Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Computers in Biology and Medicine

Medical data quality assessment: On the development of an automated framework for medical data curation

Vasileios C. Pezoulas[a], Konstantina D. Kourou[a,b], Fanis Kalatzis[a], Themis P. Exarchos[a,c], Aliki Venetsanopoulou[d], Evi Zampeli[e], Saviana Gandolfo[f], Fotini Skopouli[g], Salvatore De Vita[f], Athanasios G. Tzioufas[d], Dimitrios I. Fotiadis[a,h,*]

[a] *Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, Ioannina, GR45110, Greece*
[b] *Department of Biological Applications and Technology, University of Ioannina, Ioannina, GR45110, Greece*
[c] *Department of Informatics, Ionian University, Corfu, GR49100, Greece*
[d] *Department of Pathophysiology, School of Medicine, University of Athens, Athens, GR15772, Greece*
[e] *Institute for Systemic Autoimmune and Neurological Diseases, Athens, GR11743, Greece*
[f] *Clinic of Rheumatology, Department of Medical and Biological Sciences, Udine University, Udine, IT33100, Italy*
[g] *Department of Internal Medicine and Clinical Immunology, Euroclinic Hospital, Athens, GR11521, Greece*
[h] *Department of Biomedical Research, FORTH-IMBB, Ioannina, GR45110, Greece*

ARTICLE INFO

Keywords:
Big data
Data quality
Data quality assessment
Data curation
Data standardization

ABSTRACT

Data quality assessment has gained attention in the recent years since more and more companies and medical centers are highlighting the importance of an automated framework to effectively manage the quality of their big data. Data cleaning, also known as data curation, lies in the heart of the data quality assessment and is a key aspect prior to the development of any data analytics services. In this work, we present the objectives, functionalities and methodological advances of an automated framework for data curation from a medical perspective. The steps towards the development of a system for data quality assessment are first described along with multidisciplinary data quality measures. A three-layer architecture which realizes these steps is then presented. Emphasis is given on the detection and tracking of inconsistencies, missing values, outliers, and similarities, as well as, on data standardization to finally enable data harmonization. A case study is conducted in order to demonstrate the applicability and reliability of the proposed framework on two well-established cohorts with clinical data related to the primary Sjögren's Syndrome (pSS). Our results confirm the validity of the proposed framework towards the automated and fast identification of outliers, inconsistencies, and highly-correlated and duplicated terms, as well as, the successful matching of more than 85% of the pSS-related medical terms in both cohorts, yielding more accurate, relevant, and consistent clinical data.

1. Introduction

Data quality has been recognized as a key factor in all operating processes both in the public and private sectors [1]. It has been characterized as a multidisciplinary process since it reflects the needs for sustainable data of high quality in several domains varying from business to healthcare. The technological advances of our era combined with the fact that the structure of the current information systems is network-based, have dramatically increased the amount of digital data [2]. A crucial consequence of this evolution is that data management has become more complex and controversial. This need has increased the necessity for automated methods and rules that are able to deal with the quality assessment of big data. Lack of data quality results in bad data manipulation which makes data useless and has numerous negative effects on further processing. Thus, emphasis must be given on the development of new methods for dealing with insufficient data sources.

The most important process of a data management system is the data quality assessment [1,3]. The data quality assessment process is related to: (i) the evaluation of data protection metrics (e.g., data protection impact assessment), (ii) the organizational structure of the data, and (iii) the overall information management. Several studies [3–9] have been launched highlighting the leading role of data quality assessment in improving the information quality especially in the medical domain. To assess the quality of the data, one must first define

[*] Corresponding author. Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, GR45110, Ioannina, Greece.
E-mail address: fotiadis@cc.uoi.gr (D.I. Fotiadis).

https://doi.org/10.1016/j.compbiomed.2019.03.001
Received 18 December 2018; Received in revised form 1 March 2019; Accepted 2 March 2019

the quality requirements and metrics [1,3–9]. Examples of common data quality requirements include the: (i) accuracy, (ii) completeness, (iii) consistency, (iv) interpretability, (v) timeliness, (vi) relevancy, and (vii) ease of manipulation, among many others [1,3–9]. In general, multiple metrics can be associated with each quality requirement. For example, completeness, in the field of data science, can be defined as the degree to which a given dataset meets the pre-defined requirements of an optimal dataset. In this case, completeness is a quality requirement that can be quantified from data-driven (quality) metrics, such as, the number of missing values, incomplete terminology, etc.

Data cleaning, also referred to as data curation [10–12], is the backbone of data quality assessment which aims to clean and transform the raw data into new ones that meet the quality criteria posed by the domain of interest (e.g., the clinical domain). The migration from manual data cleaning to automated, rule-driven, data cleaning overcomes the complexity of processing huge amounts of data that are obtained in different time points and increases the scalability of such a system which is a crucial requirement. In general, automated data curation involves functionalities for: (i) automatic detection of missing values (bad features), (ii) outlier detection and removal, (iii) similarity detection, i.e., identification of duplicate fields and highly correlated distributions, and (iv) attribute identification and grouping, among many others, and therefore the enhancement of the overall quality assessment process [3]. Outlier detection involves the detection of values that vary from the standard observed distribution. Outliers can be detected using univariate and multivariate methods. In the case of clinical data which include several domains, such as, demographic data, laboratory measures, and therapies, univariate methods are preferred since the domains are independent. Similar fields can be detected by name conflicts or by highly correlated distributions.

It has been long proven that the quality of data mining results and related applications highly depend on the quality of the data [13]. For example, a clinician who wishes to apply a simple regression model on a contaminated clinical dataset (in the presence of outliers and/or incompatible values) in order to identify independent factors for a particular disease or develop a prediction model for the disease progress, will end up with a distorted model with no clinical value. In addition, the structural heterogeneity of the clinical data across different clinical centers introduces biases during the analysis of medical data. The heterogeneity and non-canonical form of medical data resulting from either the bad quality of the medical data or the lack of a standard clinical vocabulary hampers data mining and other processing tasks [13]. Data standardization is a promising solution for transforming the data into a common format [14,15]. Interoperability standards and modeling annotations are key factors towards the success of data standardization [16]. Standardization is usually performed according to a gold standard model which serves as a reference model, i.e., a set of parameters which describes the requirements (e.g., variables, types, descriptions) of a disease of interest [14]. The majority of the studies make use of gold standard models to assess the quality of different types of clinical data based on various data quality measures, such as, the accuracy, the completeness, etc.

Data standardization has many similarities with data harmonization. The latter is of great importance since it aims to overcome the heterogeneity of medical data worldwide by converting the heterogeneous data into homogeneous ones (e.g., with similar structure and terminologies) [15,16]. Harmonization involves several mechanisms including data transformation to a common format, data annotation, terminology detection and alignment, most of which are part of the data curation framework and especially of the data standardization process. It is, therefore, important to consider data standardization as a crucial part of a data quality assessment framework and a key part of data harmonization. There are two types of methods for matching heterogeneous datasets, namely the semantic matching and the lexical matching [15,16]. The latter simply seeks for string matching among the features of a dataset whereas semantic matching also accounts for associative relationships between the classes where the features belong to.

According to the literature, most of the aforementioned studies mainly focus on providing general guidelines for data quality assessment [1,3–5,9], methodological steps towards data curation [10–12] and standardization [6–8], without, however, focusing on the development and evaluation of a computational framework for data quality assessment on medical data. In an effort to address this need, we present the objectives, functionalities and methodological advances of an integrated framework for medical data curation in terms of data quality assessment. The proposed framework consists of a three-layer architecture and serves as a diagnostic tool for managing incomplete terminologies, irrelevant terms, outliers, missing values, data categorization, and duplicated terms. In the core of this framework lies data standardization. In this work, we extend data standardization as a pre-harmonization process to make data harmonization easier and faster. More specifically, we use lexical matching combined with model-based rules and external sources, i.e., vocabularies, to match and classify terms according to a pre-defined reference model which is a set of parameters which describe the requirements (variables with their types and ranges) of the clinical domain of interest. Through this procedure, we attempt to produce semantic relations between the fields of the raw dataset with those from a reference model and therefore enhance the semantic matching process for data harmonization. The proposed framework is applied on anonymized data from patients that have been diagnosed with primary Sjögren's Syndrome (pSS), yielding reliable outcomes. To our knowledge, this is the first case study on automated medical data quality assessment from a data harmonization perspective.

Our results confirm the validity of the proposed framework towards the precise identification of outliers, inconsistencies (unknown data types), and highly-correlated and duplicated terms in both cohorts, as well as, the clinical usefulness and guidance of the data quality assessment report and the curated dataset towards the improvement of the overall accuracy, consistency, and relevance of the examined clinical data. The data standardization process was able to successfully capture more than 85% of the pSS-relevant terms in both datasets using lexical matching techniques combined with rules that use knowledge from a reference model. The framework uses an XML representation of the reference model which increases its overall scalability and thus can be generalized for different types of diseases, introducing the ontologies as a preliminary step for medical data harmonization. The fact that all the computational tasks were executed in a few seconds, demonstrates the dominance of automated data curation against traditional manual curation.

2. Methods

2.1. The proposed architecture

The proposed framework for data curation consists of a three-layer architecture which receives as input a raw dataset and outputs the data evaluation report which provides information related to the data quality and the data standardization outcomes, and the curated dataset (Fig. 1). The architecture is designed to meet the data quality requirements and functionalities that were mentioned in Section 1, and is comprised of three modules: (i) the data evaluation module (Section 2.2), (ii) the data quality control module (Section 2.3), and (iii) the data standardization module which serves as a pre-harmonization step (Section 2.4).

2.1.1. Input

The framework uses as input an *mxn* raw dataset, where *m* is the number of instances (patients) and *n* is the number of attributes (features). In addition, the framework uses external sources for the purposes of the data standardization module. These sources include: (i) the

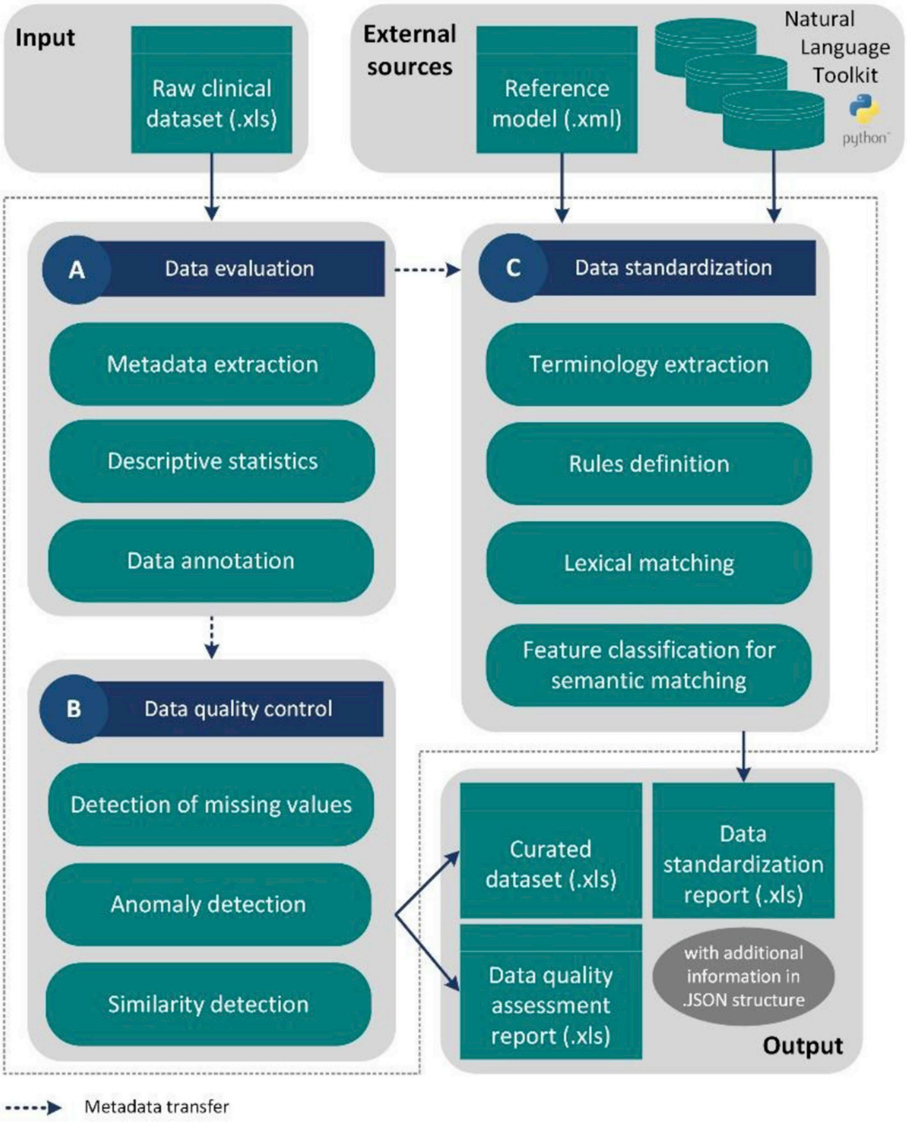


Fig. 1. The proposed three-step hierarchical architecture for data curation.

reference model (in.XML format [14,15]) which is a set of parameters that describes the requirements (e.g., variables, types, descriptions) of a disease of interest, and (ii) the NLTK's WordNet corpus reader [17] which has been attached as part of the lexical matching procedure to compute sets of synonyms.

2.1.2. Modules and related functionalities

The proposed framework has been implemented in Python 3.6 and can be executed using a Representational State Transfer (REST) service through a secure web client and consists of the following modules (Fig. 1):

- The data evaluation module which offers a first look into the structure of the raw dataset and includes functionalities for: (i) metadata extraction (e.g., range of values), (ii) computation of descriptive statistics (e.g., kurtosis), and (iii) data annotation (e.g., data types).
- The data quality control module which is responsible for data processing and includes functionalities for: (i) the detection of missing values, (ii) anomaly detection using both univariate and multivariate methods, and (iii) the detection of features with similar context using similarity metrics and detection of duplicate fields.
- The data standardization module which is responsible for matching the fields of the raw (input) dataset with those from the reference model (which is defined as the optimal dataset that fulfills the necessary requirements for describing the domain knowledge of interest) and includes functionalities for: (i) terminology extraction, (ii) rules definition based on the reference model, (iii) lexical matching to identify fields with similar terminology based on string distance measures (e.g., sequence matching) using the rules, as well as, input from external sources (i.e., the Natural Language Toolkit), and (iv) feature classification for semantic matching, i.e., feature categorization into the classes of the reference model's semantic representation (or ontology), to enable faster semantic matching.

2.1.3. Output

The framework outputs three documents (in. xls format): (i) the data quality assessment report, (ii) the curated dataset, and (iii) the data standardization report. The data quality assessment report summarizes useful information regarding the type of each feature (categorical or numeric), the value range of each feature, the number of missing values per feature, the state of each feature (according to the number of missing values), whether outliers were detected or not per feature, and finally compatibility issues (e.g., unknown data types). The curated dataset is a modified version of the raw dataset where the outliers, the problematic fields, the missing values, and the names of each feature are highlighted using appropriate color cording. The data standardization report summarizes useful information regarding the terms of the raw input dataset that were matched with those from the reference model along with the correct value range and the class where each matching term belongs to. The majority of the returned parameters from the.JSON structure are summarized in the data quality assessment report. The.JSON structure includes additional information regarding the descriptive statistics of a selected feature (e.g., kurtosis, skewness).

2.2. Data evaluation module

The first layer of the data curator aims at providing a brief introduction into the dataset's structure and quality through: (i) the extraction of structural and vocabulary information (e.g., format, attributes, range values), (ii) the computation of ordinary descriptive statistics (e.g., mean, standard deviation, skewness, kurtosis), (iii) the categorization of attributes into discrete and continuous, (iv) the detection of missing values per attribute, and (v) histogram and distribution plots.

2.2.1. Metadata extraction

Metadata extraction is the process of extracting structural and vocabulary information from the dataset under examination. Structural information provides knowledge regarding the labels (names) of the attributes, the number of features and instances, the number of meta-attributes, the number of missing and/or unknown values. On the other hand, vocabulary information provides knowledge regarding the range of the values, the data types (e.g., float, string), etc. Note that using a dataset's structural and vocabulary knowledge one can construct an ontology [18] that describes the semantic relations between the features in terms of classes, sub-classes, and object properties.

2.2.2. Descriptive statistics

A common way to get a first look into an attribute's domain is to investigate its distribution using descriptive measures. Such measures include the mean, median, minimum, maximum, standard deviation, kurtosis, and skewness. Descriptive statistics can provide valuable information that might lead to the identification of statistical differences between the features.

2.2.3. Data annotation

Data annotation refers to the categorization of features into continuous or discrete and categorical or numeric according to their data type and range values. Features can also be classified according to their quality in terms of compatibility issues and missing values, as “good” (no missing values/incompatibilities), “fair” ($<$ 25% of missing values/incompatibilities) or “bad” ($>$ 50% missing values/incompatibilities).

2.3. Data quality control module

The second layer of the data curator aims at fixing the dataset in terms of detecting outliers and terms with similar terminology and context.

2.3.1. Anomaly detection

The goal is to separate a core of regular observations from some polluting ones (i.e., the outliers), which vary from the majority. A widely used univariate method for outlier detection, is the z-score measure which quantifies the distance between a feature's value and its population mean [19,20]. It is defined as:

$$z = \frac{\boldsymbol{x} - \hat{x}}{\sigma_x}, \tag{1}$$

where $\boldsymbol{x}$ is the feature vector, $\hat{x}$ is its mean value, and σ_x is its standard deviation. Features with z-values larger than 3 or smaller than -3 are considered as outliers [19].

In the domain of the statistical tests, the Grubb's test is a widely used univariate statistical measure which tests for the hypothesis that there are outliers in the data [20]. The test statistics is given by:

$$G = \frac{max(|\boldsymbol{x} - \hat{x}|)}{\sigma_x}. \tag{2}$$

In fact, the Grubb's test statistics is defined as the largest absolute deviation from the sample mean in units of the sample standard deviation [20]. Here, we are interested in testing whether the minimum value or the maximum value of x is an outlier, i.e., a two-sided test. A value is considered to be outlier if it fulfills the following condition [20]:

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{(t_{a/(2N),N-2})^2}{N-2+(t_{a/(2N),N-2})^2}}, \tag{3}$$

where $(t_{a/(2N),N-2})^2$ is the critical value of the Student's t-distribution with $(N-2)$ degrees of freedom and a significant level equal to $a/(2N)$. The Interquartile Range (IQR), is another approach which measures the statistical dispersion using the 1st and 3rd quartiles of an attribute's range [19,20]. It is defined as the difference between the upper (Q3) and lower (Q1) quartiles of the data. Values lower than the first quartile or larger than the third quartile are outliers.

The Local Outlier Factor (LOF) [21] is a density-based approach which measures the local density of a given data point with respect to its neighboring points, where the number of nearest neighbors determines the accuracy of the model. In fact, it uses the density of a point against its neighbors to determine the degree of the whether the point is an inlier or an outlier. For a point x, the local outlier factor (LOF) is equal to:

$$LOF(x) = \frac{\sum_{x' \in N_k(x)} (lrd(x')/lrd(x)) \Big)}{N_k(x)} = \sum_{x' \in N_k(x)} lrd(x') \sum_{x' \in N_k(x)} r(x, x'), \tag{4}$$

where, $N_k(x)$ is the set of k-nearest neighbors for x, $r(x, x')$, is the reachability distance, i.e., the “true” distance between two points in the 2D space, and $lrd(x)$ is the local reachability density (lrd) of x. In fact, the LOF is equal to the average local reachability density of the neighbors divided by the point's own lrd. The higher the LOF the more likely the point is an outlier.

2.3.2. Similarity detection and de-duplication

Similarity detection refers to the identification of features that exhibit similar characteristics (i.e., a common distribution). In order to detect such features, we first compute a similarity measure that is not constrained on normally distributed data, such as, the Spearman's rank-order correlation coefficient, for each possible pair of features. Assuming that the examined dataset contains n-features, we first compute the similarity measure between each pair of features resulting to an nxn similarity matrix (adjacency matrix), where the element (i, j) is the Spearman correlation between features i and j. Then we seek for those pairs which are highly-correlated (i.e., $>$ 98% correlated) and statistically significant (i.e., $p < 0.05$) using a permutation test which tests for the hypothesis that the data are correlated. In addition, a common method towards the identification of features with duplicate terms can be performed using various string distance metrics (see Section 2.4.3) for detecting completely or partially matching terms.

2.4. Data standardization module

The third and final layer of the data curator, aims to consider for data standardization in terms of lexical matching. Terminologies are first extracted and fixed, then rules are defined and finally the lexical matching process takes place in order to match similar terms and classify them into classes based on a reference model. The importance of this module lies on the fact that it can be considered as a pre-harmonization step.

2.4.1. Terminology extraction

The feature labels are first extracted and used as input for the rules definition process. Additional constraints are performed for removing incompatibilities (e.g., parentheses, commas) per feature.

2.4.2. The reference model and related rules

The reference model is a way to describe the domain knowledge of a disease of interest [22]. It serves as a common template that contains a set of clinical parameters, such as, demographics, laboratory tests, therapies, etc. The types and ranges of each specified parameter within the template can be determined according to the guidelines from the experts of each domain. A reference model is usually constructed for a specific clinical domain according to guidelines from the clinical domain experts. The reference model can be used to construct an ontology which is helpful for solving the semantic matching problem for data harmonization. Rules definition is mainly based on the range values of each feature. For example, the term “Gender” in the reference model is a binary feature where “0” denotes male and “1” denotes female. Assuming that the corresponding feature on the input dataset has the name “Sex” and takes the value “M” for males and “F” for females, one can develop a simple rule that: (i) seeks for sequence or synonymous matches between the feature names, and (ii) records additional information regarding the correct (standardized) range of the feature (i.e., “M”:“0”, “F”:“1”) or the measurement units. The rules are based on lexical matching algorithms (see Section 2.4.3) in order to seek for terminologies with common block sequences (e.g., in the case of “lymphoma” and “lymphoma score”), high string similarity scores (e.g., in the case of “haemoglobin” and “hemoglobin”), exact string matches or synonymous terminologies (e.g., in the case of “sex” and “gender”). The rules were developed for all the terms of the model.

2.4.3. Lexical matching

For a given dataset, we first extract its structural information (metadata), i.e., terms, range values (Section 2.2). A vocabulary is then created by extracting the terms from the reference model. Our goal is to match the terms of the given dataset which are related with those from the reference model. We use the Jaro distance [23–25] as a string similarity measure to calculate the similarity between two terms. For two strings, x and y, the Jaro string similarity measure, sim_J, is equal to:

$$sim_J = \begin{cases} 0\,, & c = 0 \\ \frac{1}{3}\cdot\left(\frac{c}{|x|} + \frac{c}{|y|} + \frac{c-t}{c}\right), & o/w \end{cases}\,, \tag{5}$$

where c is the number of matching (coincident) characters, and t is half the number of transpositions. We also applied the sequence matching algorithm [26,27] to seek for identical blocks between the terms of the input data with those from the reference model. The algorithm calculates the edit distance between two strings, x and y, which is defined as the minimum number of operations, i.e., deletions, insertions, and replacements, that are needed to transform x into y:

$$D_{x,y}(i,j) = \begin{cases} i\,, & i = 0 \\ j\,, & j = 0 \\ D[i-1, i-1]\,, & i, j > 0 \; and \; x_i = y_j \\ \min\begin{Bmatrix} D[i-1, j-1] + 1 \\ D[i-1, j] + 1 \\ D[i, j-1+1] \end{Bmatrix}, & o/w \end{cases}\,, \tag{6}$$

where $D_{x,y}(i,j)$ is the distance between the first i-characters of x and the first j-characters of y. The minimum edit distance is a recursive problem which can be dynamically solved in $O(MN)$ time complexity, where M is the length of x and N is the length of y. Python's SequenceMatcher was used to find the longest common patterns between the terms of the input dataset with those from the vocabulary using the rules of Section 2.4.2. The result of the lexical matching procedure is a triplet (x_{raw}, x_{ref}, v) where x_{raw} is the term of the raw dataset, x_{ref} is the matching term from the reference model, and v is the similarity score.

2.4.4. Feature classification for semantic matching

Each identified term is finally classified into a class v which is part of the semantic representation of the reference model, i.e., the ontology. For example, the identified terms “Age”, “Gender” can be classified into the class “Demographics”. The final result of the standardization module is a tuple (x_{raw}, x_{ref}, v, c) where x_{raw} is the term of the raw dataset, x_{ref} is the matching term from the reference model, v is the (similarity) matching score, and v is the class where x_{raw} belongs to.

3. Results

3.1. Case study

The developed framework was evaluated on two cohorts with anonymized data from patients that have been diagnosed with primary Sjögren's Syndrome (pSS). The anonymized data from the first cohort include 200 patients from the University of Athens (UoA), whereas the second includes 100 patients from the Harokopio University of Athens (HUA). The cohort data were obtained under the data protection agreement version 3.7 as of August 2018 according to the Article 35 (3) (b) of the GDPR fulfilling all the necessary ethical and legal requirements for data sharing. pSS is a chronic inflammatory autoimmune disease, causing salivary gland dysfunction (dryness in the eyes, mouth, skin, vagina), affecting primarily women near the menopausal age [28,29]. In 40–60% of pSS patients, extraglandular involvement is also exhibited [29], whereas 5% of pSS patients are associated with the development of B-cell non-Hodgkin lymphoma [28].

The following three documents were generated by the REST-service: (i) a data quality assessment report, (ii) a data standardization report, and (iii) the curated dataset. Additional information is provided in JSON format during the development of the reports which is also present in the data quality assessment report. The data quality assessment report includes information related to the cohort's metadata (i.e., data types, range of values, summary statistics), detected outliers, duplicated terms, incompatibilities and, finally, the state of each feature in terms of missing values. The data standardization report includes information related to the classification of each term into the classes of the reference model, as well as, the correct set of ranges and measurement units for each term of the raw input dataset according to the reference model. The curated dataset is a modified version of the raw input dataset where the bad features, the outliers, the inconsistencies and the missing values are highlighted using appropriate color coding. In addition, the names of all the features are highlighted according to their state. Each data provider was finally asked to evaluate the validity of the two reports, as well as, the consistency of the curated dataset.

3.2. Metadata

The UoA dataset consists of 162 features (58 discrete, 73 continuous, 31 unknown) and 440 instances with 44.56% missing values in total (Table 1). Out of 162 features, 91 were characterized as problematic; 60 features with more than 50% missing values and 31 features with unknown data type. The HUA dataset consists of 204 features (104 discrete and 94 continuous) and 100 instances with 33.61% missing values (Table 1). Out of 204 features, 69 were characterized as problematic; 63 features with more than 50% missing values and 6 features

Table 1
Cohorts metadata.

Cohort	UoA	HUA
Number of features	162	204
Number of instances	440	100
Discrete features	58	104
Continuous features	73	94
Problematic features	91	69
Missing values (%)	44.56	33.61

with unknown data types.

3.3. Data quality control

3.3.1. Anomaly detection

For detecting outliers we implemented two local metrics per feature, i.e., the z-score, and the interquartile range, as well as, two global measures, i.e., the Grubb's test and the Local Outlier Factor, (see Section 2.2 for more information). An example of the boxplot for four random features is depicted in Fig. 2(A), where values higher than 75% or lower than 25% of the value range are considered as outliers. The overall LOF distribution is depicted in Fig. 2(B), where values close to 1 are considered as outliers. Since the LOF is a multivariate method, its application is constrained to features with equal number of samples (and no missing values) and thus the LOF was computed only for the "good" features, i.e., those without any missing values. According to Fig. 2(B), the LOF distribution does not indicate the existence of outliers due to the small number of "good" features. As for the rest of the methods, the missing values were ignored during the outlier detection process since they are univariate. The z-scores have also been computed for each feature. The z-score distributions of the four features of Fig. 2(A) are depicted in Fig. 3, where the features with values larger than 3 or lower than −3 are considered as outliers. The identified outliers were derived by four randomly selected features from the UoA cohort.

3.3.2. Similarity detection and de-duplication

As far as similarity detection is concerned, we used the Spearman's rank-order correlation coefficient as a similarity measure instead of the classic Pearson correlation which assumes that the dataset is normally distributed. The Spearman coefficient was computed for each pair of features resulting in a 162 × 162 adjacency matrix for the UoA cohort and an 204 × 204 matrix for the HUA cohort, with correlation values in the range $(-1, 1)$, where 0 implies no correlation and +1 implies strong correlation (Fig. 4A, C). In each adjacency matrix, the field (i, j) corresponds to the Spearman correlation between the features i and j. Each pair is also accompanied by a p-value which denotes the statistical significance of the correlation value (the confidence interval was set to 99%). Then, the pairs (i, j) having similarity value larger than 90% and $p < 0.01$, are highlighted as highly-significant and correlated features. The Jaro distance has been also computed between each pair of feature labels to seek for potential duplicate features yielding a 162x162 lexical distance matrix for the UoA cohort and an 204x204 lexical distance matrix for the HUA cohort, where 0 implies no string matching and +1 implies features have the exact same labels (Fig. 4(B, D)).

The following pairs of features were highlighted for clinical evaluation from the UoA cohort; as highly-correlated[1]: (i) {"Raynaud's phen (0–1)", "Rayanud"} with ($rho = 0.93$, $p < 0.01$), (ii) {"Ro/La", "Anti-Ro (0–1)"} with ($rho = 0.95$, $p < 0.01$), and (iii) {"Date of first biopsy", "Year of disease diagnosis"} with ($rho = 0.93$, $p < 0.01$), and as duplicate names: (i) "Lymphoma score" and "Lymphoma (0–1)" with $s = 0.95$, (iv) "Lymphadenopathy (0–1) (fixed)" and "Lymphadenopathy" with $s = 1$, (v) "Rose-Bengal Stain (0–1)" and "Rose-Bengal Stain" with $s = 0.98$. As for the HUA cohort, the following pairs of features were highlighted as highly-correlated for clinical evaluation: (i) {"Antibodies to Ro(SSA) or La(SSB) antigens, or both at diagnosis or during follow-up", "Anti-Ro positive at diagnosis or during follow-up"} with ($rho = 1$, $p < 0.01$), (ii) {"Muscle biopsy", "Myopathy at diagnosis or during follow-up related to disease"} with ($rho = 1$, $p < 0.01$), and (iii) {"Elevated serum Creatinine", "Kidney Interstitial disease at diagnosis or during follow-up related to disease"} with ($rho = 0.91$, $p < 0.01$). In this case, no duplicate names were detected.

3.4. Data quality assessment report

The data quality assessment report is one major output of the data curator which summarizes useful information regarding the value range of each feature, the type of each feature, the number of missing values, the state of each feature (based on the missing values), whether outliers were detected or not, and finally compatibility issues. An instance of the data quality assessment report can be seen in Table 2 for the UoA cohort and in Supplementary Table 1 for the HUA cohort.

Each variable (feature) is categorized into four different types of groups, namely: integer, float, date, and string. Moreover, there is an extra characterization into categorical and numeric. Categorical variables are those with binary values where the rest of the variables are denoted as numeric. This extra characterization can help the clinician to identify cases where categorical variables take values larger than 1 although such cases will be already detected as outliers. The state of each variable is denoted as good, fair, or bad, according to the number of missing values (see Section 2.2.3). For illustration purposes, we selected the z-score as a measure to detect outliers. Outlier detection is not-applicable in cases where the features have unknown or string data type, as well as, in cases where the features are completely empty. For good features, outlier detection is normally applied whereas for fair or bad features, outlier detection is applied on the non-missing values for maximizing the impact of detecting extreme values in such cases.

Examples of features with unknown type of data for the UoA cohort, include the "Rose-Bengal Stain (0–1)", where the range of the values include a symbol "+" which is unknown and probably denotes positivity, the "Urine pH at last visit" which includes a value "oj" that is probably erroneously parsed, among others. These unknown symbols are also highlighted in the curated dataset with red color as incompatibilities. An example of an outlier for the UoA cohort can be seen in the variable "Date of first biopsy", where the value range includes a minimum value of 801 and a maximum value of 19997, which denotes a discrepancy since it does not correspond to an ordinary year. A similar example, occurs for the variable "Lymphadenopathy", where the normal range is "[0, 1]" but there exists a maximum value of 1994 which probably denotes a year that has been erroneously filled. In total, 13 features were characterized as problematic due to several discrepancies (as described above) and 43 features were highlighted for outliers.

As for the HUA cohort, examples of features with unknown type of data, include the variable "Lymphoma" where the normal range is "[0, 1]" but a string "NHL" exists which is unknown. Another example, includes the variable "Other biopsies-sites" which takes values in the range "[0, 8]" but there are cases with patients having more than one values which denote that these patients have conducted biopsies in more than one sites, a fact that confuses the processing of data. An example of an outlier can be seen in the variable "HBsAg" where although the defined range is "1" for positive and "2" for negative, there are several zero values. The same applies for the variable "HCV". In total, 6 features were characterized as problematic due to several discrepancies and 82 features were highlighted for potential outliers.

The results of the data curator REST service is depicted in Fig. 5 for a pSS-related dataset. Through the REST settings, the user can define a

[1] The term "rho" denotes the Spearman correlation value whereas the term "s" denotes the Jaro distance score.

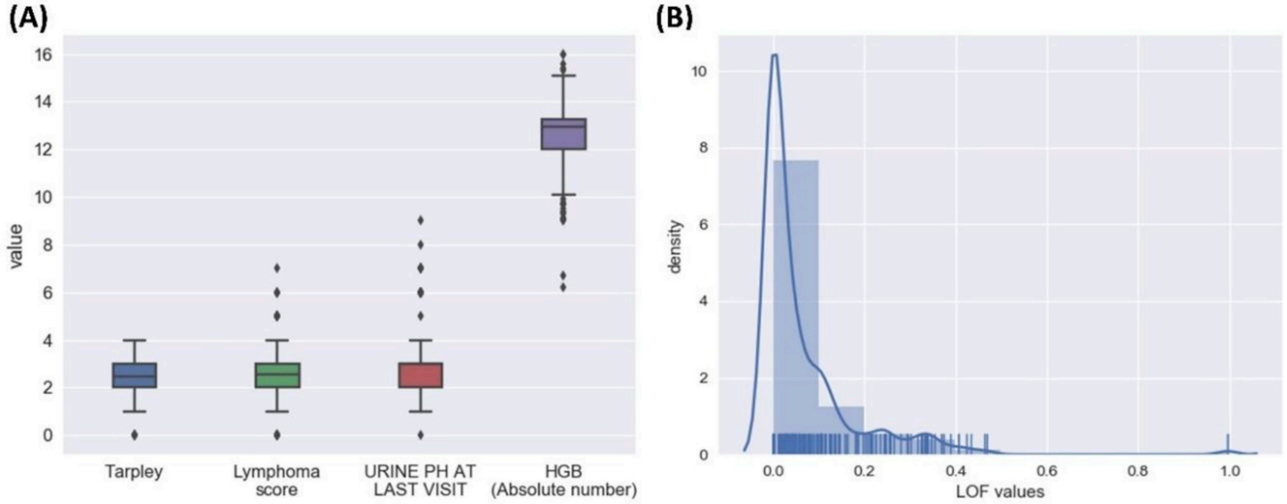


Fig. 2. Results of two methods for outlier detection: (A) A boxplot for outlier detection based on the Interquartile Range (IQR) method for four randomly selected features, and (B) the overall Local Outlier Factor (LOF) distribution across a specific group of features of the dataset, where the density is the normalized frequency and the density curve is a smooth distribution over the histogram.

local or global method for outlier detection (z-score, Interquartile range, Grubb's test, Local Outlier Factor). An example of the data quality assessment panel (are similar to those that are presented in Tables 2 and 3) can be displayed in Fig. 5(C). An example of the produced curated dataset is displayed in Fig. 5(C), where: (i) the outliers are highlighted with gold color, (ii) the problematic fields are highlighted with red color, (iii) missing values with gray, (iv) fair features with green, good features with blue and bad features with rose. The XML schema [14,15] of the reference model has been already incorporated in the service, as well as, the NLTK language toolkit [17] for data standardization purposes. In fact, the majority of the returned parameters from the.JSON structure are already summarized in the produced data quality assessment report for easiness (see Table 2).

3.5. Data standardization report

We used an updated version of a reference model that was developed in a previous study [22] where a chart describing all the necessary requirements for defining the domain knowledge of the pSS (i.e., attributes descriptions and values) was provided by the clinical experts. The updated chart includes information regarding the ranges of the attributes and the class (category) where each attribute belongs to. Using this chart, a complete reference model was developed to reflect the meaning and range of each field. This common template includes a variety of patient-related information, such as, demographics, clinical tests, therapies. The types and ranges of each specified variable within the template were determined during the development process according to the guidelines we received from the clinical experts.

Fig. 6 depicts an example of the standardization process. The reference model is depicted as an XML schema (i.e., a semantic representation or an ontology) which describes the reference model using classes, sub-classes, and object properties. Each class consists of variables where each variable has a range which serves as a set of mapping values, a type, and its parent. Thus, the ontology can be seen as a three-level hierarchical model. In the first level lies the main class "Patient" which consists of four subclasses, i.e., (i) the "Demographics", (ii) the "Clinical tests", (iii) the "Therapies", and (iv) the "ESSDAI[2] domain scores" (that belong to Level 2). Each class in Level 2 has further subclasses (i.e., "Ocular tests", "Oral tests", "Laboratory tests") or variables (e.g., "C4 (mg/dL)") that belong to Level 3. For illustration purposes, the depicted schema describes only an instance of the pSS domain.

A vocabulary was created using the pSS reference model. The terms of the reference model have been incorporated into an XML schema, so that the algorithm can automatically extract these terms and create the vocabulary. The classes denoted in Level 2 (Fig. 6) were also specified by the clinical experts. The vocabulary consists of pairs (x, y) where x is the term of the reference model (e.g., "C3 (mg/dL)") and y is the class it belongs to (e.g., "Clinical tests"). The NLTK's WordNet corpus reader was used to enrich the existing vocabulary by computing synonymous/homonymous terms for each term. For example, "gender", and "sexual relations", are indicative examples of synonymous sets (also referred to as synsets) for the term "sex". The Jaro distance measure was used to calculate the similarity between each term of the raw dataset with those from the vocabulary. Matching block methods were used to match blocks among the terms and rules were developed according to standard value descriptions. The result of the standardization procedure is a tuple (x_{raw}, x_{ref}, v, c), where x_{raw} is the term of the raw dataset, x_{ref} is the matching term from the reference model, v is the matching score, and c is the class where x_{raw} belongs to.

An illustration of the data standardization procedure is depicted in Fig. 6, for a random instance of the UoA cohort dataset. According to Fig. 6, the data standardization module receives as input the raw dataset. Then, it matches the term "SEX (female = 1)" of the input dataset with the homonymous term "gender" of the reference model and finally classifies it into the class "Demographics". An example that involves the set of mapping values (i.e., the standard range) is depicted for the rest of the terms. The algorithm not only matches the term "Abnormal Shirmer's" with the term "Schirmer's test" and classifies it into the parent class "Clinical tests", but also captures information related to the conversion of its value range from "0" and "1" to "1" and "2", respectively. In addition, the term "ANA + " is matched with the term "ANA" and classified into the class "Clinical Tests" with additional information regarding the mapping of the "0" and "1" values to "yes" and "no". Another example is shown for the term "RF ($< 20 = 0$, $> 20 = 1$) IU/mL" which is first matched with the term "RF" (Clinical Tests) and the mapping involves the conversion of the "0" and "1" values to "normal" and "high". In a similar way, the "interstitial renal disease" and "Lymphadenopathy" terms are matched with the terms "Renal domain" and "Lymphadenopathy and Lymphoma domain" of the class "ESSDAI domain", respectively, where the mapping involves the conversion of

[2] The EULAR Sjögren's syndrome (SS) disease activity index (ESSDAI) is a systemic disease activity index proposed by the European League Against Rheumatism (EULAR) that was designed to measure disease activity in patients with primary SS [30].

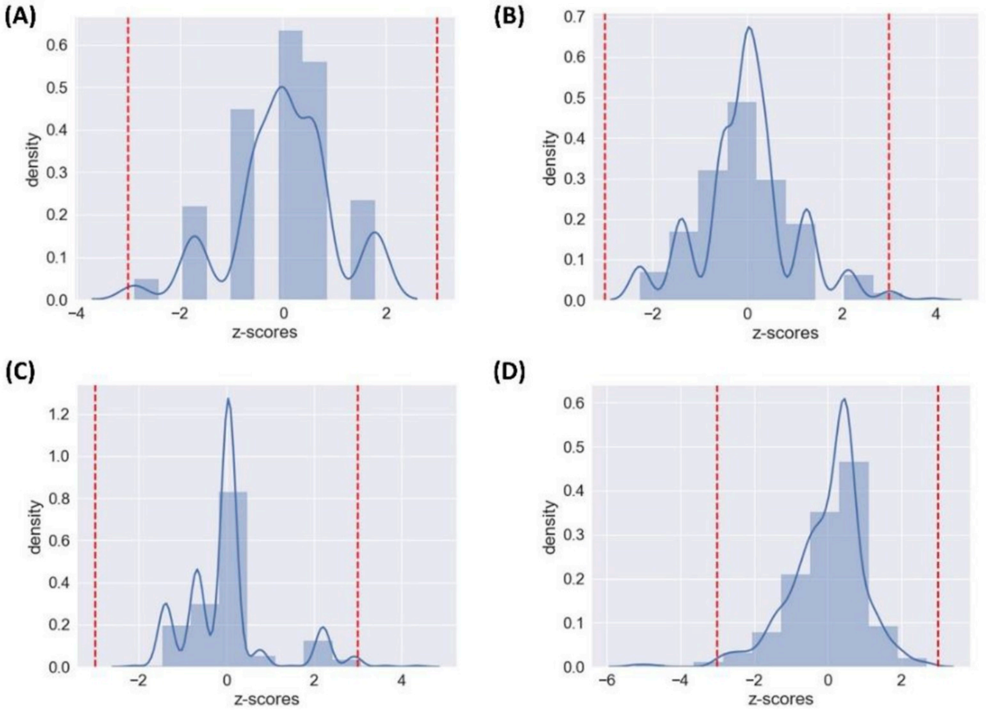


Fig. 3. Z-score distributions for the four features of Fig. 2.1(A). Values that lie outside the red vertical lines are considered as outliers: (A) Tarpley, (B) Lymphoma score, (C) Urine pH at first visit, and (D) HGB (absolute number), where HGB stands for hemoglobin. In each plot, the density is the normalized frequency and the density curve is a smooth distribution over the histogram.

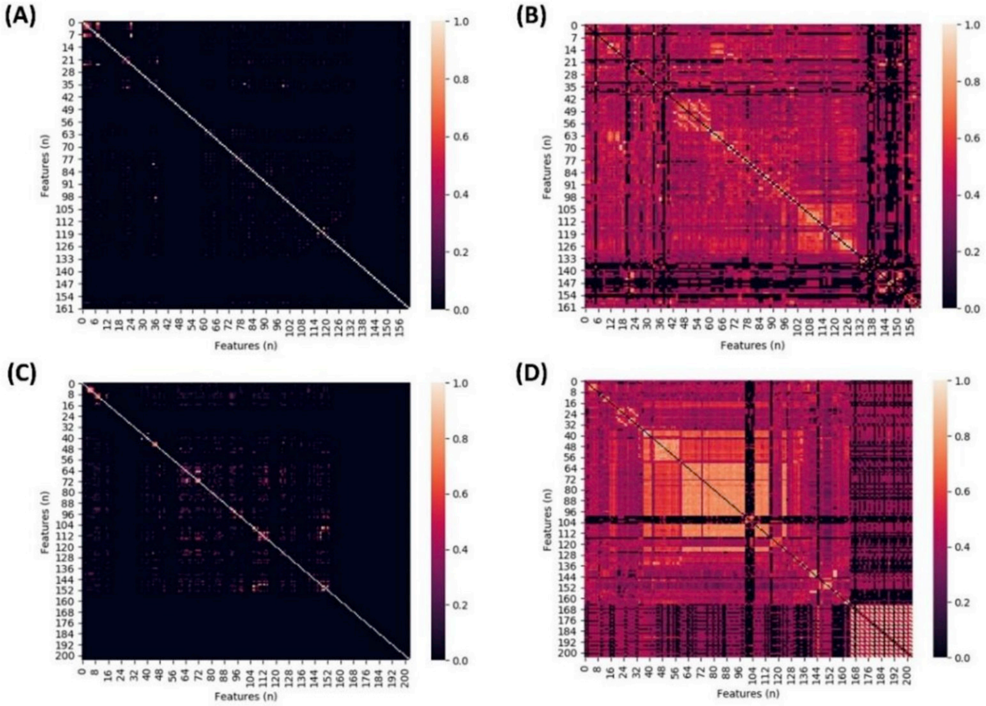


Fig. 4. Correlation and lexical distance matrices for detecting highly-correlated and duplicated terms. (A) The 162 × 162 correlation matrix for the UoA dataset along with (B) the lexical distance matrix, (C) the 204 × 204 correlation matrix for the HUA dataset along with (D) the corresponding lexical distance matrix. The colorbars in the correlation and the lexical distance matrices is used to quantify the importance of the Spearman correlation and the lexical similarity between each pair of features, respectively. A cell (i, j) that is depicted in black color denotes the absence of correlation (or lexical similarity) among the distribution of features i and j, whereas the light orange color denotes a strong correlation (> 0.9) between them.

the values “0” and “1” to “yes” and “no”. Finally, the term “C4 (mg/mL)” is matched with the term “C4 (mg/dL)” along with additional information regarding the conversion of its measurement units from “mg/mL” into “mg/dL”.

The data standardization report provides useful information that can be used for data harmonization, such as, matching terms with similarity scores, final range of values and the classes of the ontology where the matched terms belong to. At this point, it is important to note that not all terms of the dataset are pSS-relevant. The pSS reference model consists of pSS-related variables that are grouped into categories and describe the minimum requirements of the pSS domain and belong to different sub-domains, such as, demographics, laboratory tests, therapies, and ESSDAI domains. An example of a category is the salivary gland biopsy which consists of 10 variables (e.g., year of biopsy, age of patient at year of biopsy, focus score, etc.), the ESSDAI domain which consists of 12 sub-domains, including the constitutional domain, the lymphadenopathy domain, the glandular domain, the articular domain, etc. In total, there are 71 categories most of which involve more than one variable. According to the pSS minimum criteria that were posed by the clinical experts, the number of relevant terms for the UoA cohort was 81 out of 162 (Table 3) and for the HUA cohort, the number of relevant terms was 60 out of 204 (Supplementary Table 2).

Table 2
An instance of the data quality assessment report for the UoA cohort.

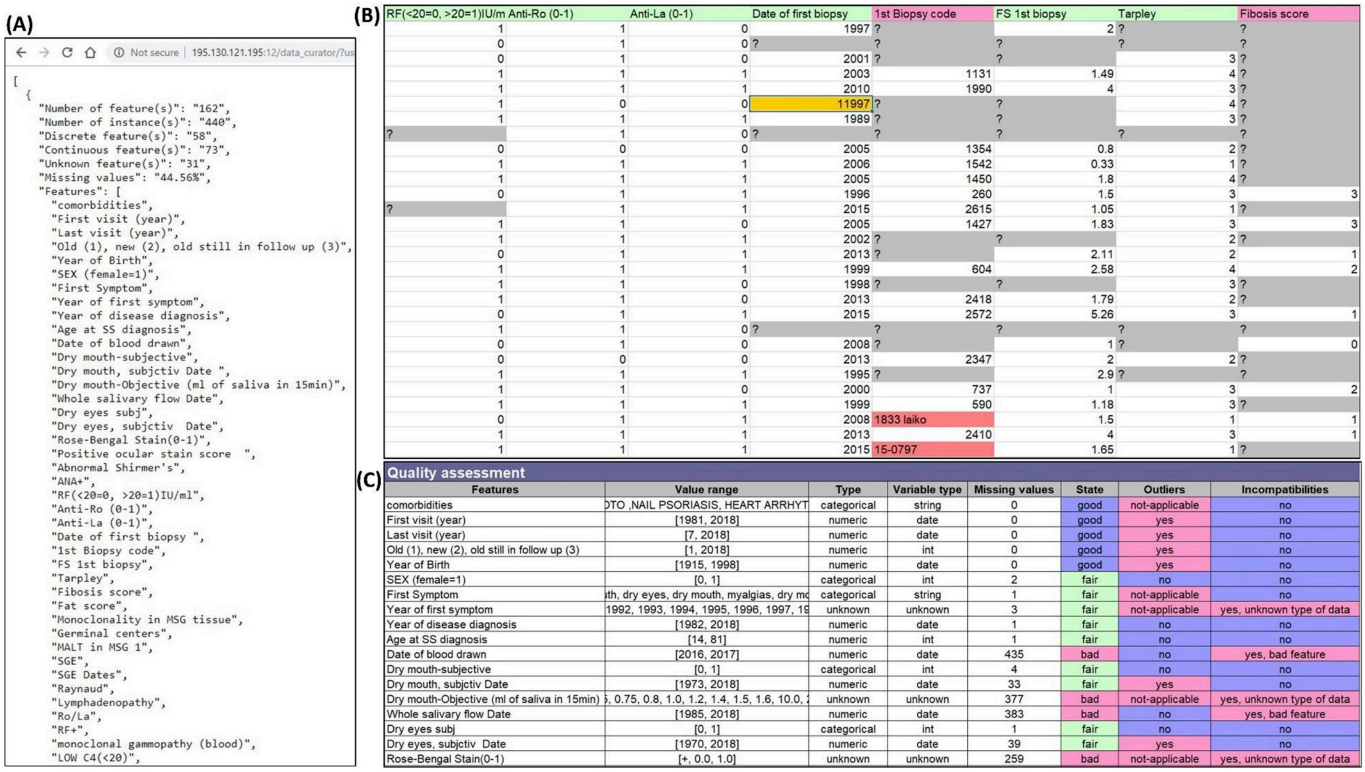
Feature	Value range	Type	Var. type	Missing values	State	Outliers[1]	Compatibility issues
SEX (female=1)	[0, 1]	categorical	int	2	fair	no	no
First visit (year)	[1981, 2018]	numeric	date	0	good	yes	no
Year of disease diagnosis	[1982, 2018]	numeric	date	1	fair	no	no
Age at SS diagnosis	[14, 81]	numeric	int	1	fair	no	no
Whole salivary flow Date	[1985, 2018]	numeric	date	178	bad	no	yes, bad feature
Dry eyes subj	[0, 1]	categorical	int	1	fair	no	no
Dry eyes, subjctiv Date	[1970, 2018]	numeric	date	39	fair	yes	no
Rose-Bengal Stain (0-1)	[+, 0, 1]	unknown	unknown	259	bad	not-applicable	yes, unknown type of data
Positive ocular stain score	[1/9, 5/9, 6/9, 8/9, 9/9]	categorical	string	429	bad	not-applicable	yes, bad feature
Abnormal Shirmer's	[0, 1]	categorical	int	96	fair	no	no
ANA+	[0, 1]	numeric	int	16	fair	no	no
RF (<20=0, >20=1) IU/ml	[0, 1]	categorical	int	52	fair	no	no
Anti-Ro (0-1)	[0, 1]	categorical	int	5	fair	no	no
Anti-La (0-1)	[0, 1]	categorical	int	10	fair	no	no
Date of first biopsy	[801, 19997]	numeric	date	16	fair	yes	no
Fat score	[0, 5]	numeric	int	345	bad	no	yes, bad feature
Germinal centers	[0, 1]	categorical	int	222	bad	no	yes, bad feature
MALT in MSG 1	[0, 1]	categorical	int	40	fair	no	no
SGE	[0, 1]	categorical	int	6	fair	no	no
Raynaud	[0, 1]	categorical	int	3	fair	no	no
Lymphadenopathy	[0, 1994]	numeric	int	5	fair	yes	no
Ro/La	[0, 1]	categorical	int	7	fair	no	no
RF+	[0, 1]	categorical	int	51	fair	no	no
LOW C4 (<20)	[0, 5]	numeric	int	38	fair	yes	no
Lymphoma score	[0, 7]	numeric	int	133	fair	yes	no
Type of monoclonal gammopathy	[IgA(κ), IgGk, IgGλ, IgMλ]	categorical	string	436	bad	not-applicable	yes, bad feature
MSG 2nd bxClonality Bx (0-1)	[0, 1, 1?]	unknown	unknown	416	bad	not-applicable	yes, unknown type of data
Urine pH at last visit	[0, 5, 5.5, 6, 6.5, 7, 8, 8.5, oj]	unknown	unknown	236	bad	not-applicable	yes, unknown type of data
Monocyte number	[42, 7540]	numeric	int	233	bad	yes	yes, bad feature
HGB (absolute number)	[6.2, 16]	numeric	float	76	fair	no	no
CRP (0,1)	[0, 1]	categorical	int	34	fair	no	no
Anti-HCV (0-1)	[0, 1]	categorical	int	237	bad	no	yes, bad feature
Anti-HTLV-1 (0-1)	[0, 0]	numeric	int	465	bad	no	yes, bad feature
ANA (titer-1)	[0, 1/80, 1, 11280, <1/160, >1/640]	unknown	unknown	24	fair	not-applicable	yes, unknown type of data
IgG	[27.3, 5580]	numeric	float	291	bad	yes	yes, bad feature
IgM	[1.5, 1711]	numeric	float	294	bad	yes	yes, bad feature
LDH	[113, 495]	numeric	int	82	fair	no	no
AMA (titer-1)	[0, 1, 1/160, 164]	unknown	unknown	297	bad	not-applicable	yes, unknown type of data
Anti-TPO (0,1)	[0, 1]	categorical	int	211	fair	no	no
Anti-TG (titer)	[0, 1]	categorical	int	226	bad	no	yes, bad feature
Lymphoma (0-1)	[0, 1]	categorical	int	1	fair	no	no

[1] The z-score was used as the outlier detection method.

Note: The highlighted rows correspond to features where outlier detection was not-applicable (either empty features or features with unknown type of data).

Note that the HUA cohort is a rich clinical-oriented database (with detailed symptomatology) instead of a research-oriented one (UoA).

Regarding the UoA cohort, the data standardization module was able to successfully match and classify 73 out of 82 (89.02%) pSS-related terms (Table 3). As far as the HUA cohort is concerned, the data standardization module was able to successfully match and classify 52 out of 60 (86.6%) pSS-related terms (Supplementary Table 2). In both tables, the matching terms are stated along with their similarity (matching) score, the final range of values and the class they belong to (1 = "Demographics", 2 = "Clinical tests", 3 = "Therapies", 4 = "ESSDAI domain"). Similarity scores were computed using the Jaro distance as a string matching metric and the sequence matcher algorithm to identify matching blocks (patterns) among the terms of the input dataset with those from the reference model. Exact matches are those due to identical matching or due to synonymous matching and the similarity score is always equal to 1. The rest of the matches are considered as partial. Partial matches are those that either achieve similarity score larger than 0.9 and/or when the sequence matcher detects matching blocks between the terms. If the sequence matcher identifies exact matching blocks among two terms, the similarity score is set to 1. The small number of pSS parameters that are observed in both cohorts comes from the fact that there exists a large group of



← → C ⌂ ⓘ Not secure | 195.130.121.195:12/data_curator/?us

```
[
  {
    "Number of feature(s)": "162",
    "Number of instance(s)": "440",
    "Discrete feature(s)": "58",
    "Continuous feature(s)": "73",
    "Unknown feature(s)": "31",
    "Missing values": "44.56%",
    "Features": [
      "comorbidities",
      "First visit (year)",
      "Last visit (year)",
      "Old (1), new (2), old still in follow up (3)",
      "Year of Birth",
      "SEX (female=1)",
      "First Symptom",
      "Year of first symptom",
      "Year of disease diagnosis",
      "Age at SS diagnosis",
      "Date of blood drawn",
      "Dry mouth-subjective",
      "Dry mouth, subjctiv Date ",
      "Dry mouth-Objective (ml of saliva in 15min)",
      "Whole salivary flow Date",
      "Dry eyes subj",
      "Dry eyes, subjctiv  Date",
      "Rose-Bengal Stain(0-1)",
      "Positive ocular stain score  ",
      "Abnormal Shirmer's",
      "ANA+",
      "RF(<20=0, >20=1)IU/ml",
      "Anti-Ro (0-1)",
      "Anti-La (0-1)",
      "Date of first biopsy ",
      "1st Biopsy code",
      "FS 1st biopsy",
      "Tarpley",
      "Fibosis score",
      "Fat score",
      "Monoclonality in MSG tissue",
      "Germinal centers",
      "MALT in MSG 1",
      "SGE",
      "SGE Dates",
      "Raynaud",
      "Lymphadenopathy",
      "Ro/La",
      "RF+",
      "monoclonal gammopathy (blood)",
      "LOW C4(<20)",
```

RF(<20=0, >20=1)IU/m	Anti-Ro (0-1)	Anti-La (0-1)	Date of first biopsy	1st Biopsy code	FS 1st biopsy	Tarpley	Fibosis score
1	1	0	1997	?	2	?	?
0	1	0	?	?	?	?	?
0	1	0	2001	?	?	3	?
1	1	1	2003	1131	1.49	4	?
1	1	1	2010	1990	4	3	?
1	0	0	11997	?	?	4	?
1	1	1	1989	?	?	3	?
?	1	0	?	?	?	?	?
0	0	0	2005	1354	0.8	2	?
1	1	1	2006	1542	0.33	1	?
1	1	1	2005	1450	1.8	4	?
0	1	1	1996	260	1.5	3	3
?	1	1	2015	2615	1.05	1	?
1	1	0	2005	1427	1.83	3	3
1	1	1	2002	?	?	2	?
0	1	1	2013	?	2.11	2	1
1	1	1	1999	604	2.58	4	2
1	1	0	1998	?	?	3	?
1	1	0	2013	2418	1.79	2	?
0	1	1	2015	2572	5.26	3	1
1	1	0	?	?	?	?	?
0	1	0	2008	?	1	?	0
0	0	0	2013	2347	2	2	?
1	1	1	1995	?	2.9	?	?
1	1	0	2000	737	1	3	2
1	1	1	1999	590	1.18	3	?
0	1	1	2008	1833 laiko	1.5	1	1
1	1	1	2013	2410	4	3	1
1	1	1	2015	15-0797	1.65	1	?

Quality assessment

Features	Value range	Type	Variable type	Missing values	State	Outliers	Incompatibilities
comorbidities	ITO ,NAIL PSORIASIS, HEART ARRHYT	categorical	string	0	good	not-applicable	no
First visit (year)	[1981, 2018]	numeric	date	0	good	yes	no
Last visit (year)	[7, 2018]	numeric	date	0	good	yes	no
Old (1), new (2), old still in follow up (3)	[1, 2018]	numeric	int	0	good	yes	no
Year of Birth	[1915, 1998]	numeric	date	0	good	yes	no
SEX (female=1)	[0, 1]	categorical	int	2	fair	no	no
First Symptom	ıth, dry eyes, dry mouth, myalgias, dry mo	categorical	string	1	fair	not-applicable	no
Year of first symptom	1992, 1993, 1994, 1995, 1996, 1997, 19	unknown	unknown	3	fair	not-applicable	yes, unknown type of data
Year of disease diagnosis	[1982, 2018]	numeric	date	1	fair	no	no
Age at SS diagnosis	[14, 81]	numeric	int	1	fair	no	no
Date of blood drawn	[2016, 2017]	numeric	date	435	bad	no	yes, bad feature
Dry mouth-subjective	[0, 1]	categorical	int	4	fair	no	no
Dry mouth, subjctiv Date	[1973, 2018]	numeric	date	33	fair	yes	no
Dry mouth-Objective (ml of saliva in 15min)	, 0.75, 0.8, 1.0, 1.2, 1.4, 1.5, 1.6, 10.0,	unknown	unknown	377	bad	not-applicable	yes, unknown type of data
Whole salivary flow Date	[1985, 2018]	numeric	date	383	bad	no	yes, bad feature
Dry eyes subj	[0, 1]	categorical	int	1	fair	no	no
Dry eyes, subjctiv Date	[1970, 2018]	numeric	date	39	fair	yes	no
Rose-Bengal Stain(0-1)	[+, 0.0, 1.0]	unknown	unknown	259	bad	not-applicable	yes, unknown type of data

Fig. 5. The results of the data curator REST service execution on the UoA cohort dataset that lies in a secure private cloud space: (A) An instance of the returned.JSON structure of the REST service call, (B) an instance of the curated dataset, and (C) an instance of the data quality assessment report.

variables which is related to the symptomatology of the different ESSDAI domains, including the “arthralgia” in the arterial domain, the “lung involvement” in the pulmonary domain, the “myositis” in the muscular domain, the “palpable purpura” and “non-palpable purpura” in the cutaneous domain, the “weight-loss” and “fever” symptoms in the constitutional domain, the “kidney involvement” in the renal domain, etc. In fact, there are 41 ESSDAI-related symptoms in the HUA cohort and 31 in the UoA cohort that are not listed in the reference model, with the purpose of creating a more research-oriented data model.

3.6. Performance

The REST service was implemented in Python 3.6 and was executed twice, one for each cohort, under a typical environment and through a secure virtual private network (VPN). The average execution time of the web service for the UoA cohort was 3.79 s whereas for the HUA cohort the execution time was equal to 1.9 s (Fig. 7). More specifically, the time for fetching data was almost equal for both cohorts (< 1 s). The average execution time for the application of the service including, data annotation and evaluation, outlier detection, similarity detection, and standardization, was 1.1 s for the UoA cohort and 1.3 s for the HUA cohort. The average execution time for constructing the data evaluation and data standardization reports along with the curated dataset was equal to 9 s for the UoA cohort and 4 s for the HUA cohort. According to Fig. 7, the execution time for the data quality operations is affected by the number of features (for the HUA cohort the number of features is larger than the number of features from the UoA cohort) whereas the time for constructing the reports (data quality assessment and standardization) and the curated dataset is affected by the number of patients (for the UoA cohort the number of patients is 4.4 times larger than the number of patients from the HUA cohort).

The small execution time demonstrates the dominance of automated data curation against traditional manual data curation where the time for identifying the outliers, and inconsistencies by both clinicians was large enough due to the size and complexity of the datasets. The curated dataset was able to highlight all the cases with unknown data types, outliers and missing values, informing the clinicians that these cases would need their attention in just a few seconds. The data evaluation report was able to summarize the metadata information in the same amount of time.

4. Discussion

The absence of medical curative treatments often leads to studies with inaccurate and incomplete data which are characterized by small statistical power. However, curation methods shall be used with caution since it is more likely to make things worse. In this work, we propose an automated framework for data curation. Furthermore, we extend data standardization as a pre-harmonization process to make data harmonization, which follows, easier and faster. Through this procedure, we attempt to produce semantic relations between the fields of the raw dataset with those from a reference dataset and therefore enhance the semantic matching process for data harmonization. The proposed framework consists of a three-layer architecture which is scalable and able to deal with incomplete terminologies, irrelevant terms, outliers, missing values, data categorization, and duplicated terms. In the core of this framework lies data standardization. The framework was evaluated on two anonymized clinical datasets from two cohorts of patients with pSS, highlighting the importance of the proposed framework for data quality assessment and data harmonization.

The data evaluation module was able to capture a first look into the dataset's structure and vocabulary. The data quality control module was able to identify outliers and missing values, as well as, detect fields with similar context and duplicated terms. The ability to choose among different outlier detection methods (z-scores, Grubb's test, IQR, LOF) increases the statistical power of the outcomes. The clinicians successfully validated the accuracy of the problematic fields which were correctly identified in both cohorts. In addition, the detected outliers

Table 3
An instance of the data standardization report for the UoA cohort.3

Feature	Matched term or category from the reference model	Score	Type of match[1]	Captured range or measurement unit	Class[2]
First visit (year)	Age at inclusion	1	partial	[1981, 2018]	1
Last visit (year)	Age at last follow-up	1	partial	[1991, 2018]	1
Year of birth	Year of birth	1	exact	[1918, 1995]	1
SEX (female=1)	Gender	1	exact	[0, 1]	1
Year of first symptom	Age at onset of first symptom	1	partial	[1971, 2016]	1
Year of disease diagnosis	Age at diagnosis of pSS	1	partial	[1983, 2018]	1
Age at SS diagnosis	Age at diagnosis of pSS	1	partial	[17, 84]	1
Dry mouth-subjective	Oral dryness	1	partial	[yes, no]	2
Dry eyes, subjctiv Date	Ocular dryness	1	partial	[1976, 2017]	2
Rose-Bengal Stain(0-1)	Rose-Bengal	1	partial	[0, 1, +]	2
Positive ocular stain score	Ocular staining score	1	partial	[1/9, 5/9, 6/9, 8/9, 9/9]	2
Abnormal Shirmer's	Schirmer's test	1	partial	[1, 2]	2
ANA+	ANA	1	partial	[yes, no]	2
RF(<20=0, >20=1) IU/ml	RF	1	partial	[normal, high]	2
Anti-La (0-1)	Anti-La	1	partial	[yes, no]	2
Monoclonality in MSG tissue	Serum monoclonal M component	1	partial	[yes, no]	2
MALT in MSG 1	Minor salivary gland biopsy	1	partial	[0, 1]	2
RF+	Rheumatoid factor	1	partial	[normal, high]	2
monoclonal gammopathy (blood)	Serum monoclonal M component	1	partial	[yes, no]	2
LOW C4 (<20)	C4	1	partial	mg/dL	2
Lymphoma score	Lymphadenopathy and lymphoma domain	1	partial	[0, 7]	4
Type of monoclonal gammopathy	Serum monoclonal M component	1	partial	[yes, no]	2
Time of 2st MSG biopsy (mm/yr)	Minor salivary gland biopsy	1	partial	[1985, 2017]	2
Code 2nd MSG Biopsy	Minor salivary gland biopsy	1	partial	[1231, …, parotid]	2
MSG 2nd bx Focus Score	Minor salivary gland biopsy	1	partial	[0.22, 12]	2
Time of 3st MSG biopsy (mm/yr)	Minor salivary gland biopsy	1	partial	[2006, 2017]	2
MSG 3nd bx Focus Score	Minor salivary gland biopsy	1	partial	[1.54, 22.84]	2
Time of 4th MSG biopsy (mm/yr)	Minor salivary gland biopsy	1	partial	[2013, 2015]	2
MSG 4th bx Focus Score	Minor salivary gland biopsy	1	partial	[1, 12]	2
Dyspareunia, subjctiv (0-1)	Dyspareunia VAS domain	1	partial	[0, 1]	4
Dyspareunia, subjctiv Date	Dyspareunia VAS domain	1	partial	[1985, 2016]	4
Abnormal Schirmer's test (0-1)	Schirmer's test	1	partial	[1, 2]	2
Schirmer's test date	Schirmer's test	1	partial	[1983, 19984]	2
Rose-Bengal Stain(0-1) 2	Rose-Bengal	1	partial	[0, 1]	2
Lymphadenopathy (0-1) (fixed)	Lymphadenopathy and lymphoma domain	1	exact	[yes, no]	4
Lymphadenopathy(fixed) date(-yr)	Lymphadenopathy and lymphoma domain	1	partial	[1992, 2013]	4
γ-globulins(11-18=0,>18=1,<11=2)	Serum immunoglobulins	1	partial	[normal, high]	2
Anti-HCV (0-1)	anti-HCV antibody	1	exact	[yes, no]	2
ANA(titer-1)	ANA titer	1	partial	[0,1,1/1250, 1/1280, 1/160, 1/2560, 1/320, 1/5120, 1/640, 1/80, <1/160]	2
IgG	IgG	1	exact	[338, 7700]	2
C3 (mg/mL)	C3	1	exact	mg/dL	2
C4 (mg/mL)	C4	1	exact	mg/dL	2
Cryo (0,1)	Cryoglobulinemia	1	partial	[yes, no]	2
Cryo (type-II, IgMk) (0,1)	Cryoglobulinemia	1	partial	[yes, no]	2
Lymphoma (0-1)	Lymphadenopathy and lymphoma domain	1	exact	[yes, no]	4
Lymphoma diagnosis date	Lymphadenopathy and lymphoma domain	1	partial	[1986, 2018]	4
Anti-Ro (0-1)	Anti-Ro	1	partial	[1, ro]	2

Note: The highlighted rows correspond to features with discrepancies (i.e., outliers and/or inconsistent).
[1] Class: 1 = "Demographics", 2 = "Clinical tests", 3 = "Therapies", 4 = "ESSDAI domain".
[2] Type of match: exact = identical or synonymous terms, partial = highly-similar terms or terms with exact matching blocks.

helped the clinicians fix several discrepancies or remove them where necessary, with a validity index of more than 90%. A large portion of the outliers in both cohorts were detected in features with binary values. In most of the cases, binary outliers do not have any clinical importance (e.g., in "Gender" the majority of values is zero which denotes females so values with one (males) are highlighted). The largest

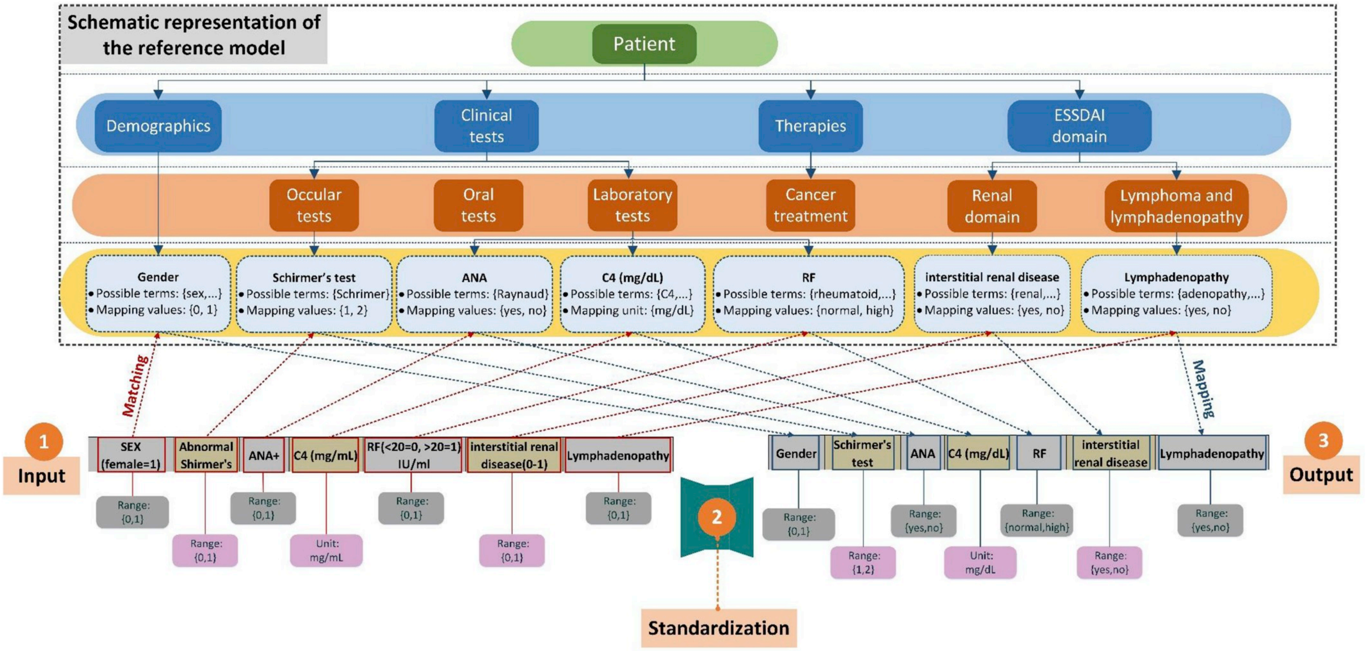

Schematic representation of the reference model
Patient
Demographics
Clinical tests
Therapies
ESSDAI domain
Occular tests
Oral tests
Laboratory tests
Cancer treatment
Renal domain
Lymphoma and lymphadenopathy
Gender • Possible terms: {sex,...} • Mapping values: {0, 1}
Schirmer's test • Possible terms: {Schrimer} • Mapping values: {1, 2}
ANA • Possible terms: {Raynaud} • Mapping values: {yes, no}
C4 (mg/dL) • Possible terms: {C4,...} • Mapping unit: {mg/dL}
RF • Possible terms: {rheumatoid,...} • Mapping values: {normal, high}
interstitial renal disease • Possible terms: {renal,...} • Mapping values: {yes, no}
Lymphadenopathy • Possible terms: {adenopathy,...} • Mapping values: {yes, no}
Matching
Mapping
1
Input
SEX (female=1)
Abnormal Shirmer's
ANA+
C4 (mg/mL)
RF(<20=0, >20=1) IU/ml
interstitial renal disease(0-1)
Lymphadenopathy
Range: {0,1}
Range: {0,1}
Range: {0,1}
Unit: mg/mL
Range: {0,1}
Range: {0,1}
Range: {0,1}
2
Standardization
Gender
Schirmer's test
ANA
C4 (mg/dL)
RF
interstitial renal disease
Lymphadenopathy
Range: {0,1}
Range: {1,2}
Range: {yes,no}
Unit: mg/dL
Range: {normal,high}
Range: {yes,no}
Range: {yes,no}
3
Output

Fig. 6. Illustration of the data standardization procedure.

portion of the missing values was physically explained by the data providers. Most of them had to do with follow-up calculations or records that have been lost in the past. Undoubtedly, the data evaluation report combined with the curated dataset have been proven useful for the clinicians during the data quality assessment process, reducing the time effort needed for manual data curation.

The data standardization module was able to identify and properly classify more than 85% of pSS-related terms for the UoA and HUA cohorts, based on knowledge from the reference model. This highlights the importance of the reference model which stands as a gold standard for matching similar terminologies across heterogeneous data and thus enables data harmonization. However, the percentage of matching terms can be greatly enhanced if the data standardization module receives as input the semantic representation of the raw dataset instead of (only) the clinical one. In addition, a semantic representation of the raw dataset can reduce information loss. An example of how an ontology can reduce information loss and improve the overall matching percentage can be seen in the HUA cohort. The HUA cohort includes nine variables which are not stated in the reference model and are related to the various therapeutic prescriptions, such as, Methotrexate (MTX), Leflunomide, Cyslosporine, Azathiprine, Hydroxychloroquine (HCQ), Mycophenolate mofetil (MMF), Anti-TNFs, Rituximab (RTX), and Belimumab. These variables could be grouped into the class "Therapies" and then the semantic matching process would be able to match this class with the homonymous class of the reference model and thus reduce the information loss by 4% with an additional increase in the matching performance by 2%. As a matter of fact, the vocabulary could be enriched by adding the detailed (sub-)symptomatology related to the different ESSDAI domains so as to increase the matching percentage, as well as, include medical acronyms related to popular laboratory tests, such as, the "HBsAg" which stands for Hepatitis B, the "WBC" which stands for white blood cells, etc.

The fact that the standardization procedure can use an XML representation of the reference model as input, increases its overall performance and introduces the ontologies and semantics as a preliminary step for achieving medical data harmonization. Having two ontologies and seeking for a way to match these two ontologies is a typical semantic matching problem which is one way to achieve data harmonization. Semantics have gained a lot of attention nowadays especially in computer science and linguistics for schema and ontology merging, data migration, query translation, agent communication, etc. [31]. Currently, the data standardization module supports the basic XML format (the basis of almost all types of ontologies and markup languages [31]) for the semantic representation of the reference model. The outcomes of this module are capable of assisting the semantic matching process which requires these matching pairs in order to semantically match each term of the input dataset with those from the reference model and thus enable data harmonization. The problem of harmonizing one dataset based on a standard one can be reduced to the standardization of the semantic representations of these two datasets so as to approximate the semantic matching problem [14,15].

While the variety of the proposed univariate and multivariate methodologies towards outlier detection [19,21] lack of an integrated approach that manages to combine both of these methodologies into a single framework, the presented framework offers an integrated service that includes outlier detection as part of its data quality control strategy. Meanwhile, the majority of the clinical studies for data quality assessment [3–9,11,12] aims to construct a gold standard model (a set of terms which describe the knowledge of a clinical domain) and then use this model to manually or semi-automatically classify the terms of a raw clinical dataset based on their accuracy, relevance, consistency, etc., with the terms of the gold standard model [6–8]. These frameworks however (Table 4): (i) do not use any automated methods towards outlier detection and de-duplication, (ii) focus only on assessing the quality of the terms that are relevant with those from the gold standard model, and (iii) do not provide re-useable data quality assessment reports. The presented framework accounts also for data standardization by producing a set of semantic relations through a rule-driven approach that is developed based on a pre-defined reference model and captures important semantic relations which enable faster data harmonization. In addition, the framework can be easily adjusted with new rules according to a provided reference model that describes the clinical domain of interest. The fact that the framework is validated through a case study on two well-established cohorts of the pSS domain, increases its reliability. The computationally efficient functionalities which are offered are capable of tracking incompatibilities and dealing with outliers and similar terms in large datasets that are stored in tabular formats. The proposed framework aims to bridge a crucial

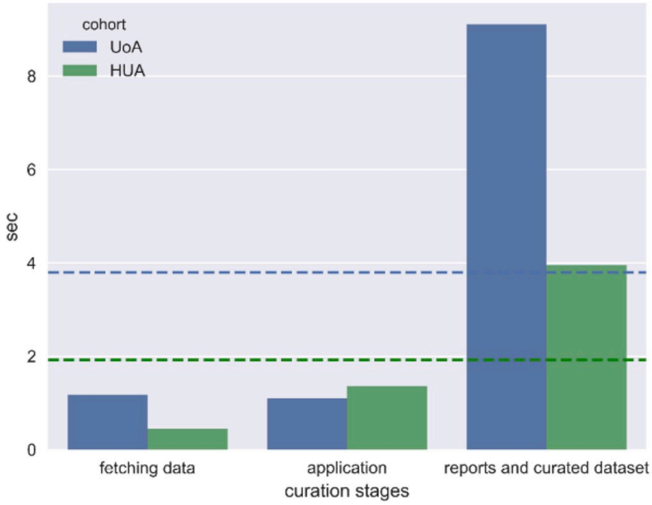


Fig. 7. Execution time (in seconds) for the different stages (i.e., fetching data, application, reports and curated dataset) of the data curator's web service for both cohorts. The average execution times are depicted in horizontal lines (blue color: UoA cohort, green color: HUA cohort).

scientific gap in automated data quality assessment, which is present in the clinical domain.

Moreover, the scope of the proposed framework can be extended to include genome wide association studies (GWAS), where data curation methods can be used to enhance the quality of genomic data. Since the proposed framework is easily scalable, it can be updated to include functionalities for outlier detection, de-duplication, and imputation in genetic data, including DNA and RNA sequencing data, among others, where the existence of missing values, outliers, and similarities can lead to falsified associations between genetic variants (e.g., single-nucleotide polymorphisms (SNPs)) and traits (e.g., diseases). This is present in case-control studies, where common variants are examined between a case group (a group of individuals under a common disease or condition) and a control group (a group of healthy individuals) [32].

5. Limitations

The main limitation of the current study was the comparison of the results of the framework with the clinicians' efforts that were needed in order to conduct a similar manual data quality assessment procedure. For this reason, the outcomes of the framework were (successfully) evaluated by the clinicians in terms of relevance, consistency, and accuracy, with the overall execution time being smaller than any manual attempt for data quality assessment.

6. Conclusions and future work

The presented framework is an integrated, web-based data quality assessment framework which offers an automated service that combines both univariate and multivariate methods for outlier detection and de-duplication, as well as, terminology-based data quality assessment based on data standardization, yielding clinician-friendly data quality assessment reports that promote the re-use of qualified data. The framework is easily scalable and can be incorporated into any medical platform that deals with big data analytics, as part of their data quality assessment strategy. For this reason, the source code of the data curator has been made publicly available under the following github repository: https://github.com/vpz4/Data-curator, along with a brief user manual. Additional applications on clinical databases are necessary to further evaluate the framework's efficacy and reliability, as well as, include more functionalities for outlier and similarity detection. Until now, the framework is executed in the form of a REST service and efforts are needed to publish the service in the form of a user-friendly front-end web interface. The fact that the proposed framework introduces the reference model as a standard model for data standardization can be generalized for different types of diseases due to the scalability it offers. The results of this framework can be also combined with semantic matching algorithms to enable data harmonization in different domains varying from autoimmune diseases to cardiovascular diseases and genomics. In fact, the standardization report can be presented in the form of a drop-down menu, where the clinician will be able to select the best match with the standard term(s) according to the Health Level-7 (HL7) standards.

Table 4
Comparison with related studies.

Study	Objective	Methodology	Comparison
[6]	Assess the quality of electronic medical records (EMR)	A six-step data quality assessment framework that aims to deal with the following data quality dimensions: completeness, correctness, concordance, plausibility and currency, by computing the percentage of matched variables across records, and matched records across patients, as well as, the type of records per patient, the presence of selected variables, and the frequency of records per patient over time.	● Lack of quantitative methods for data quality control (e.g., outlier detection, similarity detection) ● Conceptual presentation of the methodology for matching terms across patients/records
[7]	Assess the quality of electronic health record (EHR) data	A draft set of harmonized terms is presented. The set of terms was defined by the experts and was organized into three quality categories (i.e., conformance, completeness, and plausibility) so as to be compared with ten existing data quality terminologies in the context of electronic health record data, where the comparison was based in terms of coverage in the EHR domain.	● Lack of case studies to prove the superiority of the proposed set of terms against similar ones ● Only qualitative measures are defined for quality improvement ● Lack of quantitative methods for data quality control ● Lack of re-useable quality reports
[8]	Present a framework for data quality management in health care institutions	A framework that deals with the completeness, consistency, correctness, non-redundancy, and timeliness of medical data in a semi-automated way where the user defines the quality mapping criteria (e.g., completeness) and the data quality levels (e.g., acceptable) for each data source.	● The quality assessment process is exclusively based on quality criteria that are manually defined for each individual data source ● Lack of quantitative methods for data curation
Present	Present a web-based framework for medical data quality assessment	A three-step framework is presented that aims to enhance the completeness, relevance, and accuracy of clinical data by providing a set of quantitative functionalities for metadata extraction, data quality control (e.g., outlier detection, de-duplication) and data standardization (based on a set of terms that are lexically matched with those from a standard reference model).	● Produces re-useable data quality reports that can be used to fix outliers, duplicates, missing values, and inconsistencies ● Can be iteratively executed until the data quality criteria are met ● Web-based (REST service) ● Data standardization in terms of data harmonization

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731944 and from the Swiss State Secretariat for Education, Research and Innovation SERI under grant agreement 16.0210.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2019.03.001.

References

[1] L. Cai, Y. Zhu, The challenges of data quality and data quality assessment in the big data era, Data Sci. J. 14 (2015).
[2] C.H. Lee, H.J. Yoon, Medical big data: promise and challenges, Kidney Res. Clin. Pract. 36 (1) (2017) 3–11.
[3] C. Batini, M. Scannapieco, "Data and Information Quality – Dimensions, Principles and Techniques," *Data-Centric Systems and Applications*, Springer, 2016.
[4] H. Chen, D. Hailey, N. Wang, P. Yu, A review of data quality assessment methods for public health information systems, Int. J. Environ. Res. Public Health 11 (5) (2014) 5170–5207.
[5] N.G. Weiskopf, C. Wenig, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, J. Am. Med. Inform. Assoc. 20 (1) (2013) 144–151.
[6] N.G. Weiskopf, G. Hripcsak, S. Swaminathan, C. Weng, Defining and measuring completeness of electronic health records for secondary use, J. Biomed. Inform. 46 (5) (2013) 830–836.
[7] A.P. Reimer, A. Milinovich, E.A. Madigan, Data quality assessment framework to assess electronic medical record data for use in research, Int. J. Med. Inform. 90 (2016) 40–47.
[8] M.G. Kahn, T.J. Callahan, J. Barnard, A.E. Bauck, J. Brown, B.N. Davidson, et al., A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data, Egems 4 (1) (2016).
[9] L. Bai, R. Meredith, F. Burstein, A data quality framework, method and tools for managing data quality in a health care setting: an action case study, J. Decis. Syst. 1–11 (2018).
[10] X. Chu, I.F. Ilyas, S. Krishnan, J. Wang, Data cleaning: overview and emerging challenges, Proceedings of the 2016 International Conference on Management of Data, 2016, pp. 2201–2206.
[11] M. Stonebraker, D. Bruckner, I.F. Ilyas, A. Pagan, S. Xu, Data curation at scale: the Data tamer system, Conference on Innovative Data Systems Research, CIDR, 2013.
[12] P. Lord, A. Macdonald, L. Lyon, D. Giaretta, From data deluge to data curation, In Proceedings of the UK e-science All Hands meeting, 2004, pp. 371–375.
[13] H.C. Koh, G. Tan, Data mining applications in healthcare, J. Healthc. Manag. 19 (2) (2011) 65.
[14] K. Kourou, V.C. Pezoulas, E.I. Georga, T. Exarchos, P. Tsanakas, M. Tsiknakis, T. Varvarigou, S. De Vita, A. Tzioufas, D.I. Fotiadis, Cohort harmonization and integrative analysis from a biomedical engineering perspective, IEEE Rev. Biomed. Eng. 12 (2019) 303–318.
[15] C. Pang C, A. Sollie, A. Sijtsma, D. Hendriksen, B. Charbon, M. de Haan, T. de Boer, F. Kelpin, J. Jetten, J.K. van der Velde, N. Smidt, R. Sijmons, H. Hillege, M.A. Swertz, SORTA: a system for ontology-based re-coding and technical annotation of biomedical phenotype data, Database (Oxford) 18 (2015).
[16] T. Benson, Principles of Health Interoperability HL7 and SNOMED, Springer, 2010, p. 263.
[17] S. Sun, C. Luo, J. Chen, A review of natural language processing techniques for opinion mining systems, Inf. Fusion 36 (2017) 10–25.
[18] L. Otero-Cerdeira, F.J. Rodríguez-Martínez, A. Gómez-Rodríguez, Ontology matching: a literature review, Expert Syst. Appl. 42 (2) (2015) 949–971.
[19] S. SwarupaTripathy, R.K. Saxena, P.K. Gupta, Comparison of statistical methods for outlier detection in proficiency testing data on analysis of lead in aqueous solution, Am. J. Theor. Appl. Stat. 2 (6) (2013) 233–242.
[20] P.J. Rousseeuw, M. Hubert, Robust statistics for outlier detection, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 1 (1) (2011) 73–79.
[21] E. Schubert, A. Zimek, H.P. Kriegel, Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection, Data Min. Knowl. Discov. 28 (1) (2014) 190–237.
[22] V.C. Pezoulas, T.P. Exarchos, V. Andronikou, T. Varvarigou, A. Tzioufas, S. De Vita, D.I. Fotiadis, Towards the establishment of a biomedical ontology for the primary Sjögren's Syndrome, Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, 2018, pp. 4089–4092.
[23] N. Pradhan, M. Gyanchandani, R. Wadhvani, A review on text similarity technique used in IR and its application, Int. J. Comput. Appl. 120 (9) (2015).
[24] M. del Pilar Angeles, A. Espino-Gamez, Comparison of Methods Hamming Distance, Jaro, and Monge-Elkan, *DBKDA*, 2015.
[25] S.J. Gandhi, M.M. Thakor, J. Sheth, H.I. Pandit, H.S. Patel, Comparison of string similarity algorithms to measure lexical similarity, J. Syst. Inf. Technol. 10 (2) (2017) 139.
[26] M. Yu, J. Wang, G. Li, Y. Zhang, D. Deng, J. Feng, A unified framework for string similarity search with edit-distance constraint, VLDB J. 26 (2) (2017) 249–274.
[27] G.A. Rao, G. Srinivas, K.V. Rao, P.P. Reddy, Characteristic mining of mathematical formulas from document-A comparative study on sequence matcher and levenshtein distance procedure, J. Comp. Sci. Eng. 6 (4) (2018) 400–404.
[28] C.P. Mavragani, H.M. Moutsopoulos, Sjögren syndrome, Can. Med. Assoc. J. 186 (15) (2014) E579–E586.
[29] M. Ramon-Casals, P. Brito-Zerón, A. Sisó-Almirall, A.G. Tzioufas, "Topical and systemic medications for the treatment of primary Sjögren's syndrome, Nat. Rev. Rheumatol. 8 (2012) 399–411.
[30] R. Seror, S.J. Bowman, P. Brito-Zeron, E. Theander, H. Bootsma, A. Tzioufas, J.E. Gottenberg, M. Ramos-Casals, T. Dörner, P. Ravaud, C. Vitali, X. Mariette, EULAR Sjögren's syndrome disease activity index (ESSDAI): a user guide, RMD open 1 (1) (2015) e000022.
[31] H. Nacer, D. Aissani, Semantic web services: standards, applications, challenges and solutions, J. Netw. Comput. Appl. 44 (2014) 134–151.
[32] C. Gondro, S.H. Lee, H.K. Lee, L.R. Porto-Neto, Quality control for genome-wide association studies, Methods Mol. Biol. 1019 (2013) 129–147.