4

TABLE 1. Names of Relevant Features of the Clinical Part of the OC Database, Whether They are Statistically Significant (Some of Their Categories in Categorical Types or Their Means and Standard Deviations in Metric Types), and a Description of Each One

Name	Significant?	Description
Oncological_History	No	Presence of cancer in personal medical history regardless of the type.
Gynecological_Family_History	Yes	Presence of gynecological cancer in family medical history.
Status_BRCA	No	Mutational status of the BRCA1 and BRCA2 genes.
Age_at_Diagnosis	Yes	Age at diagnosis.
Anatomical_Location	Yes	Anatomical location of the tumor.
Histology_1st_Component	Yes	Type of ovarian tumor developed by the patient.
Grade	No	Constitutes a metric feature reflecting the microscopic cell appearance abnormality of the tumoral cells, being the highest score associated to more dedifferentiated or advanced tumors.
Perineural_Vascular_Invasion	Yes	Existence of vascular or perineural invasion. It is an indicative that the tumor has begun to invade the surrounding tissues.
Stage	Yes	Stage of cancer in relation to its spread throughout the body.
Surgery	Yes	Type of surgery representing if it is primary or interval.
HIPEC_in_Surgery	No	Hyperthermic intraperitoneal chemotherapy treatment.
Type_of_Primary_Surgery	No	Type of primary surgery representing an optimal (R0) or suboptimal resection (R1) of the tumor.
Neoadjuvance	Yes	Chemotherapy treatment prior to primary surgery.
Response_of_Neoadjuvance	Yes	Observed response to neoadjuvant chemotherapy.
Attitude_of_Interval_Surgery	Yes	Decision after neoadjuvancy.
Type_of_Interval_Surgery	Yes	Type of interval surgery representing an optimal (R0) or suboptimal resection (R1) of the tumor.
Adjuvance	No	Chemotherapy treatment after the primary surgery.
Response_of_Adjuvance	Yes	Observed response to adjuvant chemotherapy.
PFS	Yes	Progression free survival. It is the time from the first date of pharmacological treatment until radiological or biochemical progression.
PFI	Response variable	Platinum free interval. It is the time between the last cycle of platinum and evidence of disease progression; depending on the length of platinum drugs sensitivity, patients could be categorized as platinum resistant (<6 months) or sensitive (>6 months).
OS	Yes	Overall survival. It estimates the duration of patient survival from the date of diagnosis or treatment initiation.
Bevacizumab_Maintenance	No	Antiangiogenic treatment.

in Panel (a), CI of empirical distribution of the number of false negatives are shown for the three methods, though in this case, we can observe that distribution of Method 1 and Method 3 are quite similar. Comparison between distributions of each method can be seen in more detail in Fig. 4 in Panel (b). In this figure, it can be assessed that for false positives (first row), Method 1 introduces less false positives, since the difference distribution of the number of false positives for *Method 1 - Method 2* and for *Method 1 - Method 3* is greater than 0 in low values. However, for false negatives (second row in Panel (b)), we can observe that the difference distribution of the number of false negatives for *Method 1 - Method 3* is around 0, which means that these two methods behave similarly for false negatives. Accordingly, we can conclude that Method 1 provides with a better behavior in terms of false positives, and it will be the resampling method for our analysis framework.

B. RESULTS WITH OC DATABASE

With the aim of finding relationships between clinical and genetic factors and the OC disease progression, we explore, in this part, the database presented in Section II. For this, we make use of the analysis framework described in Section III along with the Method 1 algorithm (see Algorithm 1) as bootstrap resampling strategy.

Since the analysis framework requires difference-based statistics, we separate the OC database into two interest groups based on an indicator of disease progression. Concretely, we use the platinum-free interval (PFI), which is defined as the time (in months) between the last cycle of platinum and evidence of disease progression [20]. In this sense, depending on the length of platinum drugs sensitivity, patients could be categorized as platinum resistant (<6 months) or platinum sensitive (>6 months).

The analysis framework is implemented in custom software [9] and coded in MATLAB (MathWorks Inc.). The output of the software is $f^*_{\Delta P}(\Delta P)$ for categorical variables, and $f^*_{\Delta m}(\Delta m)$, $f^*_{\Delta\sigma}(\Delta\sigma)$, and $\Delta f^*_{M_j}$ for metric variables. With these estimations, we detect significant differences between platinum resistant and platinum sensitive groups, namely: (1) If the CI of $f^*_{\Delta P}(\Delta P)$, $f^*_{\Delta m}(\Delta m)$, $f^*_{\Delta\sigma}(\Delta\sigma)$, or $\Delta f^*_{M_j}$ overlaps 0, this denotes that the proportion, mean, standard deviation, or *pdf* in platinum sensitive group is similar to that of platinum resistant group, so this feature (or feature category) has no particular bias or additional information; (2) If the CI of $f^*_{\Delta P}(\Delta P)$, $f^*_{\Delta m}(\Delta m)$, $f^*_{\Delta\sigma}(\Delta\sigma)$, or $\Delta f^*_{M_j}$ does not overlaps 0 and it is located at positive values, this indicates that proportion, mean, standard deviation, or *pdf* is larger in platinum sensitive group, which means that this feature (or feature category) is a relevant property of this group; And (3), if the CI of $f^*_{\Delta P}(\Delta P)$, $f^*_{\Delta m}(\Delta m)$, $f^*_{\Delta\sigma}(\Delta\sigma)$, or $\Delta f^*_{M_j}$ does not overlaps 0 and it is located at negatives values, this denotes that proportion, mean, or standard deviation is bigger in platinum resistant group, and therefore this feature (or feature category) is a relevant property of this group.

All features of the OC database have been analysed with a confidence level of 95%. However, some of them have no clinical relevance, for example, the identifier of each patient or the platform on which the genome has been sequenced. For this reason, we focus on results obtained from a set of features that expert clinicians consider most relevant. Name of the each of these relevant features along with a small description of each one and with information of whether there are statistically significant differences between groups for this features (some of its categories in categorical type or its mean and standard deviation in metric type) are presented in Table 1 and Table 2 for clinical and genetic parts, respectively.

In the clinical part of the database, we find that some features do not present statistically significant differences between platinum sensitive and platinum resistant groups. This is the case of the features that represent the presence of cancer in personal medical (*Oncological_History*), the mutational status of *BRCA1* and *BRCA2* genes (*Status_BRCA*), the grade of the tumor (*Grade*), the hyperthermic intraperitoneal chemotherapy (HIPEC)

TABLE 2. Names of Relevant Features of the Genetic Part of the OC Database, Whether They are Statistically Significant (Some of Their Categories in Categorical Types or Their Means and Standard Deviations in Metric Types), and a Description of Each One

Name	Significant?	Description
HGNC_Symbol	No	Gene Nomenclature (Human Genome Nomenclature Committee).
Chr	No	Chromosome.
Genetic_Change	No	Genetic change from the reference allele to the variant allele.
Genotype	Yes	Genotype.
VarDepth	Yes	Number of times that reading a specific region, the variant allele has been read.
Conservation_Score	No	Conservation of the region under study at an evolutionary level.
Grantham_Distance	No	Variable that reflects how different are the amino acids that are changed in missense mutations.
Condel_Prediction	No	Prediction of pathogenicity of the variant according to the Condel tool.
Condel_Prediction_Score	No	Score of the degree of pathogenicity of the variant according to the Condel tool.
Sift_Prediction	No	Prediction of pathogenicity of the variant according to the Sift tool.
Sift_Prediction_Score	No	Score of the degree of pathogenicity of the variant according to the Sift tool.
PolyPhen_Prediction	No	Prediction of pathogenicity of the variant according to the PolyPhen tool.
PolyPhen_Prediction_Score	No	Score of the degree of pathogenicity of the variant according to the PolyPhen tool.
Impact	No	Pathogenicity prediction.
Amino_Acids	No	Reference amino acid to amino acid variant translated by the variant.
PFI	Response variable	Platinum free interval. It is the time between the last cycle of platinum and evidence of disease progression; depending on the length of platinum drugs sensitivity, patients could be categorized as platinum resistant (<6 months) or sensitive (>6 months).

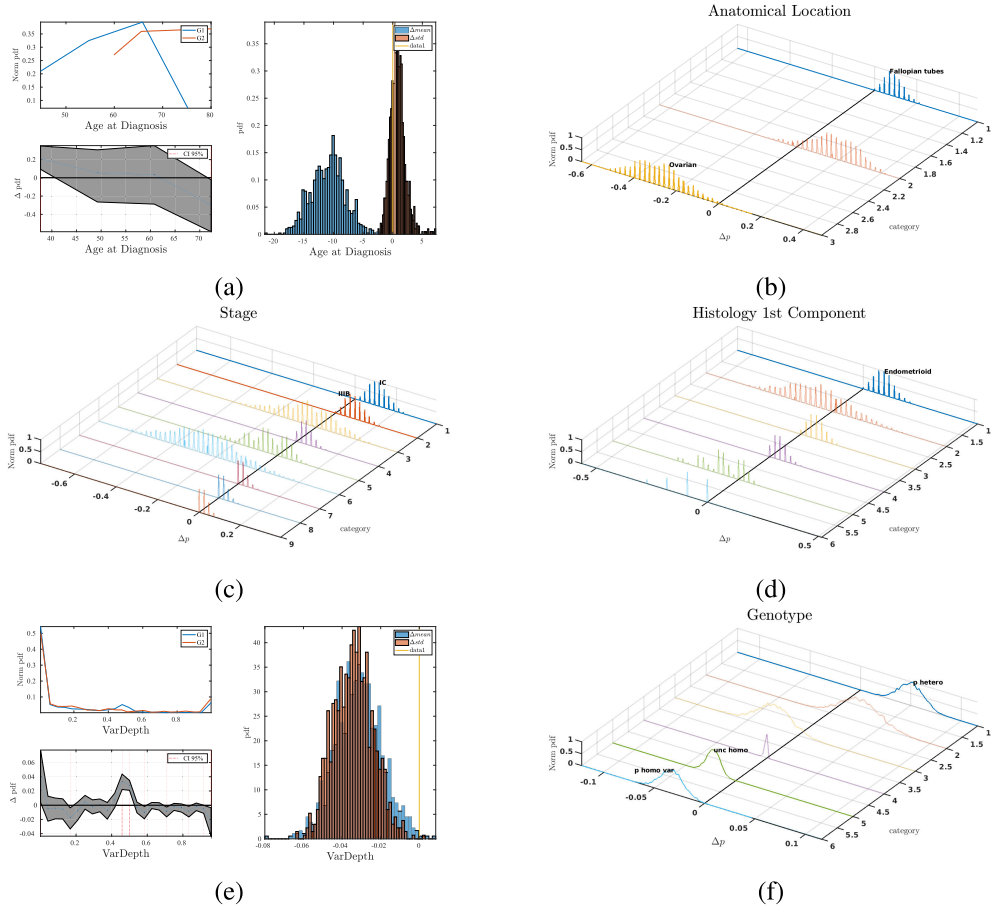

FIGURE 5. Some features of the OC database analysed with the proposed framework: (a) Age at diagnosis of patient; (b) Anatomical location of tumor; (c) Stage of the tumor; (d) Type of ovarian tumor; (e) Features representing the number of times that reading a specific region, the variant allele has been read; (f) Genotype.

treatment (*HIPEC_in_Surgery*), the type of primary surgery (*Type_of_Primary_Surgery*), or the chemotherapy treatment after the primary surgery (*Adjuvant*). However, there are many other features that do have significant differences between the two groups. For metric features, for example, the age at diagnosis (*Age_at_Diagnosis*) is significant for the difference in means between both groups, being ages greater for platinum resistant group. For the progression free survival (*PFS*) and the overall survival (*OS*), there are also significant differences in means between both groups, but, in these cases, values are higher for platinum sensitive group. In this three features, results are as expected. For categorical features, we have quite a few examples: (1) In the information about the presence of gynecological cancer in family medical history (*Gynecological_Family_History*), the *Yes* category is significant for the platinum sensitive group, and the *No* category is for platinum resistant group; (2) The anatomical location of the tumor (*Anatomical_Location*) has two

significant categories, namely, *Ovarian*, which is significant for the platinum resistant group, and *Fallopian tubes*, which is significant for the platinum sensitive group; (3) In the type of ovarian tumor (*Histology_1st_Component*), *Endometrioid* category is relevant to platinum sensitive group; (4) In the existence of vascular or perineural invasion (*Perineural_Vascular_Invasion*), *No* is a relevant category of platinum sensitive group; (5) The stage of cancer (*Stage*) has two relevant categories, both significant for the platinum sensitive group, namely, *IC* and *IIIB*. The first implies tumor growth limited to the ovary with signs of malignancy (tumor capsule ruptured or presence of malignant cells in peritoneal fluid), while the later includes local spread to the peritoneal cavity (peritoneal implants diameter <2cm); (6) The type of surgery (*Surgery*) has two significant categories, specifically, *Primary*, which represent the surgery as the first therapeutic action since there is no initial treatment with chemotherapy, and *Interval*, which represent the surgery as the therapeutic action after chemotherapy treatment (neoadjuvant treatment). The former is significant for platinum sensitive group and the latter for platinum resistant group; (7) In the chemotherapy treatment prior to primary surgery (*Neoadjuvance*), *Yes* category is significant for platinum resistant group and *No* category for platinum sensitive group; Or (8), the response to neoadjuvant chemotherapy (*Response_of_Neoadjuvance*) has one significant category, which is *PR*, means partial response to neoadjuvant chemotherapy treatment. This category is significant for platinum resistant group. In all these features, as for metric features, results are as expected.

In the genetic part of the database, we find only two features that present statistically significant differences between platinum sensitive and platinum resistant groups. These are the genotype feature (*Genotype*) and the variable which represents the average number each DNA alteration is read in the genomic sequencing profiling (*VarDepth*). In the first one, *p_homo_var* an *unc_homo* categories are significant for platinum resistant group, while *p_hetero* is significant in the sensitive subset. At this regard, *p_homo_var* reflects the exclusively presence of variant alleles (not normal genotypes) in the tumors under study and the association of such alterations with platinum agents resistance could be explained by the loss of wild type alleles (loss of heterozigosity) or genetic variants gains/amplifications in the more advanced tumors.

With respect to *VarDepth*, the higher the sequencing depth is for an specific genetic alteration, the greater the certainty for assigning such variant as real and the possibility of considering it as a clonal or trunk alteration (opposite to passenger genetic changes). The mean difference between resistant vs. sensitive patients was significant, being, as expected, higher for the platinum resistant set.

In Fig. 5, we show some of the features described above. For representation purposes in categorical features, and for significant categories, each $f^*_{\Delta P}(\Delta P)$ is plot in thicker line stile, and also the category label is displayed at the top of it. Thereby, this allows us to scrutinize the most relevant categories for each feature. For metric features, in addition to representing $f^*_{\Delta m}(\Delta m)$ and $f^*_{\Delta \sigma}(\Delta \sigma)$, normalised $f^*_{M_j}(M_j|G_1)$ (blue) and $f^*_{M_j}(M_j|G_2)$ are also represented (G_1 corresponds to platinum sensitive group and G_2 to platinum resistant group). Finally, the CI of $\Delta f^*_{M_j}$ is also represented, in grey.

V. DISCUSSION AND CONCLUSION

In the first part of the article, with the aim of unifying the univariate statistical descriptions of different types of features, an analysis framework that consists of an hypothesis test based in bootstrap resampling technique for different data types have been proposed. Three boostrap resampling strategies, called in the text Method 1, Method 2, and Method 3, have been analysed in order to determine the performance of each method. For this, it had been taken into account the amounts of false positives and false negatives generated in several experiments with synthetic data, concluding that Method 1 have a better behavior.

It is a matter of discussion why we use a non-parametric bootstrap resampling strategies if our dataset is small (54 entries) and bibliography said that parametric bootstrap methods work better, in general, with smaller sample sizes [17], [21]. This statement is based on the fact that whether the original small sample has outliers, but these are absent from the population sampled, may be reproduced in the simulated data. However, our database has been found to have a lack of extreme outliers, therefore, it is not a problem to use non-parametric bootstrap. In addition, parametric models have the problem that use an inherently arbitrary choice of model, since it is not easy to select the most appropriate mathematical function *a priori* [21]. This is the main reason for opting, in our case, for the use of non-parametric bootstrap methods.

In the second part of the article, we make use of the proposed analysis framework along with the Method 1 bootstrap resampling technique to explore the OC database. The goal is to try to discover relationships between clinical and genetic factors and the disease progression. Specifically, we explore relations between features and the platinum-free interval (PFI), categorized into platinum resistant (<6 months) or platinum sensitive (>6 months) groups. Results show that, among a set of relevant features indicated by clinicians, the clinical part of the OC database has many more significant variables than the genetic part.

The literature on OC and (Big) Data Science shows that many works are related to genetic databases and prediction of classifiers. Representative examples of this are the works of Wang *et al.* [22] and Yasodha *et al.* [23]. The former, based on gene expression data and prognostic data of OC patients from The Cancer Genome Atlas dataset, used conditional mutual information to construct a gene dependency network to identify the gene signature that can predict the prognostic risks of OC patients. The latter tried to detect OC using Big Data analysis. Specifically, they proposed an approach for identifying OC in a dataset of proteomic spectra, first, by using an algorithm for feature selection, and second,

by using this features for a classification task. Also, results were compared with other classification methods, such as Support Vector Machine (SVM), Multilayer Perceptron, and Feed Forward Neural Network. However, few works have been found in OC focusing on data quality.

In this work, some features have resulted individually in differences between both platinum sensitive and platinum resistant groups. In principle, this is an indicator that the database could be discriminatory for the hypotheses studied, being this is a good sign. However, it is necessary to make multivariate analyses. Features with relevant differences could be extremely correlated or redundant, and therefore the data classification could be reduced, or features can be coinformative and provide much better classification in a multivariate. In addition, in this work we have not analysed a very relevant type of feature, which is the text of medical comments.

It is highly convenient to expand the use of resampling techniques for textual features. Furthermore, the principles seen here should be scaled up to linear multivariate analysis, using analysis of principal components for subsets of metric variables, analysis of multiple correspondences for subsets of categorical variables, and the combination of both. We explore these points in another work [24]. Exploring non-linear multivariate analysis methods through emerging approaches such as autoencoders, as well as machine learning method specifically tailored to extract the most informative features in a database and their interactions, should also be addressed.

AUTHOR CONTRIBUTIONS

L.B.-C., S.M.-R. and J.L.R.-Á. designed and organized the article. L.B.-C. performed the data analysis and wrote the methods and part of the introduction, results, and conclusion. S.R-L. wrote the database description and parts of the introduction, results, and discussion. S.M.-R., J.G.-D, and J.L.R.-Á. contributed writing some parts and reviewing the manuscript. M.Y.-F. and A.B. provided the clinical and genetic data processing. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

Authors declare no conflict of interest.

REFERENCES

[1] M. Moschetta, A. George, S. B. Kaye, and S. Banerjee, "BRCA somatic mutations and epigenetic BRCA modifications in serous ovarian cancer," *Ann. Oncol.*, vol. 27, no. 8, pp. 1449–1455, Aug. 2016.

[2] C. Stewart, C. Ralyea, and S. Lockwood, "Ovarian cancer: An integrated review," *Seminars Oncol. Nursing*, vol. 35, no. 2, pp. 151–156, Apr. 2019.

[3] J. Millstein, T. Budden, E. L. Goode, and M. S. Anglesio, "Prognostic gene expression signature for high-grade serous ovarian cancer," *Ann. Oncol.*, vol. 31, no. 9, pp. 1240–1250, 2020.

[4] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big Data*, vol. 1, no. 1, pp. 51–59, Mar. 2013.

[5] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *J. Big Data*, vol. 2, no. 1, pp. 1–21, Dec. 2015.

[6] L. Bote-Curiel, S. Muñoz-Romero, A. Gerrero-Curieses, and J. L. Rojo-Álvarez, "Deep learning and big data in healthcare: A double review for critical beginners," *Appl. Sci.*, vol. 9, no. 11, p. 2331, Jun. 2019.

[7] J. L. Rojo-Álvarez, "Big and deep hype and hope: On the special issue for deep learning and big data in healthcare," *Appl. Sci.*, vol. 9, no. 20, p. 4452, Oct. 2019.

[8] J. R. Feijoo-Martinez, S. Munoz-Romero, C. Soguero-Ruiz, M. Castro-Fernandez, and J. L. Rojo-Alvarez, "Event analysis on power communication networks with big data for maintenance forms," *IEEE Access*, vol. 6, pp. 72263–72274, 2018.

[9] P. Talón-Ballestero, L. González-Serrano, C. Soguero-Ruiz, S. Muñoz-Romero, and J. L. Rojo-Álvarez, "Using big data from customer relationship management information systems to determine the client profile in the hotel sector," *Tourism Manage.*, vol. 68, pp. 187–197, Oct. 2018.

[10] T. Hastie, R. J. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[11] M. Greenacre and J. Blasius, *Correspondence Analysis and Related Methods*. Boca Ratón, FL, USA: Chapman & Hall, 2006.

[12] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: A statistical framework," *Int. J. Mach. Learn. Cybern.*, vol. 1, nos. 1–4, pp. 43–52, Dec. 2010.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[14] M. Rodriguez-Ibanez, S. Munoz-Romero, C. Soguero-Ruiz, F.-J. Gimeno-Blanes, and J. L. Rojo-Alvarez, "Towards organization management using exploratory screening and big data tests: A case study of the spanish red cross," *IEEE Access*, vol. 7, pp. 80661–80674, 2019.

[15] S. Muáoz-Romero, "Distribution-test estimation and topic modeling from big-data analytics from heterogeneous hospital records in ovarian cancer," in *Proc. Spanish Biometric Conf. VII Ibero-Amer. Biometric Meeting*, 2019, p. 91.

[16] S. M. Ross, *Introductory Statistics*, 4th ed. Oxford, U.K.: Academic, 2017.

[17] B. Efron and R. J. Tibshirani, *An Introduction to Bootstrap*, 1st ed. Boca Ratón, FL, USA: Chapman & Hall, 1993.

[18] D. T. Kaplan, *Resampling Stats MATLAB*, 1st ed. Arlington, VA, USA: Resampling Stats, 1999.

[19] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan, "A scalable bootstrap for massive data," *J. Roy. Stat. Soc., Ser. B Stat. Methodol.*, vol. 76, no. 4, pp. 795–816, Sep. 2014.

[20] E. Pujade-Lauraine and P. Combe, "Recurrent ovarian cancer," *Ann. Oncol.*, vol. 27, pp. 63–65, Aug. 2016.

[21] P. Hall, "Theoretical comparison of bootstrap confidence intervals," *Ann. Statist.*, vol. 6, no. 3, pp. 927–963, 1988.

[22] J.-Y. Wang, L.-L. Chen, and X.-H. Zhou, "Identifying prognostic signature in ovarian cancer using DirGenerank," *Oncotarget*, vol. 8, no. 28, pp. 46398–46413, Jul. 2017.

[23] Y. P and A. Nr, "Detecting the ovarian cancer using big data analysis with effective model," *Biomed. Res.*, pp. S309–S315, 2018.

[24] L. Bote-Curiel, "Text analytics and mixed feature extraction in ovarian cancer clinical and genetic data," *IEEE Access*, 2021.

LUIS BOTE-CURIEL received the degree in telecommunication engineering from the Universidad de Valladolid, Spain, in 2012. His research interests include machine learning and deep learning, focused on interpretability and their application to the healthcare field.

SERGIO RUIZ-LLORENTE received the degree in biology from the Universidad de Alcalá, in 2000, and the Ph.D. degree in human genetics from the Spanish National Cancer Center (CNIO, 2005). Later on, he worked in different national and international research centers (Instituto de Investigaciones Biomédicas-UAM, Memorial Sloan Kettering Cancer Center, and HM-CIOCC). He has proved experience in the use of genetic diagnostic tools, the applicability and management of high throughput molecular platforms, and the development of *in vivo* and *in vitro* preclinical assays. This research activity has resulted in the publication of 25 scientific articles in international journals, seven of them as the first author.

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

MÓNICA YAGÜE-FERNÁNDEZ received the degree in biotechnology and the master's degree in clinical and applied research in oncology from Universidad CEU San Pablo, in 2018 and 2019, respectively, and the master's degree in administration and management of pharmaceutical, biotechnological and health companies from the CESIF, in 2020. Nowadays, she is developing her career and working at the Clara Campal Comprehensive Oncology Center (CIOCC), focusing on gynecological, genitourinary, and skin tumors from a clinical and basic point of view, developing clinical trials and research studies.

ARANTZAZU BARQUÍN received the degree in medical oncology in 2019, the master's degree in medical oncology sponsored by the Spanish Medical Oncologist Association (SEOM) and in oncology molecular biology sponsored by the National Centre of Oncological Investigation (CNIO).

She is currently enrolled at Centro Oncologico Clara Campal, Spain, in the Gynaecological, Genitourinary and Skin Tumour Unit and an Active Member of the Laboratory of Translational and Innovation in Oncology. She successfully passed the ESMO Examination in 2017, and completed a short rotation at the Princess Margaret Cancer Center, Toronto, ON, Canada, in 2018. Her current research interest includes investigation in ovarian cancer.

JESÚS GARCÍA-DONAS is currently a Medical Oncologist and also the Head of the Unit of Gynecological, Genitourinary and Skin Tumors, Clara Campal Comprehensive Cancer Center. With extensive experience in the realization of clinical trials, he has participated in more than 200 studies and also designed and directed multiple clinical trials without commercial interest whose promoters have been cooperative groups. The group he leads published more than 40 articles in international scientific journals of first level. He received awards in recognition of their activity (Merck Foundation Awards to the Research in Rare Diseases) and public and private fellowships.

JOSÉ LUIS ROJO-ÁLVAREZ (Senior Member, IEEE) received the B.Sc. degree in telecommunication engineering from the Universidade de Vigo, in 1996, and the Ph.D. degree from the Universidad Politécnica de Madrid, in 2000. He is currently a Professor in the Department of Signal Theory and Communications and Telematic Systems and Computation, Universidad Rey Juan Carlos, Spain. He has coauthored more than 140 international articles and has contributed to more than 180 conference proceedings. His research interests include statistical learning methods for signal and image processing, arrhythmia mechanisms, robust signal processing methods, and data science for oncology.