Enhancing medical data quality through data curation: a case study in primary Sjögren's syndrome

V.C. Pezoulas[1], K.D. Kourou[1,2], F. Kalatzis[1], T.P. Exarchos[1,3], A.I. Venetsanopoulou[4], E. Zampeli[5], S. Gandolfo[6], F.N. Skopouli[7], S. De Vita[6], A.G. Tzioufas[4], D.I. Fotiadis[1,8]

[1]Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, Greece; [2]Department of Biological Applications and Technology, University of Ioannina, Greece; [3]Department of Informatics, Ionian University, Corfu, Greece; [4]Department of Pathophysiology, School of Medicine, University of Athens, Greece; [5]Institute for Systemic Autoimmune and Neurological Diseases, Athens, Greece; [6]Clinic of Rheumatology, Department of Medical and Biological Sciences, University of Udine, Italy; [7]Department of Internal Medicine and Clinical Immunology, Euroclinic Hospital, Athens, Greece; [8]Department of Biomedical Research, FORTH-IMBB, Ioannina, Greece.

Vasileios C. Pezoulas, MSc
Konstantina D. Kourou, MSc
Fanis Kalatzis, PhD
Themis P. Exarchos, PhD
Aliki I. Venetsanopoulou, MD
Evi Zampeli, MD
Saviana Gandolfo, MD
Fotini N. Skopouli, MD; PhD
Salvatore De Vita, MD, PhD
Athanasios G. Tzioufas, MD, PhD
Dimitrios I. Fotiadis, PhD

Please address correspondence to:
Prof. Dimitrios I. Fotiadis,
Unit of Medical Technology & Intelligent Information Systems, Dept. of Materials Science and Engineering,
University of Ioannina,
GR-45110 Ioannina, Greece.
E-mail: fotiadis@cc.uoi.gr

Received on March 20, 2019; accepted in revised form on June 18, 2019.

Clin Exp Rheumatol 2019; 37 (Suppl. 118): S90-S96.

Key words: data quality, data curation, data standardisation, primary Sjögren's syndrome

Funding: this project 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.

Competing interests: none declared.

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 the accuracy, validity, completeness, precision, integrity, and relevance of a dataset 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

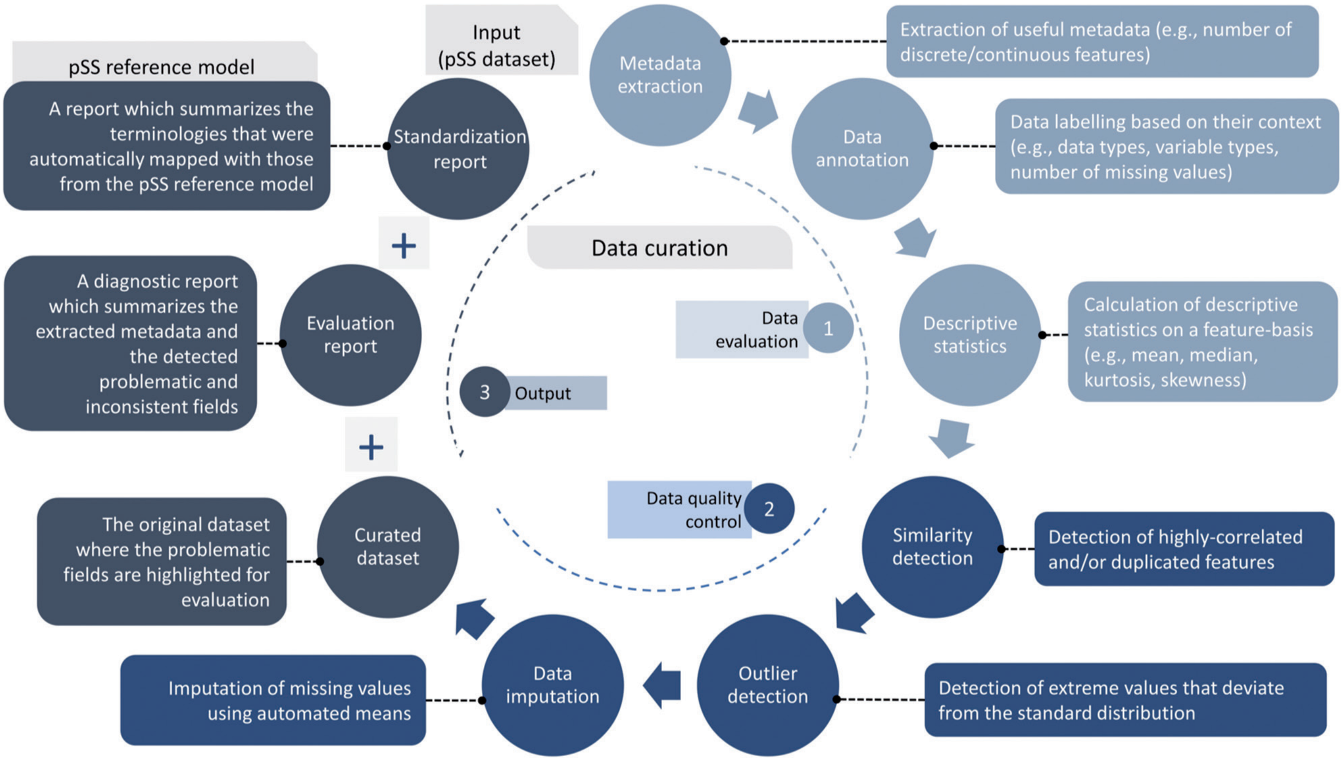

Input
(pSS dataset)
pSS reference model
A report which summarizes the terminologies that were automatically mapped with those from the pSS reference model
Standardization report
Metadata extraction
Extraction of useful metadata (e.g., number of discrete/continuous features)
Data annotation
Data labelling based on their context (e.g., data types, variable types, number of missing values)
Data curation
Data evaluation
1
A diagnostic report which summarizes the extracted metadata and the detected problematic and inconsistent fields
Evaluation report
3
Output
Descriptive statistics
Calculation of descriptive statistics on a feature-basis (e.g., mean, median, kurtosis, skewness)
Data quality control
2
The original dataset where the problematic fields are highlighted for evaluation
Curated dataset
Similarity detection
Detection of highly-correlated and/or duplicated features
Imputation of missing values using automated means
Data imputation
Outlier detection
Detection of extreme values that deviate from the standard distribution

Fig. 1. The steps towards automated data curation.

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

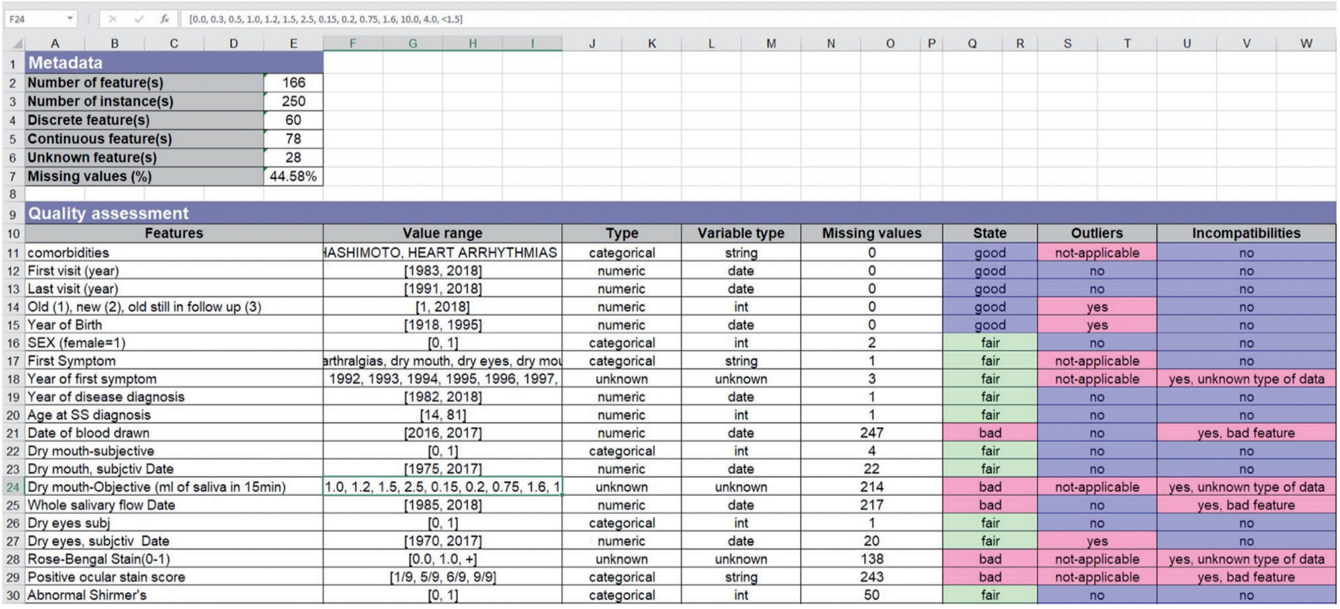
The conceptual methodology for data curation is depicted in Figure 1 (5). The data curation strategy uses as input a pSS dataset and involves the execution of the following steps: (i) metadata extraction, where feature-oriented information is extracted including the range values, the feature labels, 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*

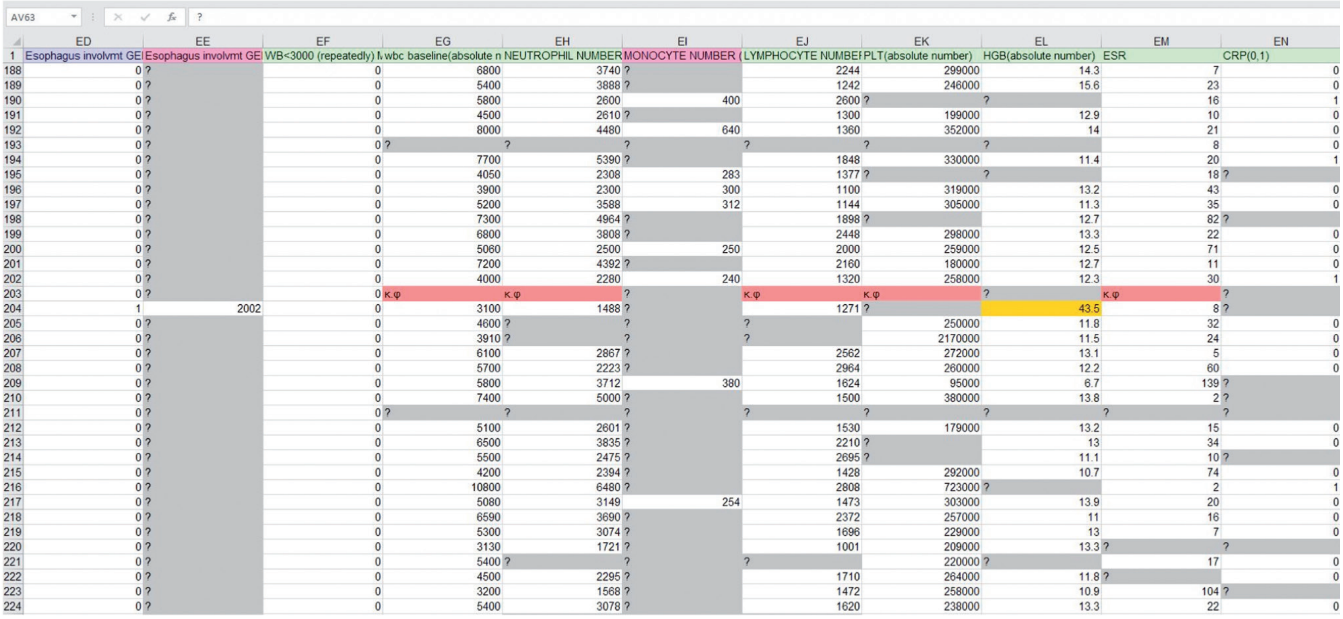
Univariate statistical methods were deployed towards the identification of the outliers, *i.e.* values that deviate from the standard population distribution. These methods include the z-score and the interquartile range (IQR). The former method: (i) normalises the distance of each sample from the mean 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).



F24 [0.0, 0.3, 0.5, 1.0, 1.2, 1.5, 2.5, 0.15, 0.2, 0.75, 1.6, 10.0, 4.0, <1.5]

1	Metadata								
2	Number of feature(s)	166							
3	Number of instance(s)	250							
4	Discrete feature(s)	60							
5	Continuous feature(s)	78							
6	Unknown feature(s)	28							
7	Missing values (%)	44.58%							
8									
9	Quality assessment								
10	Features		Value range	Type	Variable type	Missing values	State	Outliers	Incompatibilities
11	comorbidities		ASHIMOTO, HEART ARRHYTHMIAS	categorical	string	0	good	not-applicable	no
12	First visit (year)		[1983, 2018]	numeric	date	0	good	no	no
13	Last visit (year)		[1991, 2018]	numeric	date	0	good	no	no
14	Old (1), new (2), old still in follow up (3)		[1, 2018]	numeric	int	0	good	yes	no
15	Year of Birth		[1918, 1995]	numeric	date	0	good	yes	no
16	SEX (female=1)		[0, 1]	categorical	int	2	fair	no	no
17	First Symptom		arthralgias, dry mouth, dry eyes, dry mou	categorical	string	1	fair	not-applicable	no
18	Year of first symptom		1992, 1993, 1994, 1995, 1996, 1997,	unknown	unknown	3	fair	not-applicable	yes, unknown type of data
19	Year of disease diagnosis		[1982, 2018]	numeric	date	1	fair	no	no
20	Age at SS diagnosis		[14, 81]	numeric	int	1	fair	no	no
21	Date of blood drawn		[2016, 2017]	numeric	date	247	bad	no	yes, bad feature
22	Dry mouth-subjective		[0, 1]	categorical	int	4	fair	no	no
23	Dry mouth, subjctiv Date		[1975, 2017]	numeric	date	22	fair	no	no
24	Dry mouth-Objective (ml of saliva in 15min)		1.0, 1.2, 1.5, 2.5, 0.15, 0.2, 0.75, 1.6, 1	unknown	unknown	214	bad	not-applicable	yes, unknown type of data
25	Whole salivary flow Date		[1985, 2018]	numeric	date	217	bad	no	yes, bad feature
26	Dry eyes subj		[0, 1]	categorical	int	1	fair	no	no
27	Dry eyes, subjctiv Date		[1970, 2017]	numeric	date	20	fair	yes	no
28	Rose-Bengal Stain(0-1)		[0.0, 1.0, +]	unknown	unknown	138	bad	not-applicable	yes, unknown type of data
29	Positive ocular stain score		[1/9, 5/9, 6/9, 9/9]	categorical	string	243	bad	not-applicable	yes, bad feature
30	Abnormal Shirmer's		[0, 1]	categorical	int	50	fair	no	no

Fig. 2. An instance of the automatically generated data quality assessment report.



AV63 ?

	ED	EE	EF	EG	EH	EI	EJ	EK	EL	EM	EN
1	Esophagus involvmt GE	Esophagus involvmt GE	WB<3000 (repeatedly) N	wbc baseline(absolute n	NEUTROPHIL NUMBER	MONOCYTE NUMBER (	LYMPHOCYTE NUMBEI	PLT(absolute number)	HGB(absolute number)	ESR	CRP(0,1)
188	0	?	0	6800	3740	?	2244	299000	14.3	7	0
189	0	?	0	5400	3888	?	1242	246000	15.6	23	0
190	0	?	0	5800	2600	400	2600	?	?	16	1
191	0	?	0	4500	2610	?	1300	199000	12.9	10	0
192	0	?	0	8000	4480	640	1360	352000	14	21	0
193	0	?	0	?	?	?	?	?	?	8	0
194	0	?	0	7700	5390	?	1848	330000	11.4	20	1
195	0	?	0	4050	2308	283	1377	?	?	18	?
196	0	?	0	3900	2300	300	1100	319000	13.2	43	0
197	0	?	0	5200	3588	312	1144	305000	11.3	35	0
198	0	?	0	7300	4964	?	1898	?	12.7	82	?
199	0	?	0	6800	3808	?	2448	298000	13.3	22	0
200	0	?	0	5060	2500	250	2000	259000	12.5	71	0
201	0	?	0	7200	4392	?	2160	180000	12.7	11	0
202	0	?	0	4000	2280	240	1320	258000	12.3	30	1
203	0	?	0	κ.φ	κ.φ	?	κ.φ	κ.φ	?	κ.φ	?
204	1	2002	0	3100	1488	?	1271	?	43.5	8	?
205	0	?	0	4600	?	?	?	250000	11.8	32	0
206	0	?	0	3910	?	?	?	2170000	11.5	24	0
207	0	?	0	6100	2867	?	2562	272000	13.1	5	0
208	0	?	0	5700	2223	?	2964	260000	12.2	60	0
209	0	?	0	5800	3712	380	1624	95000	6.7	139	?
210	0	?	0	7400	5000	?	1500	380000	13.8	2	?
211	0	?	0	?	?	?	?	?	?	?	?
212	0	?	0	5100	2601	?	1530	179000	13.2	15	0
213	0	?	0	6500	3835	?	2210	?	13	34	0
214	0	?	0	5500	2475	?	2695	?	11.1	10	?
215	0	?	0	4200	2394	?	1428	292000	10.7	74	0
216	0	?	0	10800	6480	?	2808	723000	?	2	1
217	0	?	0	5080	3149	254	1473	303000	13.9	20	0
218	0	?	0	6590	3690	?	2372	257000	11	16	0
219	0	?	0	5300	3074	?	1696	229000	13	7	0
220	0	?	0	3130	1721	?	1001	209000	13.3	?	?
221	0	?	0	5400	?	?	?	220000	?	17	0
222	0	?	0	4500	2295	?	1710	264000	11.8	?	0
223	0	?	0	3200	1568	?	1472	258000	10.9	104	?
224	0	?	0	5400	3078	?	1620	238000	13.3	22	0

Fig. 3. An instance of the curated dataset with an outlier and five unknown data types.

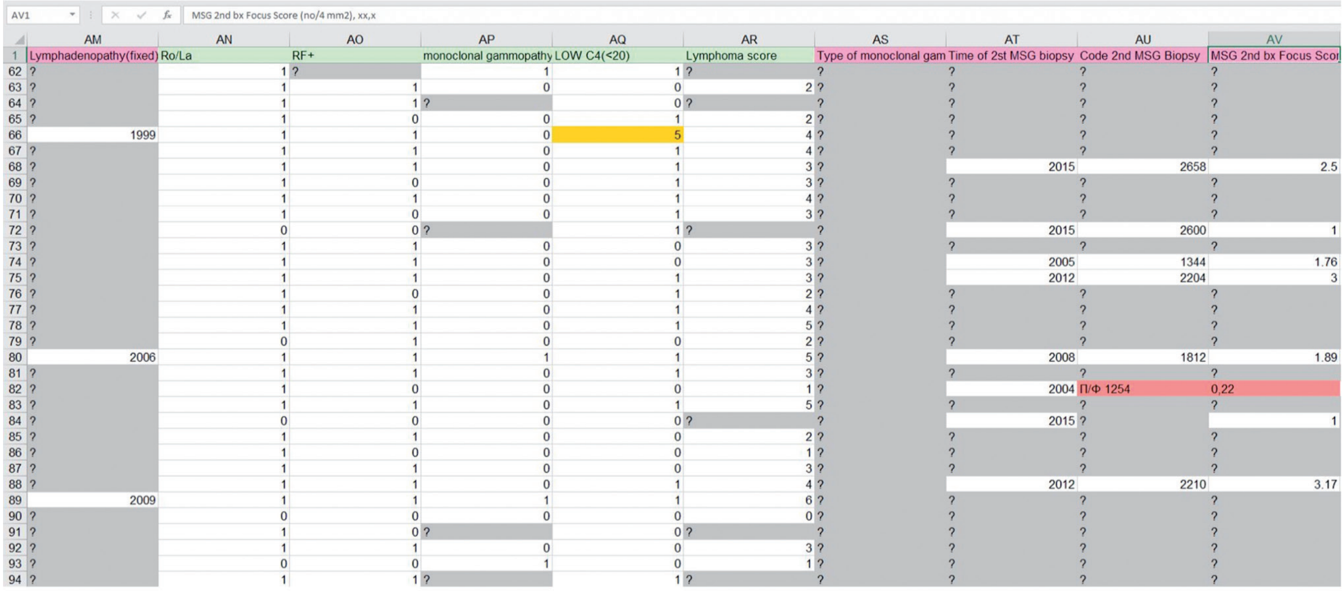
Data imputation refers to the process of replacing missing or unknown values within a dataset or simply ignoring instances with unknown values. The data imputation procedure is semi-automated and is applied only on those features that have been characterised as "fair" (*i.e.* having less than 50% missing 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.



AV1 | MSG 2nd bx Focus Score (no/4 mm2), xx,x

	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV
1	Lymphadenopathy(fixed)	Ro/La	RF+	monoclonal gammopathy	LOW C4(<20)	Lymphoma score	Type of monoclonal gam	Time of 2st MSG biopsy	Code 2nd MSG Biopsy	MSG 2nd bx Focus Scor
62	?	1	?	1	1	?	?	?	?	?
63	?	1	1	0	0	2	?	?	?	?
64	?	1	1	?	0	?	?	?	?	?
65	?	1	0	0	1	2	?	?	?	?
66	1999	1	1	0	5	4	?	?	?	?
67	?	1	1	0	1	4	?	?	?	?
68	?	1	1	0	1	3	?	2015	2658	2.5
69	?	1	0	0	1	3	?	?	?	?
70	?	1	1	0	1	4	?	?	?	?
71	?	1	0	0	1	3	?	?	?	?
72	?	0	0	?	1	?	?	2015	2600	1
73	?	1	1	0	0	3	?	?	?	?
74	?	1	1	0	0	3	?	2005	1344	1.76
75	?	1	1	0	1	3	?	2012	2204	3
76	?	1	0	0	1	2	?	?	?	?
77	?	1	1	0	1	4	?	?	?	?
78	?	1	1	0	1	5	?	?	?	?
79	?	0	1	0	0	2	?	?	?	?
80	2006	1	1	1	1	5	?	2008	1812	1.89
81	?	1	1	0	1	3	?	?	?	?
82	?	1	0	0	0	1	?	2004	П/Φ 1254	0,22
83	?	1	1	0	1	5	?	?	?	?
84	?	0	0	0	0	?	?	2015	?	1
85	?	1	1	0	0	2	?	?	?	?
86	?	1	0	0	0	1	?	?	?	?
87	?	1	1	0	0	3	?	?	?	?
88	?	1	1	0	1	4	?	2012	2210	3.17
89	2009	1	1	1	1	6	?	?	?	?
90	?	0	0	0	0	0	?	?	?	?
91	?	1	0	?	0	?	?	?	?	?
92	?	1	1	0	0	3	?	?	?	?
93	?	0	0	1	0	1	?	?	?	?
94	?	1	1	?	1	?	?	?	?	?

Fig. 4. An instance of the curated dataset with one outlier and two unknown data types.

- *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 antirheumatic 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

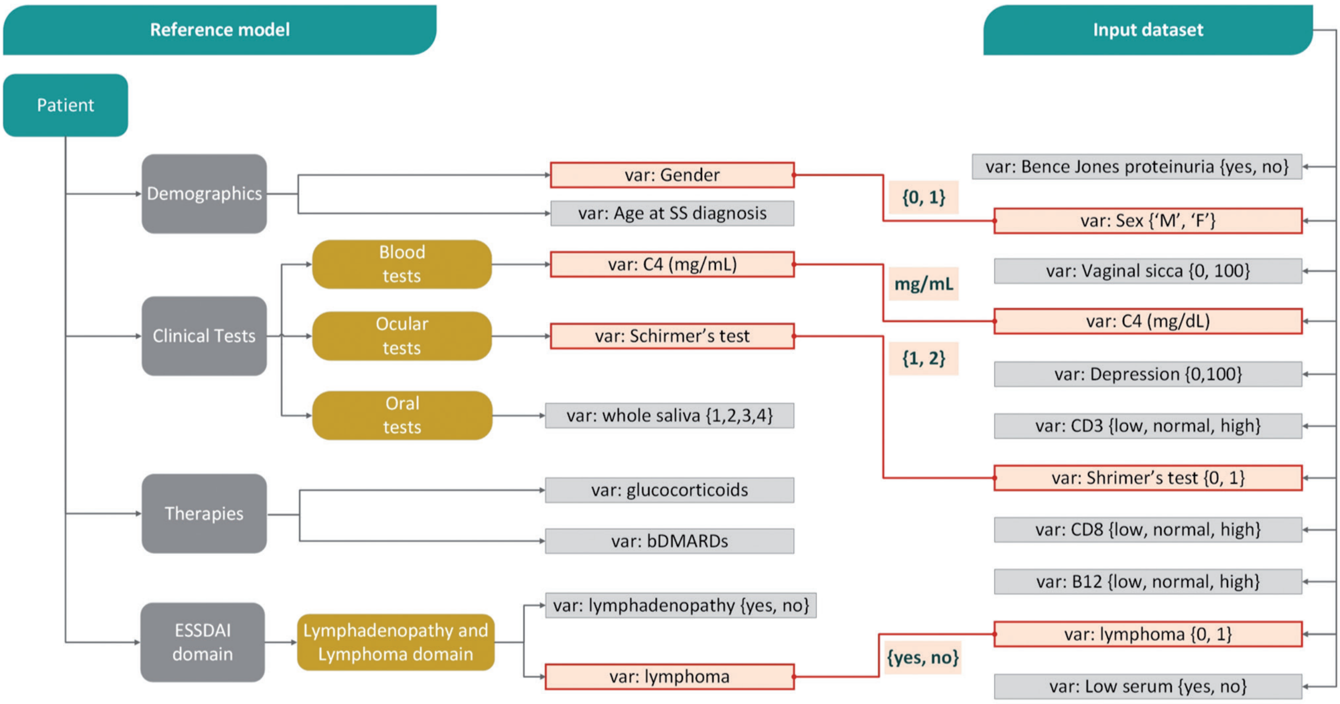
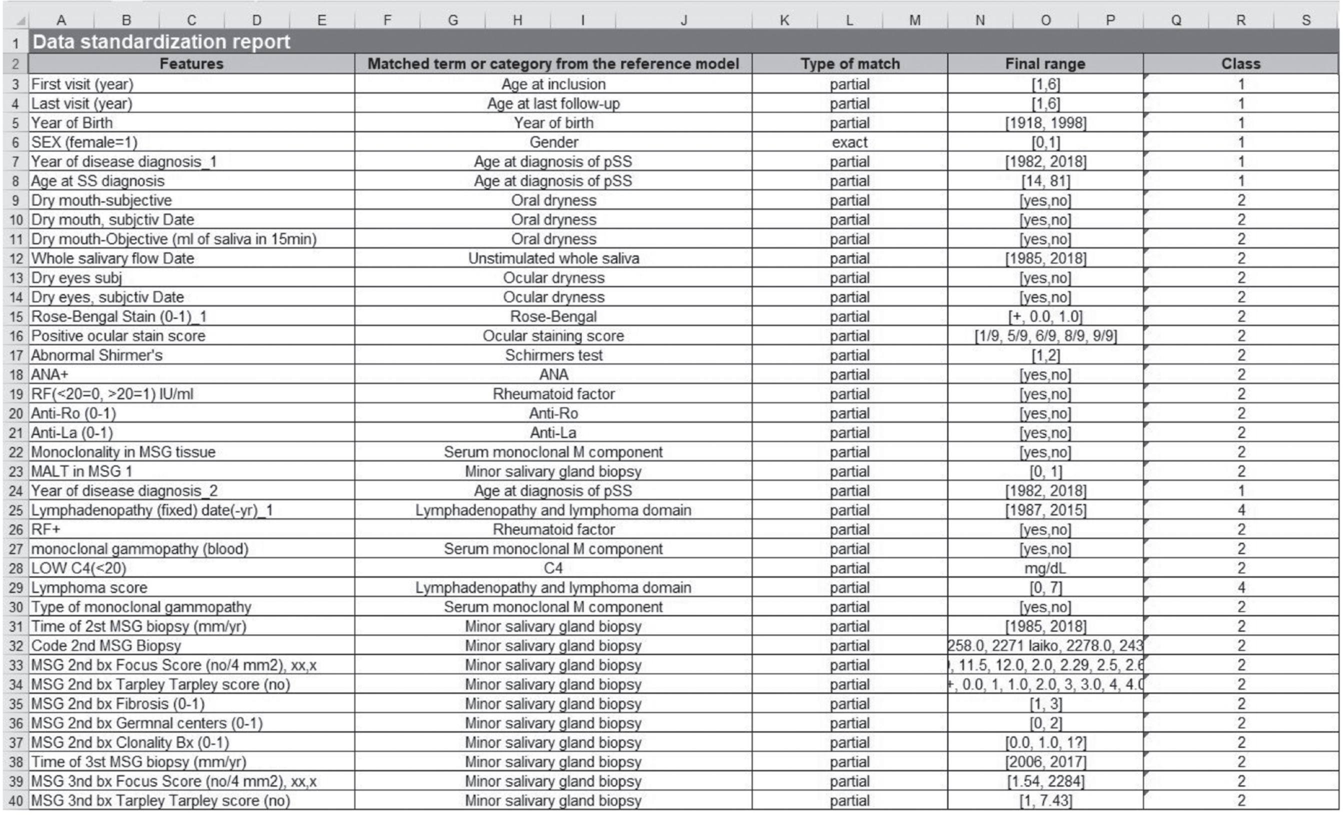


Fig. 5. An illustration of the data standardisation process.



Data standardization report

	Features	Matched term or category from the reference model	Type of match	Final range	Class
3	First visit (year)	Age at inclusion	partial	[1,6]	1
4	Last visit (year)	Age at last follow-up	partial	[1,6]	1
5	Year of Birth	Year of birth	partial	[1918, 1998]	1
6	SEX (female=1)	Gender	exact	[0,1]	1
7	Year of disease diagnosis_1	Age at diagnosis of pSS	partial	[1982, 2018]	1
8	Age at SS diagnosis	Age at diagnosis of pSS	partial	[14, 81]	1
9	Dry mouth-subjective	Oral dryness	partial	[yes,no]	2
10	Dry mouth, subjctiv Date	Oral dryness	partial	[yes,no]	2
11	Dry mouth-Objective (ml of saliva in 15min)	Oral dryness	partial	[yes,no]	2
12	Whole salivary flow Date	Unstimulated whole saliva	partial	[1985, 2018]	2
13	Dry eyes subj	Ocular dryness	partial	[yes,no]	2
14	Dry eyes, subjctiv Date	Ocular dryness	partial	[yes,no]	2
15	Rose-Bengal Stain (0-1)_1	Rose-Bengal	partial	[+, 0.0, 1.0]	2
16	Positive ocular stain score	Ocular staining score	partial	[1/9, 5/9, 6/9, 8/9, 9/9]	2
17	Abnormal Shirmer's	Schirmers test	partial	[1,2]	2
18	ANA+	ANA	partial	[yes,no]	2
19	RF(<20=0, >20=1) IU/ml	Rheumatoid factor	partial	[yes,no]	2
20	Anti-Ro (0-1)	Anti-Ro	partial	[yes,no]	2
21	Anti-La (0-1)	Anti-La	partial	[yes,no]	2
22	Monoclonality in MSG tissue	Serum monoclonal M component	partial	[yes,no]	2
23	MALT in MSG 1	Minor salivary gland biopsy	partial	[0, 1]	2
24	Year of disease diagnosis_2	Age at diagnosis of pSS	partial	[1982, 2018]	1
25	Lymphadenopathy (fixed) date(-yr)_1	Lymphadenopathy and lymphoma domain	partial	[1987, 2015]	4
26	RF+	Rheumatoid factor	partial	[yes,no]	2
27	monoclonal gammopathy (blood)	Serum monoclonal M component	partial	[yes,no]	2
28	LOW C4(<20)	C4	partial	mg/dL	2
29	Lymphoma score	Lymphadenopathy and lymphoma domain	partial	[0, 7]	4
30	Type of monoclonal gammopathy	Serum monoclonal M component	partial	[yes,no]	2
31	Time of 2st MSG biopsy (mm/yr)	Minor salivary gland biopsy	partial	[1985, 2018]	2
32	Code 2nd MSG Biopsy	Minor salivary gland biopsy	partial	258.0, 2271 laiko, 2278.0, 243	2
33	MSG 2nd bx Focus Score (no/4 mm2), xx,x	Minor salivary gland biopsy	partial	, 11.5, 12.0, 2.0, 2.29, 2.5, 2.6	2
34	MSG 2nd bx Tarpley Tarpley score (no)	Minor salivary gland biopsy	partial	+, 0.0, 1, 1.0, 2.0, 3, 3.0, 4, 4.0	2
35	MSG 2nd bx Fibrosis (0-1)	Minor salivary gland biopsy	partial	[1, 3]	2
36	MSG 2nd bx Germnal centers (0-1)	Minor salivary gland biopsy	partial	[0, 2]	2
37	MSG 2nd bx Clonality Bx (0-1)	Minor salivary gland biopsy	partial	[0.0, 1.0, 1?]	2
38	Time of 3st MSG biopsy (mm/yr)	Minor salivary gland biopsy	partial	[2006, 2017]	2
39	MSG 3nd bx Focus Score (no/4 mm2), xx,x	Minor salivary gland biopsy	partial	[1.54, 2284]	2
40	MSG 3nd bx Tarpley Tarpley score (no)	Minor salivary gland biopsy	partial	[1, 7.43]	2

Fig. 6. An instance of the automatically generated data standardisation report.

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”, “Anti-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.* “κ.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 involvmt GER (0-1)”. The rest of the features are “fair” and thus are depicted in light green color except from the “bad” features “Oesophagus involvmt” 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

1. CHEN H, HAILEY D, WANG N, YU P: A review of data quality assessment methods for public health information systems. *Int J Environ Res Public Health* 2014; 11: 5170-5207.
2. CAI L, ZHU Y: The challenges of data quality and data quality assessment in the big data era. *Data Science Journal* 2015; 14.
3. ARGYROPOULOU OD, VALENTINI E, FERRO F *et al.*: One year in review 2018: Sjögren's syndrome. *Clin Exp Rheumatol* 2018; 36 (Suppl. 112): S14-26.
4. KROESE FGM, HAACKE EA, BOMBARDIERI M: The role of salivary gland histopathology in primary Sjögren's syndrome: promises and pitfalls. *Clin Exp Rheumatol* 2018; 36 (Suppl. 112): S222-33.
5. PEZOULAS VC, KOUROU KD, KALATZIS F *et al.*: Medical data quality assessment: on the development of an automated framework for medical data curation. *Comp Biol Med* 2019; 107: 270-83.
6. PEZOULAS VC, EXARCHOS TP, ANDRONIKOU V *et al.*: Towards the establishment of a biomedical ontology for the primary Sjögren's Syndrome. *Conf Proc IEEE Eng Med Biol Soc* 2018; 2018: 4089-92.
7. ROUSSEEUW PJ, HUBERT M: Robust statistics for outlier detection. *In*: Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2011; 1: 73-9.
8. VAN BUUREN S: Flexible imputation of missing data. *Chapman and Hall*, 2018.
9. MYERS L, SIROIS MJ: Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences*, 2004; 12.
10. GANDHI SJ, THAKOR MM, SHETH J, PANDIT HI, PATEL HS: Comparison of String Similarity Algorithms to Measure Lexical Similarity. *J Syst Inform Tech*, 2017; 10: 139-54.
11. SUN S, LUO C, CHEN J: A review of natural language processing techniques for opinion mining systems. *Information Fusion* 2017; 36: 10-25.
12. MAVRAGANI CP, MOUTSOPOULOS HM: Sjögren's syndrome. *Can Med Assoc J* 2014; 186: E579-E586.
13. REIMER AP, MILINOVICH AA, MADIGAN EA: Data quality assessment framework to assess electronic medical record data for use in research. *Int J Med Inform* 2016; 90: 40-7.
14. WEISKOPF NG, HRIPCSAK G, SWAMINATHAN S, WENG C: Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform* 2013; 46: 830-6.