Data Quality Improvement of a Multicenter Clinical Trial Dataset

Data Quality Improvement of a Multicenter Clinical Trial Dataset

Gian Maria Zaccaria, Samanta Rosati, Cristina Castagneri, Simone Ferrero, Marco Ladetto, Mario Boccadoro, and Gabriella Balestra, *Member, IEEE*

***Abstract*— Medical datasets are usually affected by several problems, such as missing values, inconsistencies, redundancies, that can influence the data mining process and the extraction of useful knowledge. For these reasons, a preprocessing phase should be performed for improving the overall quality of data and, consequently, of the information that may be discovered from them. In this study we applied five steps of data preprocessing to improve the quality of a large dataset derived from a multicenter clinical trial. Our dataset included 298 patients enrolled in a prospective, multicenter, clinical trial, characterized by 22 input variables and one class variable (MIPI value). In particular, data coming from different medical centers were firstly integrated to obtain a homogeneous dataset. The latter was normalized to scale all variables into smaller and similar intervals. Then, all missing values were estimated by means of an imputation step. The complete dataset was finally discretized and reduced to remove redundant variables and decrease the amount of data to be managed. The improvement of data quality after each step was evaluated by means of the patients' classification accuracy using the KNN classifier. Our results showed that the proposed pipeline produced an increment of more than 20% of the classification performances. Moreover, the highest growth of accuracy was obtained after missing value imputation, whereas the discretization and feature selection steps allowed for a significant reduction of variables to be managed, without any deterioration of the information contained in data.**

I. Introduction

Clinical trials are clinical researches designed to produce new knowledge about a certain disease, drug or treatment [1] During these studies, a huge amount of data is collected about participants, therapies, clinical procedures, outcomes, adverse events and so on. Therefore, several variables (in some cases higher than the number of participants) are associated to each single patient enrolled in the study and all of them may contribute for a complete patient's assessment.

The analysis of the datasets obtained from clinical trials requires automatic techniques and methodologies able to retrieve informative patterns drown in data and to extract new medical knowledge. The term *Data Mining* (DM) identifies a set of tools for searching for hidden patterns of interest in large and multivariate datasets [2]. The applications of DM techniques in the medical field range from outcome prediction and patient classification [3] to image and signal analysis [4,5]. Furthermore, in recent years, several researches focused the attention on data derived from genomic medicine and molecular biology [6]. Exhaustive reviews of clinical DM applications can be found in [6–8].

Even if it is usually used as a synonym of *Knowledge Discovery in Databases* (KDD), DM constitutes only one step in the complex KDD process that aims to "extracting high-level knowledge from low-level data in the context of large datasets" [9]. In the wider scenario of KDD, a preprocessing phase should be performed before the application of DM tools, for improving the overall quality of data and, consequently, of the information that might be discovered from them [10]. Poor quality of the collected clinical data, in terms of incomplete, incorrect or improper values, can produce detrimental consequences: as an example, incorrect calculation of outcome prediction might lead to an improper medical treatment for the patient [11]. This means that "quality decision must be based on quality data" [12].

Data preprocessing techniques can be grouped into four classes, according to the problem they face with [13]. *Data integration* allows for merging data from multiple sources (for example those derived from multicenter clinical trials) into a homogeneous dataset, reducing inconsistencies. *Data transformation* techniques transform or consolidate data into forms that are appropriate for the DM processing. Two examples of data transformation methods are normalization, in which variables presenting very large and different ranges are scaled into similar intervals, and discretization, where the raw values of a numeric attribute are replaced by interval labels. *Data cleaning* (or *data cleansing*) is usually applied to handle missing values (MVs), remove noise and correct inconsistencies in data. Finally, *data reduction* reduces dataset size by eliminating redundant and irrelevant variables, without any loss of useful information. These four approaches can be applied individually or combined among them in order to solve several issues.

The aim of this study is to improve the data quality of a large dataset related to a prospective, multicenter clinical trial enrolling patients affected by mantle cell lymphoma (MCL).

G. M. Zaccaria, S. Rosati, C. Castagneri, and G. Balestra are with Department of Electronics and Telecommunications, Politecnico di Torino, Torino, Italy (e-mails: gian.zaccaria@polito.it; samanta.rosati@polito.it; cristina.castagneri@polito.it; corresponding author: G. Balestra, email: gabriella.balestra@polito.it, phone: +39-11-090-4136, fax: +39-11-090-4217).

S. Ferrero and M. Boccadoro are with the Department of Molecular Biotechnology and Health Sciences, Università degli Studi di Torino, Torino, Italy (e-mails: simone.ferrero@unito.it; mario.boccadoro@unito.it).

M. Ladetto is with the Division of Hematology, Azienda Ospedaliera SS Antonio e Biagio e Cesare Arrigo, Alessandria, Italy (e-mail: marco.ladetto@ospedale.al.it).

II. Materials and Methods

A. Clinical Trial, Population and Dataset Description

Data used in this study were collected from a phase III, multicenter, open-label, randomized, controlled trial aiming to determine the efficacy and safety of Lenalidomide (Revlimid®) as maintenance therapy versus observation in younger (< 65 years) patients affected by MCL and treated with high-dose immunochemoterapy and autologous stem cell transplantation as first-line therapy (FIL-MCL0208 trial, NCT02354313 [14]). The study was conducted in accordance with the Declaration of Helsinki and all the patients provided written informed consent for the collection and the research usage of clinical and biological data.

Forty-eight Italian medical centers were actively involved in the trial, for a total of 300 enrolled patients (age: 55±8 years). During the clinical trial several variables or features (more than 100) were acquired (derived from the compilation by the clinical centers of the electronic case report forms, eCRFs), describing the patient status at the diagnosis and at different time points. Among this huge number of variables, in this study we focused the attention on those derived from main clinical features, classical laboratory and pathological data, and several molecular data (derived from ancillary, biological investigations) recorded at the diagnosis of the disease. For each patient the corresponding MIPI (MCL international prognostic index) [15] value was assigned. MIPI is a prognostic index of overall survival that groups patients into 3 classes (1: low risk, 2: intermediate risk, 3: high risk) based on four independent clinical variables: age, ECOG (Eastern Cooperative Oncology Group) performance status [16], lactate dehydrogenase (LDH), and leukocyte count (WBC). According to the MIPI, 182 patients were classified as low risk subjects, 73 intermediate risk subjects and 43 high risk subjects. For two patients it was not possible to calculate MIPI, due to missing values in one of the four required variables. As some of the applied algorithms for quality improvement were supervised with the MIPI class, we decided to discard the four clinical variables determining the MIPI, as they could bias the final results. The definitive list of the 22 continuous input variables considered for each patient is reported in the first two columns of Table I. Therefore, our *initial* dataset contained 298 subjects characterized by 22 input variables and one class variable (MIPI value).

B. Data Quality Improvement

From a preliminary analysis of the collected data from the eCRF several problems were observed. Firstly, one variable (ß2) was measured using several devices having different thresholds for discriminating between normal and altered values. Then, the considered input features were associated to clinical parameters, with values ranges that are different one to the other. Moreover, many input values were missing. Finally, we might hypothesize that some of the acquired input variables can be irrelevant or redundant in impacting the patients' outcome. Consequently, we proposed a process for data quality improvement based on 5 steps, each aiming to solve one of the problems previously described.

- *Step 1: Data Integration*

Data integration is usually necessary for data coming from multiple sources, each collecting parameters with different orders of magnitude, units of measurement, or ranges of validity. In our case, the input variable ß2 is a laboratory measurement whose values were obtained with different devices according to the medical center responsible for the measurement. This means that the threshold for discriminating between normal and altered values (ß2_*threshold*) was different according to the center and it was provided together with the patient's ß2 value. In order to obtain a homogeneous ß2 variable, we redefined each value (ß2_*value*) as: *(*ß2_*value*-ß2_*threshold)/* ß2_*threshold*. In this way normal ß2 measurements (that means under the threshold) were represented with negative values, whereas positive values were associated to altered ß2 measurements (that means above the threshold).

TABLE I. Input Variables Description

Feature Name	Description	No of MVs	FS Result
Echo	Left ventricular ejection fraction by either bi-dimensional echocardiogram or cardiac scintigraphy [%]	42	
PLTs	Platelets or thrombocytes [10^9 per L]	8	
Hb	Hemoglobin [g/dL]	1	1
ß2M	ß2-microglobulin [mg/L]	60	1
Protein	Total amount of Albumin and Globulins [g/dL]	19	
Albumin	Serum albumin [g/dL]	37	
IgG	Immunoglobulin G [g/L]	78	1
IgA	Immunoglobulin A [g/L]	78	
IgM	Immunoglobulin M [g/L]	78	
AST	Aspartate transaminase [U/L]	14	
ALT	Alanine transaminase [U/L]	6	
gamma-GT	Gamma-glutamyl transferase [U/L]	22	
Alkaline Phosph	Alkaline phosphatase [U/L]	27	
Bilirubin	Bilirubin [mg/dL]	20	
Creatinine	Serum creatinine [mg/dL]	6	
qnt BM	Quantitative evaluation of the molecular tumor marker in the diagnostic bone marrow sample	161	1
qnt PB	Quantitative evaluation of the molecular tumor marker in the diagnostic peripheral blood sample	154	1
Ki-67	Antigen Ki-67 expression on diagnostic sample [%]	27	
Flow BM	Quantitative evaluation of the tumor invasion by Flow cytometric immune-phenotyping analysis in the diagnostic bone marrow sample	17	1
Flow PB	Quantitative evaluation of the tumor invasion by Flow cytometric immune-phenotyping analysis in the diagnostic peripheral blood sample	48	
BM Infiltration	Quantitative evaluation of the tumor invasion by morphologic and immunochemistry analysis in the diagnostic bone marrow biopsy sample [%]	71	
IgH Omology	Omology of the monoclonal immunoglobulin heavy chain gene rearrangement to the germline sequence [%]	86	

- *Step 2: Data Transformation – Normalization*

Normalization is required for variables sets presenting very different ranges, above all when DM tools involving distance metrics are going to be used subsequently. In our dataset, each input feature was related to a different clinical parameter with its own range of admissible values. To obtain comparable variability intervals, we normalized every variable using the min-max scaling: *(Var_value-Var_min)/(Var_max-Var_min),* where *Var_min* and *Var_max* are the minimum and the maximum values assumed for that variable, respectively. In this way, values between 0 and 1 were obtained.

- *Step 3: Data Cleaning – Missing Values Imputation*

Handling of MVs is one of the most recurrent problems in medical datasets, due to missing manual input or examination results. As in our case the amount of MVs was very high for every input variable, it was not possible to follow the usually used approach of discarding subjects containing missing elements. Therefore, given a patient with a certain MIPI score and a MV for a specific input variable, we replaced this element with the median of the feature values calculated across all subjects within the same MIPI class.

- *Step 4: Data Transformation – Discretization*

In this study, a discretization step was introduced to transform continuous variables into discrete features, partitioning each range of values into a set of adjacent intervals. In general, three are the aims of discretization: to reduce noise due to small variations of values, to decrease the amount of values to be memorized and managed, and to improve the classification performances [17]. Moreover, in this case it was a mandatory phase for applying the tool for data reduction in the following step. The ChiMerge algorithm [18] was chosen and implemented in this study for discretization. It is a supervised and bottom-up method that discretizes each variable separately using the $\chi 2$ statistic. It iteratively merges adjacent elements until the $\chi 2$ value exceeds a defined threshold. In this work, the threshold is determined as the $\chi 2$ value for a significance level of 0.95 and a number of degrees of freedom equal to the number of MIPI classes minus one, that is 2.

- *Step 5: Data Reduction – Feature Selection*

Feature selection (FS) identifies a group of methods for dimensionality reduction able to choose a subset of the original variables without any reduction of the amount of information contained in data. This is possible because only relevant and non-redundant features are selected according to some criteria, producing an increase of the learning accuracy and of result comprehensibility. The *QuickReduct Algorithm* (QRA) [19] was used in this study to select the most important features. It is a supervised tool based on the Rough Set Theory that allows for solving FS problems without generating all the possible subsets. QRA uses the *dependency degree* $\gamma_R(D)$ value to measure the importance of a given subset of input features *R* with respect to the class attribute *D*. The main idea of the algorithm is to iteratively add to the actual features subset those attributes producing the largest increase in the dependency degree.

C. Validation

To validate the proposed pipeline and to prove the improvement of data quality, we assess the capability of the datasets obtained after each step to correctly classify the subjects involved in the study in the pre-existing MIPI risk class. Moreover, we compared these results with the classification accuracy reached by the initial raw dataset.

The k-nearest neighbor (KNN) algorithm was used for the classification, as it is less influenced by the class imbalance than other classifiers. We have tested k values from 3 to 10 and different distance metrics such as the Euclidean, the Chebyshev and the Manhattan distances. However, as the purpose of this study is not to classify the patients but only to measure the quality variation, here we reported the best results reached with k=7 and the Manhattan distance, even if similar results were obtained with the other parameters. The leave-one-out validation was employed for assessing the classification performances.

III. Results and Discussion

Steps 1 and 2 of the proposed methodology allowed for obtaining a dataset in which all input variables showed homogeneous variability ranges. The number of MVs for each feature is reported in the third column of Table I. All these elements were imputed during step 3 and the complete dataset was used for the last two steps of the data quality improvement process. Once the dataset has been discretized in step 4, the FS algorithm was applied and the results in terms of selected features are showed in the last column of Table I (where "1" identifies a selected feature). As it emerges from the table, only six variables were selected in this phase: Hb, ß2M, IgG, qnt BM, qnt PB, Flow BM.

Table II shows the results of the validation phase in terms of classification performances, using the initial dataset and after each step of the applied methodology. The percentage of subjects that are correctly, incorrectly or not classified (no majority of votes for a class is reached) by the KNN is presented. Analyzing the second column of Table II it can be observed that, even if the first two steps do not produce any significant variation of the performances with respect to the initial dataset, the MVs imputation (step 3) allows for increasing the percentage of correct classification of about 20% with respect to step 2. Moreover, even if the variables discretization (step 4) slightly reduce the performances with respect to step 3, the FS method (step 5) produces a further improvement of the classification accuracy, meaning that the discarded variables represented a source of noise for the

TABLE II. Validation Results

	Patients correctly classified	**Patients incorrectly classified**	**Patients not classified**
Initial dataset	57.6%	31.9%	10.5%
Step 1	57.2%	31.6%	11.2%
Step 2	59.5%	32.2%	8.2%
Step 3	79.3%	14.8%	5.9%
Step 4	77.6%	16.8%	5.6%
Step 5	81.9%	13.2%	4.9%

TABLE III. FINAL CONFUSION MATRIX

		Predicted Class			
		Not Classified	**Low Risk**	**Intermediate Risk**	**High Risk**
Actual Class	**Low Risk**	2.2%	92.3%	1.6%	3.8%
	Intermediate Risk	2.7%	21.9%	65.8%	9.6%
	High Risk	7.0%	7.0%	9.3%	76.7%

identification of the patients' risk class. An opposite behavior can be observed analyzing the number of incorrectly and not classified patients (last two columns of Table II). These percentages are considerably reduced after MVs imputations, meaning that the missing information is necessary for a correct identification of the patient risk class. Furthermore, the last two steps don't produce a significant deterioration of the performances.

In Table III we report the confusion matrix obtained at the end of the data quality improvement process. Each percentage is calculated with respect to the total number of subjects belonging to a specific MIPI class. As it emerges from the table, the highest accuracy is obtained for the low risk class (92.83%), that is the largest group of subjects. Although the KNN slightly feels the effects of the class imbalance, we can suppose that this class should influence the classifier training and, as a consequence, the classification accuracy. The lowest performance is returned for the intermediate risk class (65.8%). From the MIPI point of view, this class is assigned to patients obtaining a score between 5.7 and 6.2, which is a very tight interval. This can lead to a probable misclassification of borderline subjects that can affect also the classifier accuracy.

IV. CONCLUSION

In this study we focused the attention on the phase of data preprocessing with the aim of improving the quality of data and of the knowledge retrieved from them. We applied a five-step pipeline for increasing the quality of a dataset obtained from a multicenter clinical trial on MCL patients. We evaluate the quality improvement obtained after each step in terms of patients' classification accuracy.

Our results showed that the imputation of MVs is the phase that mostly produces an increment of the data quality and classification performances. Moreover, the data discretization and the FS allow for a significant reduction of the amount of elements and variables to be managed, without any deterioration of the information contained in data. Finally, it is important to highlight that the FS algorithm used in this study is supervised with the class variable. This means that the six selected variables, that are not used for the MIPI calculation, are the most important for discriminate patients according to their MIPI value, as it was confirmed also by the classification results.

ACKNOWLEDGMENT

Authors would like to thank the FIL - Fondazione Italiana Linfomi (Italian Lymphoma Foundation), the "Centro di Prevenzione Oncologica, CPO" of the AOU "Città della Salute e della Scienza di Torino" and the Molecular Biology Laboratory of the Hematology Division of the University of Turin for sharing the data of the FIL-MCL0208 clinical trial.

REFERENCES

[1] J. Gholap, V. P. Janeja, Y. Yesha, R. Chintalapati, H. Marwaha, and K. Modi, "Collaborative data mining for clinical trial analytics," in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2015, pp. 1063–1069.

[2] M. F. Usama, P.-S. Gregory, S. Padhraic, and U. Ramasamy, Eds., "Advances in knowledge discovery and data mining," *Computers & Mathematics with Applications*, vol. 32, no. 10. American Association for Artificial Intelligence, Menlo Park (California), p. 128, 1996.

[3] G. Fiscon *et al.*, "Alzheimer's disease patients classification through EEG signals processing," in *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2014, pp. 105–112.

[4] S. Rosati, K. M. Meiburger, G. Balestra, U. R. Acharya, and F. Molinari, "Carotid wall measurement and assessment based on pixel-based and local texture descriptors," *J. Mech. Med. Biol.*, 2016.

[5] S. Rosati, G. Balestra, and F. Molinari, "Feature Extraction by Quick Reduction Algorithm: Assessing the Neurovascular Pattern of Migraine Sufferers from NIRS Signals," in *Machine Learning in Healthcare Informatics*, vol. 56, S. Dua, U. R. Acharya, and P. Dua, Eds. Springer Berlin Heidelberg, 2014, pp. 287–307.

[6] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," *Int. J. Med. Inform.*, vol. 77, no. 2, pp. 81–97, 2008.

[7] S. GraciaJacob and R. Geetha Ramani, "Data Mining in Clinical Data Sets: A Review," *Int. J. Appl. Inf. Syst.*, vol. 4, no. 6, pp. 15–26, Dec. 2012.

[8] J. Iavindrasana, G. Cohen, A. Depeursinge, H. Müller, R. Meyer, and A. Geissbuhler, "Clinical data mining: a review.," *Yearb. Med. Inform.*, pp. 121–33, 2009.

[9] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *Am. Assoc. Artif. Intell. AI Mag.*, vol. 17, pp. 37–54, 1996.

[10] R. Jensen and Q. Shen, *Computational Intelligence and Feature Selection*. Hoboken, NJ: Wiley-IEEE Press, 2008.

[11] Y. UshaRani and P. Sammulal, "A novel approach for imputation of missing values for mining medical datasets," in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2015, pp. 1–8.

[12] M. Halkidi and M. Vazirgiannis, "Quality Assessment Approaches in Data Mining," in *Data Mining and Knowledge Discovery Handbook*, New York: Springer-Verlag, 2005, pp. 661–696.

[13] J. Han, M. Kamber, J. Pei, J. Han, M. Kamber, and J. Pei, "3 – Data Preprocessing," in *Data Mining*, 2012, pp. 83–124.

[14] S. Cortelazzo *et al.*, "High dose sequentialchemotherapy with rituximab and asct as first line therapy in adult mcl patients: clinical and molecularresponse of the mcl0208 trial, a fil study," *Haematologica*, vol. 100, no. s1, pp. 3–4, 2015.

[15] E. Hoster *et al.*, "A new prognostic index (MIPI) for patients with advanced-stage mantle cell lymphoma.," *Blood*, vol. 111, no. 2, pp. 558–65, Jan. 2008.

[16] M. M. Oken *et al.*, "Toxicity and response criteria of the Eastern Cooperative Oncology Group.," *Am. J. Clin. Oncol.*, vol. 5, no. 6, pp. 649–55, Dec. 1982.

[17] S. Rosati, G. Balestra, V. Giannini, S. Mazzetti, F. Russo, and D. Regge, "ChiMerge discretization method: Impact on a computer aided diagnosis system for prostate cancer in MRI," in *2015 IEEE International Symposium on Medical Measurements and Applications (MeMeA) Proceedings*, 2015, pp. 297–302.

[18] R. Kerber, "Chimerge: Discretization of numeric attributes," *Proceedings of the tenth national conference on Artificial intelligence*. AAAI Press, San Jose, California, pp. 123–128, 1992.

[19] Q. Shen and A. Chouchoulas, "Modular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems," *Eng. Appl. Artif. Intell.*, vol. 13, no. 3, pp. 263–278, 2000.