




OPEN

Discovering HIV related information by means of association rules and machine learning


Lourdes Araujo^{1,4}, Juan Martinez-Romo^{1,4}, Otilia Bisbal²,
Ricardo Sanchez-de-Madariaga^{3,4} & The Cohort of the National AIDS Network (CoRIS)*

Acquired immunodeficiency syndrome (AIDS) is still one of the main health problems worldwide. It is therefore essential to keep making progress in improving the prognosis and quality of life of affected patients. One way to advance along this pathway is to uncover connections between other disorders associated with HIV/AIDS—so that they can be anticipated and possibly mitigated. We propose to achieve this by using Association Rules (ARs). They allow us to represent the dependencies between a number of diseases and other specific diseases. However, classical techniques systematically generate every AR meeting some minimal conditions on data frequency, hence generating a vast amount of uninteresting ARs, which need to be filtered out. The lack of manually annotated ARs has favored unsupervised filtering, even though they produce limited results. In this paper, we propose a semi-supervised system, able to identify relevant ARs among HIV-related diseases with a minimal amount of annotated training data. Our system has been able to extract a good number of relationships between HIV-related diseases that have been previously detected in the literature but are scattered and are often little known. Furthermore, a number of plausible new relationships have shown up which deserve further investigation by qualified medical experts.

According to information provided by World Health Organization (WHO), HIV/AIDS remains one of the world's most serious public health problems, particularly in low and middle-income countries. The development of AIDS (acquired immunodeficiency syndrome) disease, in patients infected with HIV, causes a progressive deterioration of the immune system and decreases the person's ability to fight many infections and other diseases as well. AIDS refers to the most advanced stages of HIV infection and is defined by the development of one or more opportunistic infections or related cancers among many other possibilities.

WHO has released a number of policy guidelines to assist countries in implementing programs to improve HIV prevention, treatment, care and support services for affected patients. Several initiatives are underway in each country along these lines. One of these initiatives in Spain has been the launching of the Spanish HIV/AIDS Research Network whose main goal consists of improving the health and quality of those affected. This network has generated the HIV/AIDS Research Network Cohort (CoRIS), which makes available to researchers the data from its main database and associated satellite databases, linking biological samples. CoRIS is an open, prospective and multicenter cohort of adult subjects with confirmed HIV infection, launched in 2004. Patients over 13 years old, and naive to antiretroviral treatment (ART) at study entry, have been recruited in HIV care units of the Spanish Public Health System and all of them have signed an informed consent form.

HIV is associated with the development of a large number of other diseases. In some cases, these diseases can be more or less mild and transient. However, in other cases, they can be very serious and long-lasting. It is essential to know the possible relationships between this diversity of diseases associated with HIV, since increasing the knowledge about them can be of great help in their diagnosis and prevention, thus improving the patients' quality of life. Knowledge about the conditions that typically appear with these diseases, and in what form, will help in making decisions about their prevention and treatment.

¹Languages and Information Systems Dpt., ETS Ingeniería Informática (UNED), Juan del Rosal 16, 28040 Madrid, Spain. ²Hospital Universitario 12 de Octubre, Instituto de Investigación I+12, Madrid, Spain. ³Telemedicine and e-Health Research Unit, Instituto de Salud Carlos III, Monforte de Lemos 5, 28029 Madrid, Spain. ⁴Instituto Mixto UNED-ISCI III IMIENS, 28029 Madrid, Spain. *A list of authors and their affiliations appears at the end of the paper. email: lurdes@lsi.uned.es

In this study, we have focused on applying machine learning (ML) techniques to extract information about the relationships between diseases associated with HIV. Specifically, we have turned to the extraction of association rules of high reliability and coverage to identify relevant relationships between HIV-related diseases.

Association rules (ARs)¹ are a data mining method that aims to discover patterns of co-occurrence between items in a transactional database. Specifically, we consider a set of n items $I = \{i_1, i_2, \dots, i_n\}$ and a set or database of transactions on these items: $T = \{t_1, t_2, \dots, t_m\}$. Each transaction is represented by a subset of items ($T = \{i_1, i_2, \dots, i_d\}$) that have occurred simultaneously.

ARs present the following form:

$$X \Rightarrow Y$$

where X and Y are two disjoint sets of items. We focus on a particular type of rules whose consequent is a single element.

An example of an association rule, that we could find in a database in which the transactions are the number of diseases suffered by the same patient, could be the following:

Urinary tract infection, abdominal pain, diabetes \Rightarrow renal failure

ARs have been frequently applied to the medical domain for different purposes. Imamura et al.² applied them to find clinical findings associated with diseases. They have also been used to analyze patterns of lifestyle risk behaviors including smoking, heavy drinking, physical inactivity³. There have also been proposals that applied ARs to find relationships between healthcare parameters and specific diseases⁴, such as antimicrobial resistance⁵, psoriasis⁶, COVID-19⁷, or Hospital-Acquired Infections⁸, among others.

Several algorithms have been proposed to systematically generate all ARs that satisfy certain favorable conditions. These conditions refer to parameters such as support (how frequent an itemset is in the transaction set) and confidence (the likeliness of occurrence of consequents in the set, given that the set already has the antecedents).

One of the algorithms that allow the generation of association rules from frequent itemsets is the Frequent-Pattern Growth or FP-Growth algorithm⁹.

By applying this or similar algorithms, we can generate the whole set of rules that satisfy the specified minimum support and minimum confidence thresholds. However, this process leads to a huge number of rules, many of which are uninteresting. Actually, discovering interesting or relevant rules is a difficult problem^{10–12} that needs to be tackled. For example, considering the following AR:

pollen allergy \Rightarrow renal failure

Since pollen allergy is a very common problem, it can appear in many rules as an antecedent or a consequent. Therefore, ARs, such as the one appearing above, may meet the required frequency threshold, and yet not provide relevant information.

A relevant rule is one that includes at least one relevant relationship between some disease of the antecedent and the one of the consequent. Relevant relationships are those validated by medical experts.

Filtering relevant or interesting rules is a difficult problem that is primarily tackled using unsupervised approaches that do not require expert-annotated training data. These methods attempt to find hidden patterns from raw unlabeled data. Among these unsupervised approaches are those based on associating to the AR a p-value (the likelihood that the association is spurious due to chance)^{13,14}. Specifically, the p-value of an AR R is the probability of observing R , or one rule stricter than R , when the two sides of R are independent. If a rule found in the data has a low p-value, it is unlikely that the two sides are independent. Rules with high p-values do not provide information about the independence of the two sides of the rule and can be discarded, as they have most likely appeared by pure chance.

However, unsupervised methods have a limit to the accuracy they can achieve. These limitations can be addressed by using supervised methods that are capable of learning from specific relationship examples, taking into account aspects beyond frequencies.

Supervised methods require labeled data that are used to extract knowledge. These methods start by applying a training process with a labeled data set and try to infer a function that fits the training data appropriately. Then, when applied on new data, this function is able to predict the output. In the case of ARs, few supervised systems have been proposed due to the lack of labeled data.

One way to circumvent this problem is to use the semi-supervised¹⁵ approach, which employs both labeled and unlabeled data in the training process. This approach typically uses a small amount of labeled data and a larger amount of raw data. Techniques based on this approach can be adjusted to improve their performance as they have larger amounts of training data, for example in a feedback process.

Sánchez-de-Madariaga et al.¹⁶ proposed a new semi-supervised data mining model, EXTRA, that combines unsupervised techniques (p-value computed as Fisher's exact test) with highly limited supervision. The training process starts with a small seed of annotated data, and the model improves its results (F-measure) using a fully supervised system (standard supervised machine learning algorithms). The key idea of this proposal is to enlarge the size of the training data by checking the agreement between the predictions of the supervised system and those of the unsupervised techniques in a series of iterative steps. The remarkable feature of this system is its ability to improve the results of purely supervised methods by combining them with unsupervised techniques. This system has been evaluated on data from the medical domain. Specifically, it has been evaluated on a set of ARs generated from primary care data, which have been manually labeled as true or false (of no interest). The diseases considered were rather common as they came from primary care data and the association rules generated are of limited interest. However, the results obtained showed the potential of the method designed.

The main goal of this study is to find poorly understood relationships between HIV-associated diseases using association rules. To this end, we propose to apply and adapt the semi-supervised algorithm with minimal supervision¹⁶ mentioned above, to filter association rules between HIV-related diseases. To apply the algorithm

FILTERING ASSOCIATION RULES (ARs) OF INTEREST

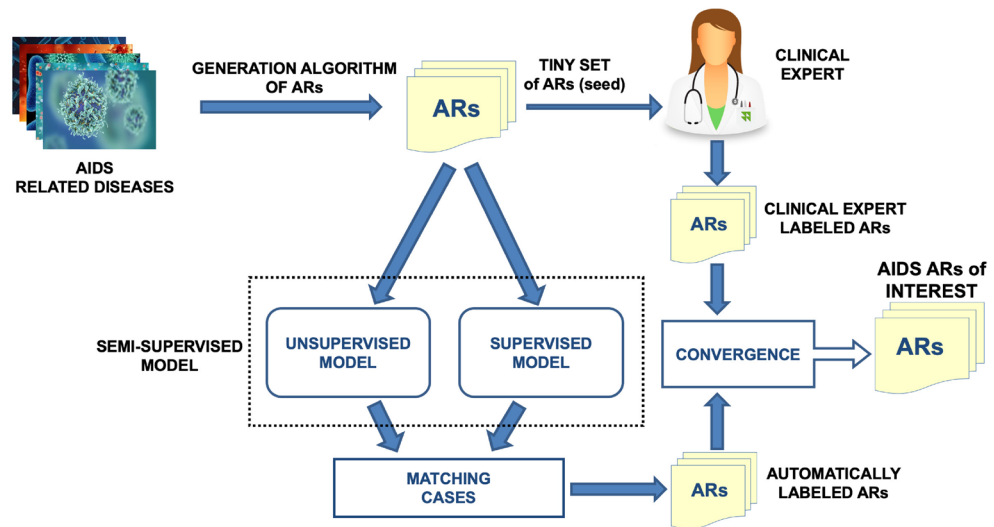


Figure 1. Scheme of EXTRAEE, the semi-supervised learning model for the filtering of relevant ARs among HIV-related diseases.

and also to be able to perform a proper evaluation, we have started by generating a dataset of association rules and then manually labeling a part of them as relevant or not by expert doctors in HIV. The rules considered by the doctor come from applying the FP-Growth algorithm to the CoRIS cohort data. The input data to our algorithm are ARs annotated with a label indicating whether they are relevant or not.

Since, to the best of our knowledge, there is little history of annotation of ARs as relevant or not, this annotation process has required the definition of a new annotation guideline, this being a contribution of the present study. Another important contribution is the associations found between diseases themselves. We also provide a highly accurate method for the classification of ARs as relevant or not, which has been evaluated on the previously created dataset.

In summary, the main contributions of the present study are outlined as follows:

- An analysis of the most appropriate conditions for assigning the relevance of the rules based on the relationships between the diseases contained in them.
- A collection of association rules between HIV-associated diseases labeled as relevant or not by experts.
- Design of a semi-supervised system that requires only a very small amount of annotated data and is able to predict with high accuracy whether new ARs for HIV-associated diseases are relevant or not.
- An expert analysis of various relationships between HIV-related diseases revealed by the ARs found.

Figure 1 shows a scheme of the applied semi-supervised system EXTRAEE. From the data on the diseases suffered by each patient in the CoRIS database, all the ARs that satisfy minimum conditions of co-occurrence frequency (support and confidence) are generated. A medical expert evaluates a small set of these rules, which serves as a seed for the algorithm, as being relevant or not. That initial seed serves as a training set for a supervised algorithm. It also allows for certain adjustments of an unsupervised algorithm. When the predictions on the ARs of both algorithms match, they are considered sufficiently reliable to be added as reference data to the training set. The process is repeated until convergence is reached when no new ARs are added to the training set.

Our analysis has uncovered multiple relationships of interest between HIV-related diseases. Some of these relationships are well known to medical experts. Others are little known and, even though they could be confirmed in the literature, they were scattered and hence their compilation may represent a breakthrough in HIV research. Finally, other relationships, although plausible, are neither confirmed nor discarded in the literature and deserve to be studied in depth by medical professionals.

Methods

In this section, we include details of the reference collection on HIV-related diseases. We also introduce the process followed to analyze and annotate the association rules to be used to train and evaluate the semi-supervised model. Finally, we present the semi-supervised model EXTRAEE applied for ARs filtering.

CORIS data. The Spanish HIV/AIDS research network (CoRIS) is an open, prospective, multicenter cohort of adult subjects with confirmed HIV infection, naïve to ART at study entry, recruited in 47 centers from 14 of 17 Autonomous Regions in Spain, from 2004 onwards. Data are organized and standardized following the HIV Cohorts Data Exchange Protocol (HICDEP) for data collection (details at <https://hicdep.org/>) and adhere to

internal strict annual quality controls. The CoRIS database collects baseline and follow-up socio-demographic, immunological and clinical data. Patients are followed periodically according to routine clinical practice. The CoRIS cohort has been described in detail elsewhere¹⁷.

Each CoRIS participant provided his or her written informed consent prior to enrolling in this study. The CoRIS cohort was approved by the Research Ethic Committee of the Gregorio Marañón Hospital. All methods were carried out in accordance with relevant guidelines and regulations.

The HIV/AIDS Research Network Cohort was established in 2004 in conjunction with the HIV Biobank. These platforms are fundamental resources for improving knowledge about HIV. The coordination team is a multidisciplinary group: epidemiologists, statisticians, clinicians, pharmacists, microbiologist among others. The aim of CoRIS is to collect information on HIV-positive patients in order to study the epidemiological characteristics, the progression of the infection and its determinants, as well as the response to treatment and its influencing factors. CoRIS is organized around three structures:

- Coordination center: it organizes and coordinates data collection, supporting hospitals in these procedures and carries out data processing for statistical analysis.
- Clinical centers: hospitals and health centers where information and biological samples are collected.
- CoRIS Scientific Committee: it decides on the scientific development of the cohort. It also evaluates requests from research groups that need to use CoRIS data in their projects.
- Biobank: blood samples are collected (at baseline visit and annually thereafter) from all patients who give specific informed consent, from which aliquots of serum and cells are separated and stored in a centralized biobank.

For all patients enrolled in the cohort, updates of clinical data and biological parameters were requested at a periodicity of 6 ± 2 months. Follow-up was terminated upon death, change of follow-up center to one outside the cohort, or failure of the patient to appear for scheduled visits.

A database including all baseline and follow-up variables was created with the patient data and made available to all centers. After receiving the data, the coordinating center transfers them to a series of files with a common structure for all the centers. In order to connect all the information referring to the same patient and avoid duplication, a unique code combining the patient's initials, date of birth and sex is used. The files are updated with the data sent by the hospitals every 12 months, both for new patients and for the follow-up of patients already included.

To guarantee the validity of the information and homogeneity between centers, several quality controls are carried out. All the information received at the data coordination center is subjected to a program that automatically detects inconsistencies, out-of-range data and duplicates. The cohort protocol was approved by the ethics committee of each participating hospital. When a patient is recruited, informed consent is requested. All information sent from the hospitals is anonymous.

In addition to clinical and epidemiological information, blood samples are collected from all patients who give specific informed consent, from which samples of serum and cells are separated and stored in a centralized BioBank for the entire cohort.

Manual evaluation of association rules. The development of ML systems requires training data that allows the system to find the appropriate model and configuration to make predictions about new data. In our case, the objective is to have a ML system capable of discriminating whether an AR is relevant or not. For sake of simplicity, sometimes we will call them true (relevant) or false (otherwise). We therefore need a collection of ARs classified as true or false, which will allow us to perform the training. The first problem we faced was to establish the criteria for deciding whether an AR is relevant or not. After making a thorough study of the collection of initial rules extracted by the FP-Growth algorithm, we established the following set of criteria:

An AR $R (X \Rightarrow Y, X = (x_1, \dots, x_n))$ is relevant (True) if

$$\exists i, i \in \{1, n\}, \text{ such that } \text{relation}(x_i, Y) \text{ is not trivial.}$$

that is, if there is any non-trivial relationship between any of the antecedent and consequent diseases.

During the annotation process the following considerations have been taken into account:

- The association with the consequent can be either directly causal or inversely causal ($x_i \rightarrow Y$ and $Y \rightarrow x_i$ are considered equivalent):
For example:
acute myocardial infarction \rightarrow high blood pressure: true
but also:
high blood pressure \rightarrow acute myocardial infarction: true
We assume that often the cause-effect relationship is not clearly established, so we simply consider the relationship between the two elements, without establishing which is the cause and which is the effect.
- If there are several antecedents and some of them are related to the consequent, even if they are not related to each other, the rule is considered true, that is
for the AR $R: X \Rightarrow Y, X = (x_1, \dots, x_n)$, $\text{relation}(x_i, x_j)$ are not taken into account.
For example:
Arterial hypertension, Diabetes mellitus Dislipidemia \rightarrow Acute myocardial infarction

Data	Description
Number of ARs	1000
ARs of interest	613
ARs without interest	387
Shortest AR	1 antecedent and 1 consequent
Longest AR	4 antecedents and 1 consequent
N. of health disorder involved	141
Most freq. health disorders in ARs	Non-defining neoplasm AIDS (387 times)

Table 1. Some data on the Association Rules HIV/AIDS Dataset (ARAIDS), the manually validated collection of ARs.

- If the association with the consequent is that it includes the antecedent or is a part of the antecedent (in a non-trivial way), the AR is true, that is
if for the AR $R: X \Rightarrow Y$, $X = (x_1, \dots, x_n)$ and $\exists i$, such that $i \in \{1, n\}$, such that x_i is a disease that generalizes Y or vice versa, then the AR R is considered relevant:
lung neoplasm \rightarrow *Non-AIDS neoplasm*

In the first phase of this study, we considered that, if any of the elements included in the antecedent did not have an established relationship with the consequent, the rule was false. However, this led to eliminate many interesting rules in which an item could be included by chance, but which established a relevant relationship among the rest of the items. Therefore, we chose to classify as true those rules in which there were potentially relevant relations, even if they included some irrelevant items.

Once the general annotation guidelines were established, a HIV expert clinician (Dr. Otilia Bisbal) examined the rules, by gathering information to establish relationships between elements of the antecedent and the consequent based on their own experience in many cases, and also based on scientific literature in cases of less common relationships. Table 1 shows some data about the manually annotated ARs collection, Association Rules HIV/AIDS Dataset (ARAIDS). It is composed of 1000 rules of which, according to the adopted annotation criteria, 613 are relevant. The longest AR includes 4 antecedents and one consequent. The ARs in the collection involve 141 different health disorders, of which the one that occurs most frequently is *Non-defining neoplasm AIDS*, appearing 387 times.

Semisupervised method for filtering HIV/AIDS associated rules. This section presents the EXTRA algorithm¹⁶, a semi-supervised algorithm which requires an extremely small amount of data to be trained. This system is made up of two modules or components: one that implements an unsupervised method and another that implements a supervised one.

The set of ARs are extracted from CoRIS data by means of the FP-Growth algorithm. This algorithm generates all the possible ARs with certain constraints related to the support and confidence conditions and to the form of the rules.

The support of an AR $X \Rightarrow Y$ is the fraction of transactions that include the set of items in the antecedent or consequent of the rule:

$$\text{support}(X \Rightarrow Y) = \text{support}(X \cup Y) = \frac{\text{count}(X \cup Y)}{N}$$

where N is the number of transactions in the database, and $\text{count}(X \cup Y)$ the number of transactions containing all items in X (antecedent) or Y (consequent).

On the other hand, the confidence of a rule is defined as the fraction of transactions in which itemsets X and Y appear:

$$\text{conf}(X \Rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}$$

It can be interpreted as how often a transaction that contains the itemset X also contains itemset Y .

The parameters used for this algorithm are: Min. Support of 0.0001; Min. confidence of 0.6; Max. antecedent length unlimited; Max. consequent length of 1.

Unsupervised component of EXTRA algorithm. This component applies Fisher's exact test to obtain the p-value corresponding to the ARs. Specifically, the p-value is used to rank a set of ARs. We rank them in ascending order and establish a threshold. Then, the n rules above that threshold (lower value) are considered as true, and the rules below that threshold (higher value) as false.

In order to compute the p-value, the data set is split into two halves called exploratory (50%) and holdout (50%). Then, the FP-Growth algorithm is applied to extract the ARs in both sets.

These two sets of rules allow us to apply the Fisher test to obtain the p-values for the rules in the holdout set. This is done by building, for each rule $R: A \rightarrow B$ in the holdout set, a contingency table with the following data for the rule collected in the exploratory dataset:

	Rules with B (n_2)	Rules without B ($N - n_2$)
Rules with A (n_1)	Rules with A and B (k)	Rules with A (without B) ($n_1 - k$)
Rules without A ($N - n_1$)	Rules with B (without A) ($n_2 - k$)	Rules without A and B ($N - n_1 - n_2 + k$)

The p-value is computed as the hypergeometric distribution of the numbers contained in the cells of the table:

$$p(R) = \frac{\binom{n_1}{k} \binom{N - n_1}{n_2 - k}}{\binom{N}{n_2}} \quad (1)$$

where N is the number of rules in the exploratory set, K is the number of rules in this set containing A and B , n_1 is the number of rules containing A , n_2 is the number of rules containing B .

The value of the threshold is key to the performance of this unsupervised algorithm. Other studies using Fisher's test for filtering the relevant ARs¹⁴ set the threshold following some heuristic such as taking the value which provides a certain number of relevant rules. However, in our case, it can be set with high accuracy using the data from the training set. The threshold is chosen so that the number of hits in the set of ARs of the training set is maximized. In this way, the unsupervised component becomes supervised to a certain degree. This is a great improvement for this method, which is later used to improve the results of the supervised component, resulting in a semi-supervised system that improves its two components.

Supervised component of EXTRA algorithm. Thanks to the availability of manually annotated ARs, we can make use of classical ML systems to build a classifier that indicates whether a new rule is relevant. Specifically, we applied a Random Forest (default parameters provided by WEKA v3.8.2) with the following two groups of features:

- The first set of features have been extracted from the content of the association rules and the output of the FP-Growth algorithm:
 - Support.
 - Confidence.
 - Lift. The lift value is the quotient of the posterior and prior confidence of an association rule. That is, if “ $\emptyset \rightarrow \text{flu}$ ” has a confidence of 60% and “ $\text{cough} \rightarrow \text{flu}$ ” has a confidence of 72%, then the lift value (of the second rule) is $72/60 = 1.2$.
 - Number of antecedents. The number of antecedents of an AR “ A and $B \rightarrow C$ ” is the number of elements of the set $S = \{A, B\}$.
 - Number of consequents. The number of consequents of an AR “ A and $B \rightarrow C$ ” is the number of elements of the set $S = \{C\}$.
- The second group of features attempts to capture medical information on diseases:
 - CDC. The Centers for Disease Control and Prevention (CDC) is the national public health agency of the United States and provides a list of diseases and conditions. This feature provides the normalised value of the association rule items that belong to this list.
 - DIS. In this feature, the items of the association rule that are diseases but do not appear in the CDC list are detected and the normalised value is provided.
 - ADD. The CoRIS dataset provides 34 adverse event types. These adverse events are a very serious type of disease like heart attack, lymphoma, cancer, etc.). This rule provides the normalised value of the presence of this type of adverse disease among the items of the association rule.
 - COD. The CoRIS dataset provides 123 types of cause of death. This feature identifies whether any item in the association rule belongs to this list and uniquely identifies it.
 - CIE. For each item of the association rule we have identified its ICD10 code. Due to the large number of possible labels of the complete code (more than 71K) we have only used the first character corresponding to the “chapter”. Therefore we have created 24 features in which each rule indicates the number of items belonging to that ICD10 chapter.

Semi-supervised EXTRA algorithm. The semi-supervised combination of the components described above results in the semi-supervised EXTRA system capable of classifying relevant ARs with high accuracy while requiring a minimal amount of training data. This system starts from a small seed dataset S of annotated rules. According to previous studies¹⁶, this initial seed can be set to around 10 rules. This seed set is used to train the supervised component leading to a ML model. This ML model is then used to predict the class (i.e. relevant or not) for each AR in the rest of the training set. The last computed seed dataset is also used to compute an accurate p-value threshold for the unsupervised component. After sorting S according to their p-value, we choose

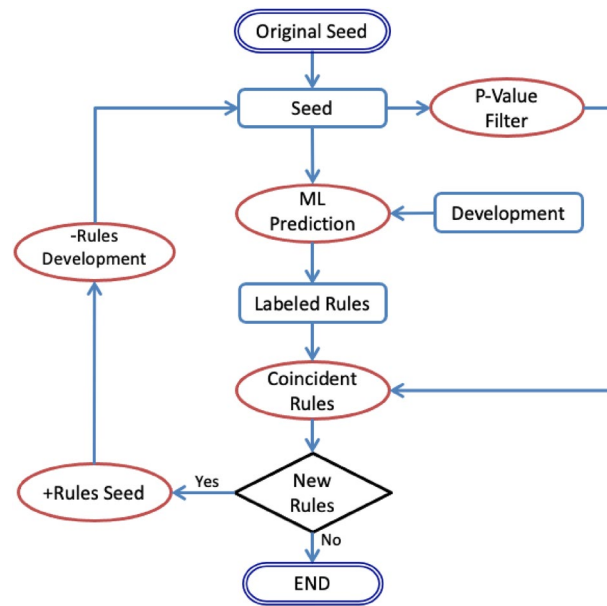


Figure 2. Flow diagram of incremental learning in EXTRAE semi-supervised algorithm for filtering relevant ARs. Rounded rectangles show the beginning and the end of the iterations, rectangles are the rule sets, ovals are processes, and the diamond represents a condition.

as threshold the p -value that maximizes the hits for the seed set. Next, the unsupervised component is applied to the development set to filter the relevant ARs. Afterwards, the results of both components, supervised and unsupervised, are applied to the development dataset for enlarging the seed set. Specifically, the ARs for which the predictions of both components match, i.e. both are true or both are false (coincident set of ARs), are added to the seed set and removed from the development set. The new seed set is used to train the supervised module again, as well as to adjust the threshold of the unsupervised component. This process is repeated until the coincident set of ARs is empty (i.e. the seed set cannot grow anymore). Figure 2 shows a scheme of the system.

Results

In this section we present the experimental results obtained by the system on the ARAIDS dataset along with a parameter analysis.

For the evaluation of our system we have used a set of standard evaluation measures which focus on different aspects of the results. *F-measure* is a combination of the system precision and recall. *AUC ROC* estimates the area under the *ROC* curve for machine learning model comparison, where the *ROC* curve is a graph representing the performance of the system as its discrimination threshold for binary classification varies. We also use the *PR Curve (PRC)*, the result of drawing the graph between the precision and the recall. This graph shows from which recall we have a degradation of the precision and vice versa. The area under this curve (*AU-PRC*) provides a value to compare different systems.

We have carried out a fivefold cross-validation for evaluation. Since EXTRAE is a semi-supervised system, only a portion of the training rules are used for training (seed subset), depending on the different configurations analyzed.

Table 2 shows the results of the semi-supervised method based on Incremental Learning (EXTRAE Algorithm). In order to evaluate the results, the three evaluation measures that state-of-the-art systems typically employ (*F-measure*, *AUR-ROC*, and *AU-PRC*) are provided. As for other values shown in the table, seed size is the original size of the training set from which the set is automatically increased. Iterations show the number of times that new rules need to be added to the seed set, in order that a set is reached to which no new rule can be added. The p -value is calculated from the seed set. Results show the performance of the system after n iterations. From the results shown in this Table, the best seed size is 35. A p -value threshold of $1.12\text{E}-13$ is calculated on this seed size and after 8 iterations an *f-measure* of 0.84 is obtained.

Table 3 shows the partial results of the EXTRAE Algorithm in each iteration for the best configuration shown in Table 2. These results correspond to one of the 5 partitions used as part of the fivefold cross validation method. In the first iteration, 478 new rules are added and an *F-measure* of 0.72 is obtained. From the fifth iteration, the number of matching rules is greatly reduced and, in this way, the performance increases slowly until it reaches an *F-measure* of 0.84. In only eight iterations it increases its performance by 75%, which proves the high quality of the added rules. If we look at the *AUC-ROC* and *AU-PRC*, the final results are even higher than the *F-measure*. This shows that the system performance is robust when using different evaluation criteria.

Table 4 shows the results of the EXTRAE system compared to the two supervised and unsupervised systems of which EXTRAE is composed. For the supervised system, results are provided when using the same number

ARAIDS dataset					
Seed size	Iterations	p-value	F-measure	AUC-ROC	AU-PRC
10	5	1.03E-12	0.62	0.69	0.70
15	4	1.03E-12	0.68	0.71	0.71
20	7	1.03E-12	0.72	0.77	0.77
25	5	1.12E-13	0.79	0.84	0.83
35	8	1.12E-13	0.84	0.88	0.88
50	9	1.12E-13	0.81	0.85	0.85
75	4	1.25E-13	0.80	0.84	0.83
100	6	1.25E-13	0.75	0.79	0.80
125	7	1.25E-13	0.76	0.81	0.81
150	5	1.25E-13	0.71	0.74	0.74

Table 2. Results of EXTRAE algorithm on ARAIDS dataset using different seed sizes, based on their F-measure, AUC-ROC, and AU-PRC. Iterations is the max. number of iterations reached and *p-value* is obtained automatically using the filter approach on the seed set. Best results appear in boldface.

Iteration	Coincident rules	F-measure	AUC-ROC	AU-PRC
0	–	0.48	0.53	0.54
1	478	0.72	0.79	0.78
2	156	0.79	0.83	0.84
3	41	0.81	0.84	0.84
4	18	0.81	0.85	0.85
5	9	0.82	0.85	0.85
6	6	0.82	0.86	0.86
7	5	0.83	0.87	0.88
8	2	0.84	0.88	0.88

Table 3. Evolution of learning from a seed set with 35 rules from one of the 5 partitions used as part of the fivefold cross validation, based on their F-Measure, AUC-ROC, and AU-PRC. Coincident rules are those from the development set that have the same prediction and label based on the *p-value* filter. Best results appear in boldface.

ARAIDS dataset						
System	Seed/training size	Iterations	p-value	F-measure	AUC-ROC	AU-PRC
EXTRAE	35 rules	8	1.12E-13	0.84	0.88	0.88
Supervised	35 rules	–	–	0.52	0.56	0.55
Supervised	800 rules	–	–	0.81	0.86	0.85
Unsupervised	–	–	5.32E-14	0.68	–	–

Table 4. Results of the best performance for the EXTRAE algorithm, and both supervised and unsupervised systems on ARAIDS dataset using the best configuration of EXTRAE, based on their F-measure, AUC-ROC, and AU-PRC. Iterations is the max. number of iterations reached and *p-value* is obtained automatically using the filter approach on the seed set for EXTRAE. Best results appear in boldface.

of seeds (35) as used by EXTRAE, and also for the maximum number of association rules (80%) of the training set. Note that the test set is 20% of the dataset. Finally, the results of the unsupervised system are shown taking into account the best performing *p-value*. The unsupervised system performs significantly worse than EXTRAE, which shows the complexity of the problem. In the case of the supervised system trained with 35 rules, it was clear that the amount of training data was too low for a supervised system, but we wanted to reflect the power of EXTRAE when using the same number of rules.

Finally, EXTRAE obtains better results than the supervised system trained with 800 rules by a slight difference. This comparison is most interesting because, despite initially training with a seed of 35 rules, EXTRAE manages to select the most discriminating rules from the training set to end up with a higher quality rule set than the supervised system and, thus, improve its performance.

Discussion

In this study, we have proven that it is possible to perform filtering of relevant ARs with high accuracy, using a semi-supervised system capable of operating on an extremely small amount of training data. Specifically, in this case, the optimal results have been reached with only 35 annotated ARs. The results for the data considered here are even better than those obtained by the EXTRA algorithm applied to other data in a previous study¹⁶. In the latter, the semi-supervised algorithm applied to primary care data obtained an F-measure of 0.75, AUC-ROC of 0.80 and AU-PRC of 0.81 respectively.

We have demonstrated that it is advantageous to perform association rule filtering with a semi-supervised system, and that the results of such a system are able to outperform both unsupervised and supervised systems when using a reduced amount of training data.

Association rules are a fairly simple artifact in their form, unlike, for example, texts. Therefore, a small number of parameters is enough to characterize their form, as well as aspects related to the frequency of the diseases involved and their combinations (captured by the features of support and confidence).

The medical aspects of the system are to a great extent captured by the unsupervised part. In this part, the system includes the most statistically significant ARs according to the diseases that make up each rule. The combination of both parts and their joint evolution leads the system to select the most relevant ARs.

The prediction of relevant ARs provided by the proposed model and its validation by a HIV medical expert has provided an interesting collection of HIV-related disease relationships. We describe below a number of relationships that have appeared during the validation process.

The relationship of depression with cardiovascular disease, diabetes and cancer is described in the literature and is probably due to immune factors, toxic habits, drugs, etc. and has, therefore, been considered to be true. Similarly, the relationship between psychosis with cardiovascular risk and diabetes has also been described for the same reasons and has also been considered to be true^{18–21}.

The relationship between acute myocardial infarction (AMI) and dementia (especially when it is of vascular origin) has also been described in the literature, so it has been considered true^{22,23}.

Likewise, the relationship between diabetes and dementia is also reported in the literature, so it has been considered true²⁴.

A doubtful case is the relationship between fracture and neoplasm: in extended neoplasms when bone metastases occur there may be pathological fractures, but this is not frequent, and it is not the type of fracture referred to in the CoRIS database, so it has been considered false.

The relationship between Kaposi's sarcoma and bronchial neoplasm has been considered true since Kaposi's sarcoma is a neoplasm that can affect the lung.

The relationship between cachectic syndrome (cachectic sd) and neoplasms has been considered false, because even though the latter can produce a cachectic sd, in the CoRIS database it is specified that it refers to the former due to HIV or wasting syndrome in particular.

Regarding the term “secondary malignant neoplasm of other specified sites”, if it appears in an AR whose consequent is lung or bladder cancer or head and neck cancer, the AR has been considered true because it can be a metastasis of these neoplasms. However, it has been considered false if the consequent of AR is recurrent bacterial pneumonia, because although the latter increases the risk of lung neoplasm, it does not increase the risk of any neoplasm.

The relationship between lactic acidosis and diabetes and cancer has been described²⁵ so it has been considered true.

Appendix A includes a series of relationships considered to be true since they have been confirmed in the literature, together with some references to them. The appendix also gathers relationships that have been considered false either because they lack consistent support in the literature or because they have been ruled out by the literature.

Data availability

The datasets supporting the conclusions of this article are included within the article and its tables, as well as in the supplementary material. The original HIV data from the CoRIS Cohort can be requested to this organization under the corresponding agreement.

Received: 28 June 2022; Accepted: 18 October 2022

Published online: 28 October 2022

References

1. Agrawal, R., Imielinski, T. & Swami, A. N. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC, USA, May 26–28, 1993.*, 207–216 (1993).
2. Imamura, T. *et al.* A technique for identifying three diagnostic findings using association analysis. *Med. Biol. Eng. Comput.* **45**, 51–59 (2007).
3. Park, S. H., Jang, S. Y., Kim, H. & Lee, S. W. An association rule mining-based framework for understanding lifestyle risk behaviors. *PLoS ONE* **9**, e88859 (2014).
4. Rao, P. S. & Devi, T. U. Applicability of apriori based association rules on medical data. *Int. J. Appl. Eng. Res.* **12**, 9451–9458 (2017).
5. Manolitsis, I. *et al.* Using association rules in antimicrobial resistance in stone disease patients. *Stud. Heal. Technol. Inform.* 462–465 (2022).
6. Ou, J. & Zhang, J. Data mining and meta-analysis of psoriasis based on association rules. *J. Healthc. Eng.* **2022** (2022).
7. Babu, S. A., Raj, R. J. S., Varalatchoumy, M., Gopila, M. & Justin, B. V. F. Novel approach for predicting covid-19 symptoms using arm based apriori algorithm. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, 1577–1580 (IEEE, 2022).

8. Nasiri, M. *et al.* Risk factors affecting death from hospital-acquired infections in trauma patients: Association rule mining. *J. Health Manag. Inform.* **8**, 27–33 (2021).
9. Han, J., Pei, J. & Yin, Y. Mining frequent patterns without candidate generation. *ACM Sigmod Rec.* **29**, 1–12 (2000).
10. Prajapati, D. J., Garg, S. & Chauhan, N. Interesting association rule mining with consistent and inconsistent rule detection from big sales data in distributed environment. *Futur. Comput. Inform.* **2**, 3 (2017).
11. García, E., Romero, C., Ventura, S. & Calders, T. Drawbacks and solutions of applying association rule mining in learning management systems. In *Proceedings of the International Workshop on Applying Data Mining in e-Learning (ADML 2007)*, Crete, Greece, 13–22 (sn, 2007).
12. Dahbi, A., Jabri, S., Ballouki, Y. & Gadi, T. A new method to select the interesting association rules with multiple criteria. *Int. J. Intell. Eng. Syst.* **10**, 191–200 (2017).
13. Liu, G., Zhang, H. & Wong, L. Controlling false positives in association rule mining. *Proc. VLDB Endow.* **5**, 145–156 (2011).
14. Webb, G. I. Discovering significant patterns. *Mach. Learn.* **71**, 131 (2008).
15. Saravanan, R. & Sujatha, P. A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 945–949 (IEEE, 2018).
16. Sánchez-de-Madariaga, R., Martínez-Romo, J., Escribano, J. M. C. & Araujo, L. Semi-supervised incremental learning with few examples for discovering medical association rules. *BMC Med. Inform. Decis. Mak.* **22**, 20 (2022).
17. Sobrino-Vegas, P. *et al.* La cohorte de la red española de investigación en sida y su biobanco: Organización, principales resultados y pérdidas al seguimiento. *Enfermedades Infecciosas y Microbiol. Clínica* **29**, 645–653 (2011).
18. De Hert, M. *et al.* Cardiovascular disease and diabetes in people with severe mental illness position statement from the European Psychiatric Association (EPA), supported by the European Association for the Study of Diabetes (EASD) and the European Society of Cardiology (ESC). *Eur. Psychiatry* **24**, 412–424 (2009).
19. Holt, R. I. Association between antipsychotic medication use and diabetes. *Curr. Diabetes Rep.* **19**, 1–10 (2019).
20. Howell, S., Yarovova, E., Khwanda, A. & Rosen, S. D. Cardiovascular effects of psychotic illnesses and antipsychotic therapy. *Heart* **105**, 1852–1859 (2019).
21. Sánchez, M. C., Escurriola, M. F., Baquero, D. B., Arno, A. G. & Callol, J. A. V. Psicosis, riesgo cardiovascular y mortalidad asociada: vamos por el buen camino?. *Clínica e Investig. en Arter.* **26**, 23–32 (2014).
22. Deckers, K. *et al.* Coronary heart disease and risk for cognitive impairment or dementia: Systematic review and meta-analysis. *PLoS ONE* **12**, e0184244 (2017).
23. Sundbøll, J. Depression, stroke, and dementia in patients with myocardial infarction. *Dan. Med. J.* **65**, B5423 (2018).
24. Xue, M. *et al.* Diabetes mellitus and risks of cognitive impairment and dementia: A systematic review and meta-analysis of 144 prospective studies. *Ageing Res. Rev.* **55**, 100944 (2019).
25. Kraut, J. A. & Madias, N. E. Lactic acidosis. *New Engl. J. Med.* **371**, 2309–2319 (2014).

Acknowledgements

This study has been partially supported by the Spanish Ministry of Science and Innovation within the DOTT-HEALTH Project (MCI/AEI/FEDER, UE) under Grant PID2019-106942RB-C32, the OBSER-MENH Project (MCIN/AEI/10.13039/501100011033 and UE (“NextGenerationEU”/PRTR)) under Grant TED2021-130398B-C21 and the project RAICES (IMIENS 2022), PI18CIII/00004 “Infobanco para uso secundario de datos basado en estándares de tecnología y conocimiento: implementación y evaluación de un infobanco de salud para CoRIS (Info-bank for the secondary use of data based on technology and knowledge standards: implementation and evaluation of a health info-bank for CoRIS) - SmartPITeS” and PI18CIII/00019 - PI18/00890 - PI18/00981 “Arquitectura normalizada de datos clínicos para la generación de infobancos y su uso secundario en investigación: solución tecnológica (Clinical data normalized architecture for the generation of info-banks and their secondary use in research: technological solution) - CAMAMA 4” from Fondo de Investigación Sanitaria (FIS) Plan Nacional de I+D+i.

The RIS cohort (CoRIS) is supported by the Instituto de Salud Carlos III through the Red Temática de Investigación Cooperativa en Sida (RD06/006, RD12/0017/0018 and RD16/0002/0006) as part of the Plan Nacional R+D+I and co-financed by ISCIII-Subdirección General de Evaluación and el Fondo Europeo de Desarrollo Regional (FEDER). The list of members of the Cohort of the Spanish HIV Research Network (CoRIS) is included in the Supplementary Material. Additional relationships between Spanish HIV-related diseases confirmed or discarded are included as Supplementary Material.

This study would not have been possible without the collaboration of all patients, medical and nursing staff and data managers who have taken part in the Project.

Author contributions

L.A., J.M. and R.S. designed the semi-supervised model, J.M. performed the experiments, L.A. and J.M. wrote a draft of the article, O.B. evaluated the relationships between diseases encountered. All authors analyzed the results and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-22695-y>.

Correspondence and requests for materials should be addressed to L.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

The Cohort of the National AIDS Network (CoRIS)

Joaquín Portilla⁵, Irene Portilla⁵, Esperanza Merino⁵, Gema García⁵, Iván Agea⁵, José Sánchez-Payá⁵, Juan Carlos Rodríguez⁵, Livia Giner⁵, Sergio Reus⁵, Vicente Boix⁵, Diego Torrus⁵, Verónica Pérez⁵, Julia Portilla⁵, Juan Luís Gómez⁶, Jehovana Hernández⁶, Ana López Lirola⁶, Dácil García⁶, Felicitas Díaz-Flores⁶, M. Mar Alonso⁶, Ricardo Pelazas⁶, M. Remedios Alemán⁶, Víctor Asensi⁷, María Eugenia Rivas Carmenado⁷, Tomás Suarez-Zarracina⁷, Federico Pulido⁸, Rafael Rubio⁸, Otilia Bisbal⁸, M. Asunción Hernando⁸, David Rial⁸, María de Lagarde⁸, Octavio Arce⁸, Adriana Pinto⁸, Laura Bermejo⁸, Mireia Santacreu⁸, Roser Navarro⁸, Candela Gonzalez⁸, Jose Antonio Iribarren⁹, M. José Aramburu⁹, Xabier Camino⁹, Miguel Ángel von Wichmann⁹, Miguel Ángel Goenaga⁹, M. Jesús Bustinduy⁹, Harkaitz Azkune⁹, Maialen Ibarguren⁹, Xabier Kortajarena⁹, Ignacio Álvarez-Rodríguez⁹, Leire Gil⁹, Lourdes Martínez⁹, Félix Gutiérrez¹⁰, Catalina Robledano¹⁰, Mar Masiá¹⁰, Sergio Padilla¹⁰, Araceli Adsuar¹⁰, Rafael Pascual¹⁰, Marta Fernández¹⁰, Antonio Galiana¹⁰, José Alberto García¹⁰, Xavier Barber¹⁰, Vanessa Agullo¹⁰, Javier Garcia Abellán¹⁰, Reyes Pascual¹⁰, Guillermo Telenti¹⁰, Lucia Guillén¹⁰, Ángela Botella¹⁰, Roberto Muga¹¹, Arantza Sanvisens¹¹, Daniel Fuster¹¹, Juan Berenguer¹², Isabel Gutierrez¹², Juan Carlos López¹², Margarita Ramírez¹², Belén Padilla¹², Paloma Gijón¹², Teresa Aldamiz-Echevarría¹², Francisco Tejerina¹², Cristina Diez¹², Leire Pérez¹², Chiara Fanciulli¹², Saray Corral¹², Francesc Vidal¹³, Anna Martí¹³, Joaquín Peraire¹³, Consuelo Viladés¹³, Montserrat Vargas¹³, Montserrat Olona¹³, Anna Rull¹³, Verónica Alba¹³, Elena Yeregui¹³, Jenifer Masip¹³, Graciano García-Pardo¹³, Frederic Gómez Bertomeu¹³, Sonia Espineira¹³, Marta Montero¹⁴, Sandra Cuéllar¹⁴, Marino Blanes¹⁴, María Tacias¹⁴, Eva Calabuig¹⁴, Miguel Salavert¹⁴, Juan Fernández¹⁴, Inmaculada Segarra¹⁴, Juan González-García¹⁵, Ana Delgado-Hierro¹⁵, José Ramón Arribas¹⁵, Víctor Arribas¹⁵, Jose Ignacio Bernardino¹⁵, Carmen Busca¹⁵, Joanna Cano¹⁵, Julen Cardifanos¹⁵, Juan Miguel Castro¹⁵, Luis Escosa¹⁵, Iker Falces¹⁵, Pedro Herranz¹⁵, Victor Hontañón¹⁵, Milagros García¹⁵, Alicia González-Baeza¹⁵, Ma Luz Martín-Carbonero¹⁵, Mario Mayoral¹⁵, Ma Jose Mellado¹⁵, Rafael Micán¹⁵, Rosa de Miguel¹⁵, Rocío Montejano¹⁵, Ma Luisa Montes¹⁵, Victoria Moreno¹⁵, Luis Ramos¹⁵, Berta Rodés¹⁵, Talía Sainz¹⁵, Elena Sendagorta¹⁵, Eulalia Valencia¹⁵, Jose Ramón Blanco¹⁶, Laura Pérez-Martínez¹⁶, José Antonio Oteo¹⁶, Valvanera Ibarra¹⁶, Luis Metola¹⁶, Mercedes Sanz¹⁶, Piedad Arazo¹⁷, Gloria Sampéris¹⁷, David Dalmau¹⁸, Marina Martínez¹⁸, Angels Jaén¹⁸, Montse Sanmartí¹⁸, Mireia Cairó¹⁸, Javier Martínez-Lacasa¹⁸, Pablo Velli¹⁸, Roser Font¹⁸, Mariona Xercavins¹⁸, Noemí Alonso¹⁸, Francesco Aiello¹⁸, María Rivero¹⁹, Beatriz Piérola¹⁹, Maider Goikoetxea¹⁹, María Gracia¹⁹, Carlos Ibero¹⁹, Estela Moreno¹⁹, Jesús Repáraz¹⁹, Gemma Navarro²⁰, Manel Cervantes Garcia²⁰, Sonia Calzado Isbert²⁰, Marta Navarro Vilasaro²⁰, Belen Lopez Garcia²⁰, Ignacio de los Santos²¹, Alejandro de los Santos²¹, Jesús Sanz²¹, Lucio García-Fraile²¹, Enrique Martín²¹, Ildefonso Sánchez-Cerrillo²¹, Marta Calvet²¹, Ana Barrios²¹, Azucena Bautista²¹, Carmen Sáez²¹, Marianela Ciudad²¹, Ángela Gutiérrez²¹, Santiago Moreno²², Santos del Campo²², José Luis Casado²², Fernando Drona²², Ana Moreno²², M. Jesús Pérez²², Sergio Serrano²², Ma Jesús Vivancos²², Javier Martínez-Sanz²², Alejandro Vallejo²², Matilde Sanchez²², Jose Antonio Pérez-Molina²², José Manuel Hermida²², Enrique Bernal²³, Antonia Alcaraz²³, Joaquín Bravo²³, Ángeles Muñoz²³, Cristina Tomás²³, Mónica Martínez²³, M. Carmen Villalba²³, Federico García²⁴, Clara Martínez²⁴, José Hernández²⁴, Leopoldo Muñoz Medina²⁴, Marta Álvarez²⁴, Natalia Chueca²⁴, David Vinuesa²⁴, Adolfo de Salazar²⁴, Ana Fuentes²⁴, Emilio Guirao²⁴, Laura Viñuela²⁴, Andrés Ruiz-Sancho²⁴, Francisco Anguita²⁴, Jorge Del Romero²⁵,

Montserrat Raposo²⁵, Carmen Rodríguez²⁵, Teresa Puerta²⁵, Juan Carlos Carrió²⁵, Mar Vera²⁵, Juan Ballesteros²⁵, Oskar Ayerdi²⁵, Begoña Baza²⁵, Eva Orviz²⁵, Antonio Antela²⁶, Elena Losada²⁶, Melchor Riera²⁷, María Peñaranda²⁷, M. Angels Ribas²⁷, Antoni A. Campins²⁷, Mercedes Garcia-Gazalla²⁷, Francisco J. Fanjul²⁷, Javier Murillas²⁷, Francisco Homar²⁷, Helem H. Vilchez²⁷, Luisa Martín²⁷, Antoni Payeras²⁷, Jesús Santos²⁸, María López²⁸, Crisitina Gómez²⁸, Isabel Viciano²⁸, Rosario Palacios²⁸, Luis Fernando López-Cortés²⁹, Nuria Espinosa²⁹, Cristina Roca²⁹, Silvia Llaves²⁹, Juan Manuel Tiraboschi³⁰, Arkaitz Imaz³⁰, Ana Karina Silva³⁰, María Saumoy³⁰, Sofía Catalina Scévola³⁰, Adrián Curran³¹, Vicenç Falcó³¹, Jordi Navarro³¹, Joaquin Burgos³¹, Paula Suanzes³¹, Jorge García³¹, Vicente Descalzo³¹, Patricia Álvarez³¹, Bibiana Planas³¹, Marta Sanchiz³¹, Lucía Rodríguez³¹, Julián Olalla³², M. José Sánchez³², Javier Pérez³², Alfonso del Arco³², Javier de la Torre³², José Luis Prada³², Onofre Juan Martínez³³, Lorena Martínez³³, Francisco Jesús Vera³³, Josefina García³³, Begoña Alcaraz³³, Antonio Jesús Sánchez Guirao³³, Alvaro Mena³⁴, Angeles Castro³⁴, Berta Pernas³⁴, Pilar Vázquez³⁴, Soledad López³⁴, Sofía Ibarra³⁵, Guillermo García³⁵, Josu Mirena³⁵, Oscar Luis Ferrero³⁵, Josefina López³⁵, M. Mar Cámara³⁵, Mireia de la Peña³⁵, Miriam Lopez³⁵, Iñigo Lopez³⁵, Itxaso Lombide³⁵, Victor Polo³⁵, Joana de Miguel³⁵, Carlos Galera³⁶, Marian Fernández³⁶, Helena Albendin³⁶, Antonia Castillo³⁶, Asunción Iborra³⁶, Antonio Moreno³⁶, M. Angustias Merlos³⁶, Asunción Vidal³⁶, Concha Amador³⁷, Francisco Pasquau³⁷, Concepcion Gil³⁷, Jose Tomás Algado³⁷, Inés Suarez-García³⁸, Eduardo Malmierca³⁸, Patricia González-Ruano³⁸, M. Pilar Ruiz³⁸, José Francisco Pascual³⁸, Elena Sáez³⁸, Luz Balsalobre³⁸, M. Villa López³⁹, Mohamed Omar³⁹, Carmen Herrero³⁹, M. Amparo Gómez³⁹, Miguel Alberto de Zarraga⁴⁰, Desiré Pérez⁴⁰, Vicente Estrada⁴¹, Nieves Sanz⁴¹, Noemí Cabello⁴¹, Jorge Vergas García⁴¹, María Jose Núñez⁴¹, Iñigo Sagastagoitia⁴¹, Miguel Górgolas⁴², Alfonso Cabello⁴², Beatriz Álvarez⁴², Laura Prieto⁴², Irene Carrillo⁴², José Sanz⁴³, Alberto Arranz⁴³, Cristina Hernández⁴³, María Novella⁴³, M. José Galindo⁴⁴, Ana Ferrer⁴⁴, Antonio Rivero Román⁴⁵, Inma Ruíz⁴⁵, Antonio Rivero Juárez⁴⁵, Pedro López⁴⁵, Isabel Machuca⁴⁵, Mario Frias⁴⁵, Ángela Camacho⁴⁵, Ignacio Pérez⁴⁵, Diana Corona⁴⁵, Ignacio Pérez⁴⁵, Diana Corona⁴⁵, Miguel Cervero⁴⁶, Rafael Torres⁴⁶, Juan Antonio Pineda⁴⁷, Pilar Rincón⁴⁷, Juan Macías⁴⁷, Luis Miguel Real⁴⁷, Anais Corma⁴⁷, Marta Fernández⁴⁷, Alejandro Gonzalez-Serna⁴⁷, Eva Poveda⁴⁸, Alexandre Pérez⁴⁸, Luis Morano⁴⁸, Celia Miralles⁴⁸, Antonio Ocampo⁴⁸, Guillermo Pousada⁴⁸, Lucía Patiño⁴⁸, Carlos Dueñas⁴⁹, Sara Gutiérrez⁴⁹, Elena Tapia⁴⁹, Cristina Novoa⁴⁹, Xjoylin Egües⁴⁹ & Pablo Tellería⁴⁹

⁵Hospital General Universitario de Alicante, Alicante, Spain. ⁶Hospital Universitario de Canarias, San Cristóbal de la Laguna, Spain. ⁷Hospital Universitario Central de Asturias, Oviedo, Spain. ⁸Hospital Universitario 12 de Octubre, Madrid, Spain. ⁹Servicio de Enfermedades Infecciosas, Hospital Universitario Donostia, Instituto de Investigación BioDonostia, Donostia-San Sebastián, Spain. ¹⁰Hospital General Universitario De Elche, Elche, Spain. ¹¹Hospital Universitari Germans Trias i Pujol (Can Ruti), Badalona, Spain. ¹²Hospital General Universitario Gregorio Marañón, Madrid, Spain. ¹³Hospital Universitari de Tarragona Joan XXIII, Tarragona, Spain. ¹⁴Hospital Universitario y Politécnico de La Fe, Valencia, Spain. ¹⁵Hospital Universitario La Paz/IdiPAZ, Madrid, Spain. ¹⁶Hospital San Pedro Centro de Investigación Biomédica de La Rioja (CIBIR), Logroño, Spain. ¹⁷Hospital Universitario Miguel Servet, Zaragoza, Spain. ¹⁸Hospital Universitari Mutua Terrassa, Terrassa, Spain. ¹⁹Complejo Hospitalario de Navarra, Pamplona, Spain. ²⁰Parc Taulí Hospital Universitari, Sabadell, Spain. ²¹Hospital Universitario de La Princesa, Madrid, Spain. ²²Hospital Universitario Ramón y Cajal, Madrid, Spain. ²³Hospital General Universitario Reina Sofía, Murcia, Spain. ²⁴Hospital Nuevo San Cecilio, Granada, Spain. ²⁵Centro Sanitario Sandoval, Madrid, Spain. ²⁶Hospital Clínico Universitario de Santiago, Santiago de Compostela, Spain. ²⁷Hospital Universitario Son Espases, Palma de Mallorca, Spain. ²⁸Hospital Universitario Virgen de la Victoria, Málaga, Spain. ²⁹Hospital Universitario Virgen del Rocío, Seville, Spain. ³⁰Hospital Universitario de Bellvitge, Hospitalet de Llobregat, Spain. ³¹Hospital Universitario Valle de Hebrón, Barcelona, Spain. ³²Hospital Costa del Sol, Marbella, Spain. ³³Hospital General Universitario Santa Lucía, Cartagena, Spain. ³⁴Complejo Hospitalario Universitario a Coruña (Chuac), A Coruña, Spain. ³⁵Hospital Universitario Basurto, Bilbao, Spain. ³⁶Hospital Universitario Virgen de la Arrixaca, El Palmar, Spain. ³⁷Hospital de la Marina Baixa, La Vila Joiosa, Spain. ³⁸Hospital Universitario Infanta Sofía, San Sebastián de los Reyes, Spain. ³⁹Hospital Universitario de Jaén, Jaén, Spain. ⁴⁰Hospital Universitario San Agustín, Avilés, Spain. ⁴¹Hospital Clínico San Carlos, Madrid, Spain. ⁴²Hospital Universitario Fundación Jiménez Díaz, Madrid, Spain. ⁴³Hospital Universitario Príncipe de Asturias, Alcalá de Henares, Spain. ⁴⁴Hospital Clínico Universitario de Valencia, Valencia, Spain. ⁴⁵Hospital Reina Sofía, Córdoba, Spain. ⁴⁶Hospital Universitario Severo Ochoa, Leganés, Spain. ⁴⁷Nuestra Señora de Valme, Seville, Spain. ⁴⁸Hospital Álvaro Cunqueiro, Vigo, Spain. ⁴⁹Hospital Clínico Universitario de Valladolid, Valladolid, Spain.