ABSTRACT

Objective. To address the need for automatically assessing the quality of clinical data in terms of accuracy, relevance, conformity, and completeness, through the concise development and application of an automated method which is able to automatically detect problematic fields and match clinical terms under a specific domain.

Methods. The proposed methodology involves the automated construction of three diagnostic reports that summarise valuable information regarding the types and ranges of each term in the dataset, along with the detected outliers, inconsistencies, and missing values, followed by a set of clinically relevant terms based on a reference model which serves as a set of terms which describes the domain knowledge of a disease of interest.

Results. A case study was conducted using anonymised data from 250 patients who were diagnosed with primary Sjögren’s syndrome (pSS), yielding reliable outcomes that were highlighted for clinical evaluation. Our method was able to successfully identify 28 features with detected outliers, and unknown data types, as well as, identify outliers, missing values, similar terms, and inconsistencies within the dataset. The data standardisation method was able to match 76 out of 85 (89.41%) pSS-related terms according to a standard pSS reference model which has been introduced by the clinicians.

Conclusion. Our results confirm the clinical value of the data curation method towards the improvement of the dataset quality through the precise identification of outliers, missing values, inconsistencies, and similar terms, as well as, through the automated detection of pSS-related relevant terms towards data standardisation.

Introduction

Data quality assessment lies in the basis of any healthcare system that deals with medical data analytics. Data quality is most commonly defined in terms of accuracy, validity, completeness, precision, integrity, and relevance of a data set according to a gold standard model (1, 2). In the clinical domain, the gold standard can be a set of parameters that describe the knowledge of a disease of interest, including standard measurement units and standard data formats. For example, a complete and relevant clinical dataset is a dataset where the majority of the parameters of the standard dataset (in the same clinical domain) is met. In addition, the presence of outliers and inconsistent values (e.g. unknown data types and symbols) obscures the quality of the data by introducing significant biases. As a matter of fact, the lack of actions for data quality improvement result in clinical data that are useless, irrelevant, and incomplete, a fact that introduces numerous undesirable implications towards their analysis. The most common way for assessing the quality of a dataset is manual data curation, according to which the clinician deals with missing values, inconsistencies, unknown data types, and outliers, by visually inspecting the whole dataset to deal with problematic fields. Manual data curation, however, is extremely time consuming and sometimes is even impossible, especially in the case where the volume of the clinical data is very large (2). In this work, we deploy an automated method which is able to assist the clinician during the data quality assessment procedure by providing clinician-friendly information concerning the precise detection of outliers, missing values, inconsistencies, highly-correlated terms, and unknown data types that are...
present in contaminated clinical datasets, as well as, by automatically matching a group of terms of the input dataset with those from a standard reference model. A case study was conducted to demonstrate the clinical importance of our method on a clinical dataset which consists of 250 patients who have been diagnosed with primary Sjögren’s syndrome (pSS) (3, 4). A reference model was used as a set of clinical-oriented parameters that were defined by the clinical experts as the minimum requirements that are needed to sufficiently describe the clinical domain knowledge of pSS. Our method was able to successfully identify any outliers, unknown data types, and further inconsistencies within the dataset, as well as, successfully match 89.41% of the pSS-related terms according to the pSS reference model. This along with the fast execution and the adequacy it offers, boosts its clinical importance towards data curation and enhances the quality of the clinical data in terms of relevance, completeness, conformity, and accuracy.

Materials and methods

Steps towards clinical data curation

The conceptual methodology for data curation is depicted in Figure 1 (5). The data curation strategy uses as input a pSS dataset and involves the execution of the following steps: (i) metadata extraction, where feature-oriented information is extracted including the range values, the feature labels, the number of instances (patients), and the number of missing values, (ii) data annotation, where the features are classified into discrete and continuous, as well as, into good, fair or bad, depending on the number of missing values, (iii) calculation of descriptive statistics, where various statistical measures are calculated including the mean, median, kurtosis, and skewness, (iv) similarity detection, where the highly-correlated pairs of features are identified, (v) outlier detection, where the values that deviate from the standard distribution are identified, on a feature-basis, using univariate methods, and (vi) data imputation, where the missing values of a specific subset of features are treated using automated means. The workflow’s output is: (i) the curated dataset, where the problematic fields are marked using color coding for easier clinical evaluation, (ii) the data evaluation report, which summarises the extracted metadata and the detected problematic fields per feature, and (iii) the data standardisation report, which summarises the features of the input dataset that where matched with those from the pSS reference model (6).

Data curation mechanisms

This section provides an insight on the methodology that has been deployed to realise the data curation strategy.

• Outlier detection and data imputation

Univariate statistical methods were deployed towards the identification of the outliers, i.e. values that deviate from the standard population distribution. These methods include the z-score and the interquartile range (IQR). The former method: (i) normalises the distance of each sample from the mean with the standard deviation (z-score), and (ii) isolates the absolute z-score values that are larger than 3 times the standard deviation (7). The IQR method calculates the difference between the 1st quartile (Q1) and the 3rd quartile (Q3) of a feature’s distribution, i.e. the interquartile range (IQR), and isolates the values below Q1-1.5*IQR or above Q3+1.5*IQR (7).
Data imputation refers to the process of replacing missing or unknown values within a dataset or simply ignoring instances with unknown values. The data imputation procedure is semi-automated and is applied only on those features that have been characterised as “fair” (i.e. having less than 50% missing values) with the absence of any outliers and/or unknown data types to avoid any further contamination. Standard data imputation methods were employed (8), including: (i) the average/most frequent method, where the missing values of the continuous features are replaced with the average value and the most frequent value in the case of the discrete features, and (ii) the random method, where random values are drawn from the feature’s distribution to replace each missing value.

• **Similarity detection**
  The Spearman correlation coefficient (9) was computed for each pair of features as a rank-order correlation measure to search for highly-correlated pairs of features and a 98% threshold was applied to isolate the most prominent pairs of features. As for the detection of features with the exact same or similar terminologies, the Jaro string similarity score (10) was computed between each pair of features labels. A 98% threshold was applied once more to discriminate the duplicated pairs.

• **Data standardisation**
  This process involves the mapping of the terms of a pSS clinical dataset with those from the pSS reference model.
• The pSS reference model
The pSS reference model was developed in co-operation with the clinical experts and includes a set of pSS-related parameters (including information regarding the range values and data types) that were identified by the experts as the minimum requirements that are needed to describe the clinical domain knowledge of pSS (6). The reference model consists of four main pSS-related categories, namely the (6): (i) “Demographics”, (ii) “Clinical tests”, (iii) “EULAR Sjögren’s syndrome disease activity index (ESSDAI) domain” scores, and (iv) “Past/current therapies”. Each category may include additional sub-categories and parameters. For example, the category “Demographics” includes various parameters, such as, the age at diagnosis, the age at the onset of first symptoms, the gender, and the education level. The main category “Clinical tests” includes different types of SS-related clinical tests which serve as sub-categories, such as, the oral tests (e.g. the unstimulated whole saliva and the oral dryness), the ocular tests (e.g. Schirmer’s test, van Bijsterveld/Rose-Bengal test), and the blood tests (e.g. white blood cell count, number of platelets), salivary gland ultrasonography tests, and biopsies-related parameters (e.g. Tarpley score). The main category “ESSDAI domain” includes sub-categories that are related to the various ESSDAI domains, such as, the pulmonary domain, the muscular domain, the renal domain, the constitutional domain, the cutaneous domain, and the lymphadenopathy and lymphoma domain, among others. Finally, the category “Past/current therapies” include various administered medications, such as, the glucocorticoids, the biological disease-modifying anti-rheumatic drugs (bDMARDs), the concomitant DMARDs (cDMARDs), etc.

• Terminology mapping
The parameters of each category within the reference model are used for the terminology mapping process. The latter is conducted automatically and involves the execution of three steps: (i) terminology extraction, where the labels (terminologies) of the pSS-related features are first extracted, (ii) terminology mapping, where the features that exhibit lexical and conceptual similarities with those from the reference model, i.e. features with common blocks or sequences, are matched, and (iii) classification of the matched terms into the classes of the pre-defined pSS reference model including the proper range values to accomplish standardisation. The Natural Language Toolkit (NLTK) database (11) is used as a medical index repository which includes clinical terms that are used to enhance the accuracy of the terminology mapping process through the identification of homonymous terms (e.g. “gender”, and “sex”).

Results
Case study
We acquired anonymised data from patients that have been diagnosed with primary pSS (10). The anonymised data include 250 patients from the University of Athens (UoA) cohort. 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.

Data quality assessment
An instance of the data quality assessment report is depicted in Figure 2. The report was automatically generated in a conventional .xls format, and consists of: (i) a metadata panel, where useful metadata are presented, and (ii) a quality assessment panel where the diagnostic results are presented. More specifically, the total number of features was equal to 166 and the number of patients was equal to 250. Out of 166 features, 60 features were characterised as discrete and 78 as continuous. The number of unknown features was equal to 28 and the total number of missing values was equal to 44.58%. The value ranges in each feature include the

Fig. 4. An instance of the curated dataset with one outlier and two unknown data types.
minimum and the maximum values that exist on each feature’s space. In the case where the feature has unknown or string data type, the complete set of unique values is presented in the value range. For example, the feature “First visit (year)” has a variable type date in the range “[1983, 2018]”. On the other hand, the feature “comorbidities” has a variable type string and thus all the unique string values are recorded (e.g., “Heart Arrhythmias”, “Hashimoto”). In the same feature, the outlier detection method is not applicable since it has a string data type. The same occurs
for the features “First Symptom”, and “Year of first symptom”. The outlier detection method is also not applicable for the bad features with unknown data types, such as, for: (i) the “Rose-Bengal Stain (0-1)”, which includes an unknown symbol “+” that probably denotes positivity, (ii) the “Positive ocular stain score”, which includes values that are recorded as fractions (e.g. “1/9”), and (iii) the “Dry-mouth-Objective (ml of saliva in 15 min)”, which includes an incompatible value <1.5”.

Three pairs were detected as highly-correlated: (i) {“Raynaud’s phen (0-1)”, “Raynaud”}, (ii) {“Ro/La”, “Antti-Ro (0-1)”), and (iii) {“Date of first biopsy”, “Year of disease diagnosis”}, and one pair was identified as duplicate: (“Rose-Bengal Stain (0-1)”, “Rose-Bengal Stain”).

Curated dataset
An instance of the curated dataset is depicted in Figure 3, where the same incompatible value (i.e. “±φ”) has been detected for the features “wbc baseline (absolute number)”, “Neutrophil number (absolute number)”, “Lymphocyte number (absolute number)”, “PLT (absolute number)” which is the number of platelets, and “ESR” which stands for the erythrocyte sedimentation rate. An abnormal absolute value 43.5 was detected as an outlier for the feature “HGB (absolute number)” which stands for haemoglobin. In this instance, a good feature is also depicted in light blue color, namely the “Oesophagus involvm Ger (0-1)”. The rest of the features are “fair” and thus are depicted in light green color except from the “bad” features “Oesophagus involvm” and “Monocyte number (absolute number)”. A final instance of the same curated dataset is depicted in Fig. 4, where an outlier value 11997 has been detected by the service for the feature “Date of first biopsy” along with an incompatible value “>1” for the feature “FS 1st biopsy”. The former value implicates an erroneously parsed year whereas the former value denotes a value which might be larger than 1 but it is not properly recorded.

Data standardisation
An illustration of the data standardisation process is presented in Figure 6, where the main categories of the reference model are presented along with sub-categories and indicative parameters on the left and the parameters of the input dataset are presented on the right. The parameter “Sex” of the input dataset is matched with the parameter “Gender” of the reference model and classified into the main category “Demographics”, with further information regarding the conversion of its values from (“M”, “F”) to {0, 1}, where “0” stands for male and “1” for female. The parameter “C3 (mg/dL)” is matched with the parameter “C3 (mg/mL)” and classified into the main category “Blood tests” which is a sub-category of the main category “Clinical tests”, with further information regarding the conversion of its measurement units from “mg/dL” into “mg/mL” (pre-defined). The parameter “Shirmer’s test” is matched with the parameter “Schirmer’s test” of the reference model and classified into the class “Ocular tests” which is a sub-category of the main category “Clinical tests”, with additional information regarding the conversion of its values from {0, 1} to {1, 2}, where “1” stands for positive and “2” for negative. The parameter “lymphoma” is matched with the homonymous parameter from the reference model and classified into the sub-category “Lymphadenopathy and lymphoma domain” which belongs to the main category “ESSDAI domain”, with further information regarding the conversion of its values from {0, 1} to {yes, no}.

The data standardisation process was able to successfully link 76 out 85 (89.41%) pSS-related terms from the input dataset. An instance of the data standardisation report is depicted in Figure 6. The first column includes the labels of the features that exist in the input dataset. The second column includes the matched parameters (terms) from the reference model. The third column provides information regarding the type of match (either partial or exact). The fourth column includes the final range that the feature should have according to the reference model and finally, the last column includes the main category where the feature was assigned to (1: Demographic, 2: Clinical test, 3: ESSDAI domain, 4: Therapies). For example, the feature “ANA+” (Fig. 6, row 18) was: (i) partially matched with the parameter “ANA” of the reference model, (ii) classified into the main category “Clinical Test” of the reference model, and (iii) marked for conversion from {0, 1} to {yes, no}. The feature “Rose-Bengal Stain (0-1)” was partially matched with the related parameter of the reference model and classified into the main category “Clinical Test”. However, the final range was not parsed since it includes an unknown value “+” (Fig. 2). The same occurs for the rest of the features with unknown data types (e.g. “MSG 2nd bx Focus Score”).

Discussion
Data quality management constitutes the core of a healthcare data management system. In this work, we have applied a data quality workflow to enhance the quality of pSS data in terms of: (i) relevance, and conformity, by automatically isolating a specific set of pSS-relevant terms according to a reference model, and (ii) completeness and accuracy, by highlighting problematic fields (e.g. outliers, missing values, unknown data types) that are present in the dataset. The proposed method overcomes significant limitations that are present in similar attempts for data curation (13, 14), which do not make use of automated methods for outlier detection and data imputation, and focus only on assessing the quality of the terms that are relevant with those from a gold standard model either manually or semi-automatically, without providing any re-usable reports. Our method was able to automatically identify any problematic fields that were present in the data (28 features with outliers and unknown data types) and isolate 89.41% of the pSS-related terms using a reference model that was developed by the clinical experts, thus reducing the time effort needed for manual data curation. Additional case studies are needed to evaluate the usability of the method and extend the standardisation process to include more pSS-related symptoms.
References