Je-LKS
Journal of e-Learning and Knowledge Society
The Italian e-Learning Association Journal
PEER REVIEWED PAPERS
NEW TRENDS, CHALLENGES AND PERSPECTIVES ON HEALTHCARE COGNITIVE COMPUTING: FROM INFORMATION EXTRACTION TO HEALTHCARE ANALYTICS
Vol. 14, n.1, 2018
ISSN: 1826-6223 | eISSN: 1971-8829

PRELIMINARY DATA ANALYSIS IN HEALTHCARE MULTICENTRIC DATA MINING: A PRIVACY-PRESERVING DISTRIBUTED APPROACH

Andrea Damiani[1], Carlotta Masciocchi[2], Luca Boldrini[2], Roberto Gatta[2], Nicola Dinapoli[1], Jacopo Lenkowicz[2], Giuditta Chiloiro[2], Maria Antonietta Gambacorta[2], Luca Tagliaferri[1], Rosa Autorino[1], Monica Maria Pagliara[3], Maria Antonietta Blasi[3], Johan van Soest[4], Andre Dekker[4], Vincenzo Valentini[2]

[1] Polo Scienze Oncologiche ed Ematologiche, Università Cattolica del Sacro Cuore - Fondazione Policlinico Universitario Agostino Gemelli, Roma - Italy
[2] Polo Scienze Oncologiche ed Ematologiche, Istituto di Radiologia, Università Cattolica del Sacro Cuore - Fondazione Policlinico Universitario Agostino Gemelli, Roma - Italy
[3] Polo Scienze dell'invecchiamento, neurologiche, ortopediche e della testa-collo, Istituto di Oftalmologia, Università Cattolica del Sacro Cuore, Fondazione Policlinico Universitario Agostino Gemelli, Roma - Italy
[4] Maastricht University Medical Center, Radiation Oncology MAASTRO-GROW School for Oncology and Development Biology, Maastricht, Netherlands

Keywords: distributed learning, distributed preliminary analysis, privacy-preserving, healthcare, data mining

The new era of cognitive health care systems offers a large amount of patient data that can be used to develop prediction models and clinical decision support systems. In this frame, the multi-institutional approach is strongly encouraged in order to reach more numerous samples for data mining and more reliable statistics. For these purposes, shared ontologies

for citations:
Damiani A. *et al.* (2018), *Preliminary Data Analysis in Healthcare Multicentric Data Mining: a Privacy-preserving Distributed Approach,* Journal of e-Learning and Knowledge Society, v.14, n.1, 71-81. ISSN: 1826-6223, e-ISSN:1971-8829
DOI: 10.20368/1971-8829/1454

need to be developed for data management to ensure database semantic coherence in accordance with the various centers' ethical and legal policies. Therefore, we propose a privacy-preserving distributed approach as a preliminary data analysis tool to identify possible compliance issues and heterogeneity from the agreed multi-institutional research protocol before training a clinical prediction model. This kind of preliminary analysis appeared fast and reliable and its results corresponded to those obtained using the traditional centralized approach. A real time interactive dashboard has also been presented to show analysis results and make the workflow swifter and easier.

1 Introduction

In the new era of cognitive health care systems, a massive amount of previously unavailable clinical variables is available for each patient and needs to be managed (e.g. by Electronic Health Records, clinical research, pathology reports, medical reports etc.) (Patel & Kannampallil, 2014) . These data can be successfully used to develop prediction models in order to produce decision support systems for clinicians (Lambin, *et al.,* 2016). However, even if a high number of covariates represents an opportunity to investigate new relations, it also poses new challenges, starting with the higher number of patient records needed in order to achieve an adequate level of statistical significance and to enable researchers to perform model validation (Lambin, *et al.,* 2013). A sufficient number of patient records is usually available only via multi-institutional data sharing. With this approach, datasets coming from different institutions are sent to a central repository and consolidated into a single database. In order to achieve this goal, data sharing is performed by a standardized data collection system, with a shared terminological system ensuring semantic coherence (Meldolesi, *et al.,* 2014), in a privacy-preserving environment, thus achieving both usability and safety of health data, in accordance with ethical and legal requirements by local and international regulations, as in the U.S. (Korn, 2002) and in the EU (Carey, 2009). Several techniques have been proposed, such as encryption for data anonymization (Gkoulalas-Divanis *et al.,* 2014), randomization methods or k-anonymity models and l-diversity (Aggarwal & Philip, 2008). Unfortunately, such methods reduce the granularity of representations in order to increase the privacy preservation of data (Aggarwal & Philip, 2008).

Distributed Learning (DL) techniques may be a good solution to privacy-related issues in performing data analytics through the use of multi-institutional big data: they preserve patients' privacy and data ownership in training prediction models by leaving all data within the originating institutions. This approach, under some conditions, obtains the same results as the classical centralized approach (Deist, *et al.,* 2017). Boyd *et al.* (Boyd *et al.*, 2011) developed a significant class of algorithms implementing a distributed method for the support vector machine, LASSO and logistic regression using the Alternating Direction Me-

thod of Multipliers (ADMM). In practice, several applications on clinical data analytics (Lu *et al.*, 2015; Jochems *et al.*, 2016; Deist *et al.*, 2017; Damiani *et al.,* 2015) have recently been published using this approach; the architecture typically involves two components: site and master. In these applications, patient data were stored at each site and only cumulative statistics were exchanged with the central server (the master). The master then computed the new parameters, which were sent to the single site, and tuned the computation until a convergence criterion was reached (figure 1). This is an example of server-client architectures (Dai, *et al.*, 2018).

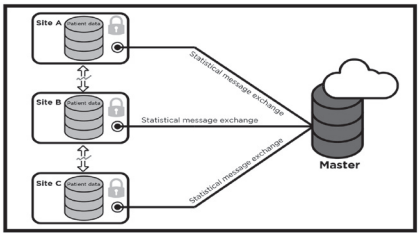

Fig. 1: Statistical message exchanges among sites and master. Patient data are stored at each site and only statistical messages are exchanged between each site and master

The large majority of publications in the DL field focus mainly on learning algorithm development and usually omit all aspects relative to preliminary data analysis. Especially in the health care field, preliminary data analysis is an essential step before training prediction models.

In any multicentric research effort in the field of health care, an initial analysis phase on enrolled patients, in which the researcher inspects available data in order to detect abnormal behaviors, contradictory trends of a given covariate across sites, or peculiar correlations of pairs of covariates (e.g. two continuous covariates exhibit a positive correlation at site A while simultaneously showing a negative correlation at site B) is necessary. These abnormalities, if left undetected, could lead to a reduced quality of the predictive model learned by the main algorithm and ultimately undermine the value of the whole research effort. In a distributed setting, such need is amplified as the researcher can only have direct access to his/her own local dataset. This means that a preliminary analysis step with a distributed approach is very important in the development of data value awareness and data quality enhancement.

1.1 Data issue

The majority of the prediction models proposed in clinical literature are de-

veloped on well-defined cohorts of patients with specific protocol designs (Steyerberg, 2008). In the case of traditional multi-institutional studies, the protocol of the study defines the policies by enrolling a subgroup of patients and is shared among the centers in order to ensure a homogeneous set of clinical cases is recruited. The level of homogeneity is generally verified during the descriptive statistics phase through the application of statistical tests such as Chi-Square, T-test or Mann Whitney test. The choice of the most appropriate statistical test depends on several factors, such as sample number or the type (numerical, ordinal, categorical, etc.) of covariate to be analyzed. This preliminary distribution verification is essential to detect possible bias in patient recruitment by a specific center. Furthermore, it ensures the reliability and reproducibility of the model across new samples coming from the same predefined target population. Some possible aims of investigation for this step are: are one or more covariates distributed in the same way across sites? Shouldn't these two covariates show the same kind of mutual correlation across sites?

The distributed privacy-preserving version of two tools adopted in descriptive statistics are proposed: a distributed version of Chi-Square test and an investigation tool using linear or logistic regression models across sites. These algorithms have been tested on real clinical data in order to analyze covariate distributions and correlations across sites, verifying the homogeneity of the data of a selected protocol and enhancing their quality. During the experiment, these tools were used by an expert in order to detect data anomalies.

2 Material and Methods

2.1 Chi-Square test and linear and logistic regression model

The Chi-Square test is a statistical test which compares the distribution between two binned numeric or categorical variables. Its purpose is to accept or reject a null hypothesis (Ho). This assumption is quantified by a p-value statistic parameter. If the Ho is rejected ($p<0.05$) a statistically significant difference between two covariate distributions is observed. In the traditional approach and in case of multi-institutional data sharing, datasets are sent to a central repository and are centralized into a unique big dataset. In this case, where the data are accessible, each numeric covariate is divided into predefined intervals. In case of categorical or binary variables the interval corresponds to the variable categories. Accordingly with each binning, the number of occurrences in each interval value is calculated. Considering two binned datasets, let S_i be the number of occurrences in bin *i-th* for the first dataset and D_i the number of occurrences in bin *i-th* for the second dataset, the Chi-Square statistics is:

$$X^2=\sum_{i=1}^{k}\frac{\left(S_i*\sqrt{\frac{D}{S}}-D_i*\sqrt{\frac{S}{D}}\right)^2}{\frac{S}{D}} \quad (1)$$

where

$$S=\sum_{i=1}^{k}S_i \qquad D=\sum_{i=1}^{k}D_i \quad (2)$$

And the degrees of freedom (*dof*):

$$dof=(NumberofDataset-1)*(TotalBin-1) \quad (3)$$

Given the (1) and *dof* (3) values, the Chi-Square table distribution or its relative function available in the statistical analysis tool (such as the *pchisq* function available in "R" statistical software) can be used to evaluate the p-value statistic parameter.

2.2 Data access and distributed infrastructure

The three simulated sites collected the clinical data using an in-house software called BOA (Tagliaferri *et al.*, 2016). The aim of this software is an ontology-based standardized data collection able to improve data quality and allow cooperation among different institutions.

To this purpose, data was stored in a PostgreSQL database (version 9.4.1) and a learning connector was used for each simulated sites (Varian Medical Systems, Palo Alto, USA). Registry data (e.g. name and surname) were stored locally into a different database after a de-identification of the Patient's ID in order to increase privacy preservation of data. Through the learning connector, each site then queried the local data using SQL language and exchanged messages with the master.

The connection between each learning connector and the general server was guaranteed by a server-client architecture called Varian Learning Portal (VLP), developed by Varian Medical System company, through which intermediate statistic results were asynchronously exchanged among the master and sites. The researchers interact with the VLP using a web-based interface[1] in which they can upload their distributed algorithms and run simulations. The VLP automatically transmits the single site's algorithm to each other site and the master's to the cloud service. The site algorithm communicates with the learning connector and the master's algorithm, which runs on the VLP, can exchange intermediate statistic results, backwards and forwards with each single simulated site.

[1] (https://www.varianlearningportal.com/VarianLearningPortal/)

2.3 Distributed Chi-Square test and linear and logistic models implementation

As mentioned in section 2.1, the traditional Chi-Square test requires patient data to be accessible. In this section, we propose a distributed Chi-Square test in which patient data never leave the single originating site and health data security is assured. This approach has been used in order to analyze the distribution of the same covariates across each combination of sites.

Each site calculates the occurrences S_i and D_i by accessing its local database and without sharing patient data at each iteration. The statistical parameters of the test (e.g. Chi-Square statistics and p-values) are calculated on the master's side and a result, identical to that obtained using a centralized approach, is generated. Supposing that "M" sites are given, the occurrences S_i and D_i are calculated for each combination of two sites and sent back to the master. The master will then aggregate the statistics received from the sites, calculating the results of the Chi-Square statistic based on equation 1, 2 and 3 and the corresponding p-values. A value lower than 0.05 is considered statistically significant. Details of the proposed distributed Chi-Square test for numeric covariates iteration by iteration are listed in table 1. Regarding binary covariates, only iteration 3 and iteration 4 are used. Correlations between couples of covariates were evaluated either with local logistic or linear models, according to covariate types: linear (for both the numeric covariates) or logistic (for numeric and/or binary covariates) regression models were trained at each site. The beta coefficients and p-value parameters were then sent back to the master. These methods, in addition to distributed 2, could help researchers identify cohort differences among sites.

The proposed code was entirely developed using R version 3.3.1. The results were visualized using a dashboard called Web-based dIstributed statistics REsults (WIRE). WIRE allows the interactive visualization of the distributed descriptive statistic results in real time, offering graphical tools (see figure 2 as an example). The dashboard consists of two parts: site distributed and master distributed statistic results. In the first section, the distributed base statistic results for each site are reported in terms of the number of patients and covariate ranges. In the second section, the distributed base statistic results for the master are visualized in terms of the total number of patients, cumulative covariate ranges, distributed Chi-Square test and local linear and logistic regression models.

WIRE is supported by several browsers: Google Chrome, Mozilla Firefox, Safari and Internet Explorer and its design and development were implemented using the "Shiny" R package which develops interactive web applications simply.

Table 1
MESSAGE EXCHANGES AMONG MASTER AND SITES: ITERATION BY ITERATION

	Message exchanges among master and sites: iteration by iteration
Iteration 1	Each site calculates covariate ranges (in terms of minimum and maximum value for each covariate) and number of patients, then sends the intermediate statistic results to the master.
Iteration 2	The master receives intermediate statistics from each site. Considering each combination of two sites each time (e.g. site A and site B), the master aggregates and sends back to each site 3 values for each covariate: the maximum absolute value, the minimum absolute value and the number of bin calculated as $numBin = \mathrm{int}\left(1 + max\left(\log\left(N_a\right), \log\left(N_b\right)\right)\right)$ where N_a and N_b are the number of patients of site A and site B respectively.
Iteration 3	Each site calculates the predefined interval values by creating a normal distribution using the number of bin, maximum and minimum values received from the master for each covariate and each combination. The number of occurrences S_i and D_i are evaluated for each value of the interval. These values are finally sent back to the master.
Iteration 4	The master receives the occurrences from all sites. For each covariate and each combination the Chi-Square statistics (based on equation 1), S, D (based on equation 2) and dof parameters (based on equation 3) are evaluated. The final p-values are then calculated using R statistical software.

3 Case study

Clinical standardized data from 234 uveal melanoma patients treated with brachytherapy were used for the purpose of our investigation. The inclusion criteria were: dome-shaped melanoma, distance to the Fovea>1.5 mm, tumor thickness>2 mm and follow-up>4 months. Three variables were used in this experience: the presence of diabetes (binary variable: yes versus no), the tumor volume (numeric variable) and the tumor distance to the fovea (numeric variable). The collected dataset was then randomly split into three databases to simulate three different sites (Site A, Site B and Site C). 83 patients were assigned to Site A, 119 to Site B and 32 to Site C. Each dataset was then archived on an independent workstation with a proper learning connector installed to simulate the existence of 3 different institutions and the learning environment described in section 2.2 was recreated. The primary aim of our experiment was to simulate the event in which 3 centers are developing a common predictive model using this infrastructure. The algorithms reported in section 2.3 were applied in order to test the distributed databases' homogeneity before the application of a distributed predictive model. The distributed Chi-Square test (see section 2.3) was applied for each covariate and for each pair of sites (e.g. combination site A-site B; site A-site B and combination site B- site C). The results of such analysis, visualized through the WIRE interface, showed some heterogeneity in terms of distribution of the "volume", "distance to fovea" and "diabetes" covariates as shown in table 2. P-values were calculated for the different combination of sites and appeared to be lower than 0.01 for the "volume" covariate in the combination of

sites A-C B-C, lower than 0.01 for the "distance to fovea" covariate for the combination for sites A-C B-C and for the "diabetes" for all combinations analyzed.

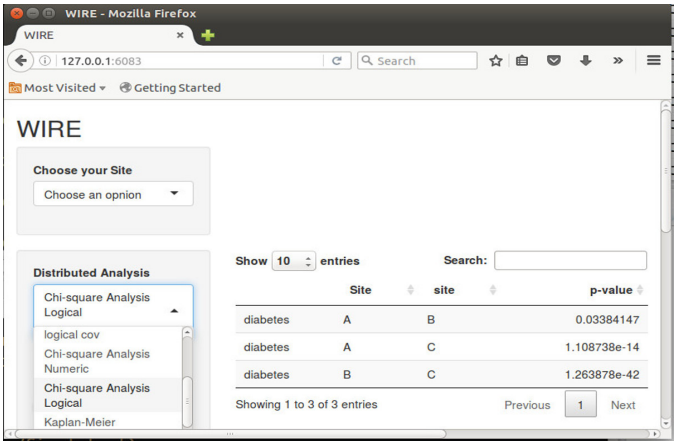

Fig. 2 - Snapshot of WIRE graphical tools. Users can visualize site and master distributed statistic results in real-time in order to check covariate distributions.

Table 2

$\chi 2$ P-VALUE DISTRIBUTED ANALYSIS FOR EACH COVARIATE AND FOR EACH COMBINATION OF SITES ARE REPORTED. THE STATISTICAL SIGNIFICANT VALUES ARE REPORTED AS BOLD VALUES

Combination sites	$\chi 2$ p-value		
	Volume	Distance to Fovea	Diabetes
A-B	0.07	0.2	**0.03**
A-C	**< 0.01**	**< 0.01**	**< 0.01**
B-C	**< 0.01**	**< 0.01**	**< 0.01**

The local linear and logistic regression model was then trained. A statistically linear correlation between "volume" and "distance to fovea" was also observed and more specifically, site B and C showed an inverse correlation when compared to site A (negative slope parameter). These results may or may not represent an alarming signal about the data and suggest that further investigation is needed before proceeding with the use of the combined data. Having observed these differences, we decided to start the aforementioned iterative data processing in order to identify and solve the heterogeneity causes, allowing the researchers to employ previously unusable data. Furthermore, thanks to the help of an expert not involved in the patient enrollment phase who checked the

descriptive statistics of single centers' covariates on the WIRE interface (and therefore without accessing the single databases), we realized that the sites for numerical covariates ("volume" and "distance to fovea") had not fully respected the shared enrollment protocol, including some non-eligible patients. Having removed those patients from the database, we re-ran the tests on a total of 197 patients (site A:68; site B:65; site C:64), obtaining three homogeneous datasets in which no differences in terms of distributions among the single covariates or covariate correlation across the sites were found. Using this last patient subset, we compared the Chi-Square statistics and the observed p-value results through the distributed and centralized approaches. The results are reported in table 3. The difference between the distributed and centralized p-value was less than 10^{-16}. Finally, the response time for the distributed preliminary analysis tool appeared to be very short ($t < 0.05$ s), allowing to support a real time dataset investigation.

Table 3
THE CENTRALIZED AND DECENTRALIZED ALGORITHMS WERE IMPLEMENTED USING R. BOTH EXPERIMENTS WERE PERFORMED ON THE SAME DATASET. RESULTS SHOW THAT BOTH MODELS RESULTED IN IDENTICAL CHI-SQUARE STATISTICS AND P-VALUE COEFFICIENTS.

Features	Distributed chi-square		Centralized chi-square		Combination
	$\chi 2$	p-value	$\chi 2$	p-value	
Volume	1.783	0.878	1.783	0.8782	A-B
Distance to Fovea	5.363	0.373	5.363	0.3731	
Diabetes	0.079	0.778	0.007	0.7784	
Volume	2.741	0.739	2.741	0.739	A-C
Distance to Fovea	3.026	0.695	3.026	0.695	
Diabetes	1.540	0.214	1.540	0.214	
Volume	9.869	0.007	9.869	0.079	B-C
Distance to Fovea	5.694	0.337	5.694	0.337	
Diabetes	1.131	0.287	1.131	0.287	

4 Discussion

The proposed study addresses just one of the potential applications of distributed preliminary analysis on data before training a distributed prediction model. A mathematical approach that applies a strict privacy- preserving policy has been proposed which makes those privacy preservation barriers less difficult to manage. The results were visualized using the WIRE dashboard. It was successfully used by an expert to detect discrepancies compared to the agreed-upon research protocol. The very short running time and the achievement of the same results when compared to the centralized approach suggest that the application of this solution is workable. This approach will greatly facilitate

the collaboration among institutions characterized by different ethical, legal requirements and policies on clinical data management. Some limitations of this approach are: the installation of statistical software R is necessary to run the algorithms and to compile the WIRE interface, which does not allow for running the algorithms manually. These are launched using command line codes and involvement of information technologists is therefore required. The three different sites were only simulated.

Conclusion

DL techniques may be a good solution to privacy-related issues in performing data analytics through the use of multi-institutional big data. Chi-Square test and integrating logistic and regression models were proposed as a necessary step in order to detect data heterogeneity. The technology discussed in this paper allowed clinicians to detect major abnormalities in the covariate distributions across sites, just by looking at the dashboard and without actually accessing the data. In future works, the application of these methods with model development by using real distributed sites will be mandatory.

REFERENCES

Aggarwal C.C., & Philip S.Y.(2008), *A general survey of privacy-preserving data mining models and algorithms*, In Privacy-preserving data mining(pp. 11–52), Boston, Springer.

Boyd S., Parikh N., Chu E., Peleato B., et al.(2011), *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends in Machine Learning, 3(1), 1–122.

Carey P.(2009), *Data protection: a practical guide to uk and eu law*, Oxford, Oxford University Press.

Dai W., Wang S., Xiong H., & Jiang X.(2018), P*rivacy preserving federated big data analysis*, In Guide to big data applications (pp. 49–82), Springer International Publishing AG.

Damiani A., Vallati M., Gatta R., Dinapoli N., et al. (2015), *Distributed learning to protect privacy in multi-centric clinical studies*, In Conference on artificial intelligence in medicine in europe 65– 75.

Deist T.,Jochems A.,van Soest J., Nalbantov G.,et al.(2017) *Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT*, 4, 24-31.

Gkoulalas-Divanis A., Loukides G., & Sun J.(2014). *Publishing data from electronic health records while preserving privacy: A survey of algorithms*, Journal of biomedical informatics, 50, 4–19.

Jochems A., Deist T.M., Van Soest J., Eble M., et al.(2016), *Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital–a real life proof of concept*, Radiotherapy and Oncology, 121(3), 459–467.

Korn D.(2002), *The effect of the new federal medical-privacy rule on research*, The New England journal of medicine, 346(3), 201.

Lambin P., Zindler J., Vanneste B.G.L., Van De Voorde L., et al.(2016), *Decision support systems for personalized and participative radiation oncology,* Advanced Drug Delivery Reviews, 109, 131-153.

Lambin P.,Roelofs E.,Reymena B.,Velazquez E.R.,Buijsen J., et al. (2013), 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy', 109, 159-164.

Lu C.L., Wang S., Ji Z., Wu Y., Xiong L., et al.(2015), *Webdisco: a web service for distributed cox model learning without patient-level data sharing*, Journal of the American Medical Informatics Association, 22(6), 1212–1219.

Meldolesi E.,van Soest J.,Alitto A.R.,Autorino R., et al.(2014), *VATE: VAlidation of high TEchnology based on large database analysis by learning machine*, 3(5), 435-450.

Patel V.L. & Kannampallil T.G.(2014), *Cognitive informatics in biomedicine and healthcare,* 53, 3-14.

Steyerberg E.W.(2008). *Clinical prediction models: a practical approach to development, validation, and updating*. Springer Science & Business Media.

Tagliaferri L., Kovács G., Autorino R., Budrukkar A., et al.(2016), *ENT COBRA(consortium for brachytherapy data analy- sis): interdisciplinary standardized data collection system for head and neck patients treated with interventional radiotherapy(brachytherapy)*, Journal of contemporary brachytherapy, 8(4), 336.