FINAL

Deliverable D1.5

Data mining on the EPIRARE survey data

A. Coi[1], M. Santoro[2], M. Lipucci[2], A.M. Bianucci[1], F. Bianchi[2]

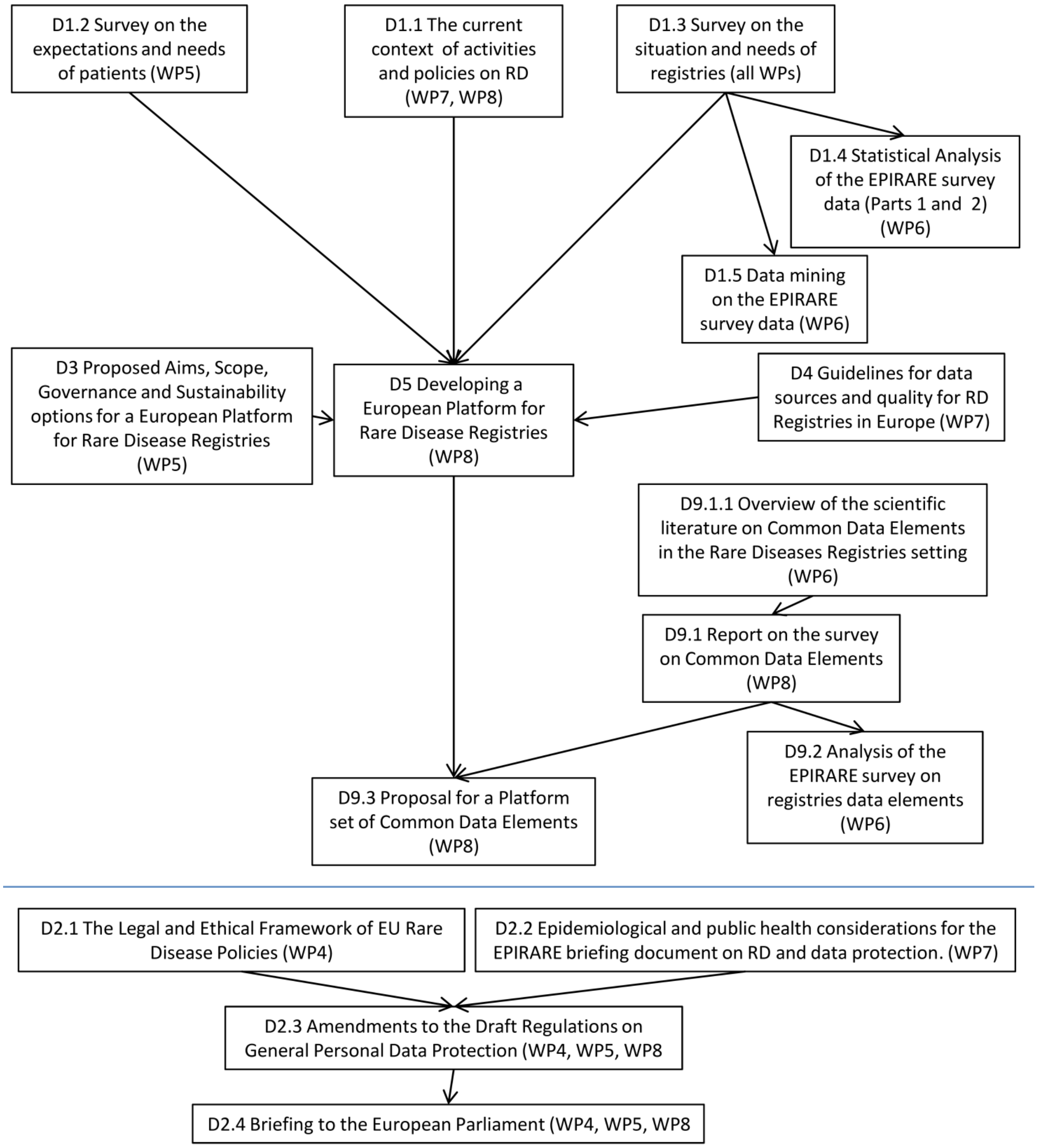
[1] Department of Pharmacy, Unit of Research of Bioinformatic and Computational Chemistry, University of Pisa
[2] Unit of Environmental Epidemiology and Disease Registries, IFC- CNR, Pisa - Epirare WP6 Team

EPIRARE Work Package 6 Team in collaboration with Department of Pharmacy, Unit of Research of Bioinformatic and Computational Chemistry, University of Pisa

CONFIDENTIAL

CONTENTS

Overview of the documents produced by EPIRARE


D1.2 Survey on the expectations and needs of patients (WP5)
D1.1 The current context of activities and policies on RD (WP7, WP8)
D1.3 Survey on the situation and needs of registries (all WPs)
D1.4 Statistical Analysis of the EPIRARE survey data (Parts 1 and 2) (WP6)
D1.5 Data mining on the EPIRARE survey data (WP6)
D3 Proposed Aims, Scope, Governance and Sustainability options for a European Platform for Rare Disease Registries (WP5)
D5 Developing a European Platform for Rare Disease Registries (WP8)
D4 Guidelines for data sources and quality for RD Registries in Europe (WP7)
D9.1.1 Overview of the scientific literature on Common Data Elements in the Rare Diseases Registries setting (WP6)
D9.1 Report on the survey on Common Data Elements (WP8)
D9.2 Analysis of the EPIRARE survey on registries data elements (WP6)
D9.3 Proposal for a Platform set of Common Data Elements (WP8)
D2.1 The Legal and Ethical Framework of EU Rare Disease Policies (WP4)
D2.2 Epidemiological and public health considerations for the EPIRARE briefing document on RD and data protection. (WP7)
D2.3 Amendments to the Draft Regulations on General Personal Data Protection (WP4, WP5, WP8
D2.4 Briefing to the European Parliament (WP4, WP5, WP8

Disclaimer

The contents of this document is in the sole responsibility of the Authors; The Executive Agency for Health and Consumers is not responsible for any use that may be made of the information contained herein.

Introduction

Data mining encompasses a set of techniques that allow information to be extracted from a complex dataset to make it more easily interpretable.

The Epirare Survey represents a classic example of a complex dataset. In fact it consists of a number of samples (220 Registries) characterized by a large number of variables.

Objectives

Exploiting data mining techniques, as part of WP6, met the following specific objectives:

- confirm the clustering hypothesis via the analysis of the variables and the subsequent characterization of the three types of registries;
- identify a subset of variables, selected from those derived from queries of the survey, which are decisive for classifying these registries in accordance with the hypothesis made by the cluster analysis.

Methods

The study was carried out starting from the cluster analysis that allowed the registries to be divided up into three classes: public health, clinical / genetic research, and treatment[1].

The cluster analysis was performed using the query "q9", regarding the stated objectives of the registries, and this variable was removed from the dataset prior to data mining.

The survey was thus transformed into a 219x272 numerical matrix. It was then subjected to data mining mainly with the aim of identifying a subset of variables, selected among the 272 queries derived from the survey, which could then be used to classify such registries in accordance with the hypotheses made by the cluster analysis.

The purpose of this step was to confirm the cluster hypothesis through the analysis of the variables and the subsequent characterization of the three types of registry.

Classification models were thus developed using two variants of the random forest method, namely the traditional algorithm (randomForest)[2] and an evolution of this algorithm (cforest), developed by Hothorn et al.[3-5].

Both these methods enable the importance of variables to be measured against their ability to predict the properties investigated, i.e. the correct classification of each registry into the respective category.

Three experiments were performed, aimed at developing predictive models of classification, using the following methods:

- Traditional randomForest algorithm (out-of-bag - OOB - with resampling), based on rpart
- Cforest algorithm (OOB without resampling), based on cpart
- Cforest algorithm (OOB without resampling) based on cpart with unbiased control.

The predictive power of the models, namely their ability to correctly classify the registries, which from time to time are chosen randomly in the OOB sample, was evaluated using scientifically consolidated statistical criteria, such as sensitivity, accuracy and correct classification rate.

From the model that performed best for each method, information was extrapolated concerning the importance of variables, using a permutation method (variable permutation importance), according to which, a decrease in the accuracy of the prediction of the OOB samples following the permutation of a variable, indicates that variable is important for the correct prediction of the class. In each experiment, a univocal criteria[6] was chosen as the critical threshold for a variable to be considered "important" and informative.

Each experiment involved a tuning phase in which dozens of models were produced by intervening on some critical parameters. In this way the robustness and stability of the results was ensured.

Finally, three models were selected, one for each experiment, which showed an accuracy of 69.4%, 66.2% and 63%, and from which three subsets of 76, 45 and 57 variables were identified from the 272 used.

Three other experiments were then performed, one for each subset (of 76, 45 and 57 variables), with similar procedures to those used in the first step, but using 219x76, 219x45 and 219x57 matrices.

Dozens of models were thereby produced, the best of which showed an accuracy of 73.5%, 69.4% and 69.9%. The improved predictive power (3.5 percentage points on average for the first two experiments, nearly 7 percentage points in the third) confirmed the importance of the selected variables from an informational point of view, thus showing that subsets selected from a starting point of 272 variables are sufficient to obtain models of classification with a higher performance than those obtained using the full set of variables.

Furthermore, the analysis of the three subsets identified a set of intersection of 41 variables that were common to all three experiments. This is a clear indication of the great stability of the models obtained and the resulting reliability of the results.

Analysis of variables

An analysis of the 41 "informative" variables that were common to the three experiments data mining revealed clear differences in the three types of registries. The full list of the 41 variables is reported in Appendix.

In general, registries can be characterized above all on the basis of related variables: the nature, quality and dissemination of the data collected and the patient information sheet.

As regards the nature of the data reported (q16), most treatment registries collect:

- data related to medications, devices and health services,
- anthropometric information,
- data on health status and quality of life of patients
- data on patient death (q17). This data is reported less frequently in the registries of clinical research / genetics and even less in those of public health; in the latter, the date of death of the patient is more often reported with links to other data sources.

The collection of data on births is far more frequent in the treatment registries than in the other two types.

Clinical research / genetic registries and treatment registries have some common characteristics; most of them in fact show:

- genetic and clinical data
- data on family history

Socio-demographic information is reported to a greater extent by the public health registries and clinical research / genetics registries.

As regards the quality of the data reported, most public health registries have an automatic control system to avoid errors when entering data (q22). This type of control takes place to a lesser extent in the other two types of registries.

In the patient information sheet, most registries of clinical research / genetic and treatment report (q36):

- specific research objectives
- the patient's right to withdraw their data from the registry.

In addition, the patient information sheet for most of the clinical research / genetics registries is reviewed by the local ethics committee (q37). This also happens in the treatment registries with a significant percentage, albeit lower.

The public health registries differ in terms of the dissemination of data, since most of them inform the public health policy makers of their activities, and collaborate and share data with other registries. Such data sharing is also true of the clinical research / genetic registries, to a considerably lesser extent, and in a little more than a third of the treatment registries.

From an ethical point of view, the clinical research / genetics registries and the treatment registries usually require written consent from the patient for the data to be included in the registry (q35), unless the data are collected anonymously. This takes place less in the public health registries than in the other two types of registry, partly because more data are collected anonymously.

Finally, of the services potentially accessible from the EU platform (q64), most of the clinical research / genetic registries indicate the importance of having the models (e.g. the template for informed consent). On the other hand, most of the treatment registries considers rate easy access to data sources as being important.

For all the types of registries there are standardized criteria for inclusion / exclusion defined for RD cases (q14); this percentage is more marked in the treatment registries than in the public health and clinical research / genetics registries.

Conclusions

The use of data mining based on the creation of forests of decision trees has allowed the “extraction" of information that confirms a division of registries into three categories: public health, clinical research/genetics, and treatment.

The data mining techniques revealed a subset of "informative" variables which were enough to identify a registry in its respective category. The stability and reliability of the results were confirmed using three separate methods that led to the identification of a common subset of variables.

The analysis of these variables allowed the characterization of the three categories of registry, and the characteristics that emerged from the study seem to be in agreement with the nature of the type of the registry. Some variables related to the nature, quality and dissemination of the data reported are sufficient to confirm the result of cluster analysis based on variables that were not included in the data mining study. This aspect in particular may be considered an important indicator of the stability of the result and consequently confirms the need to consider the definition of three types of registries, each of which characterized by data of a different nature, quality and level of diffusion.

In addition to this, the pooled results of the cluster analysis and the models obtained with decision trees emphasize the prospective power of this dual approach: the identification of a subset of variables (the ones used for the cluster analysis and the 41 identified by the data mining) which can be considered as being sufficient for the categorization of registries.

The results of the multivariate statistical analysis and data mining analysis provide a bulk of knowledge useful to generate hypotheses for the definition of the informative needs of registries. These analyses allowed a powerful informative synthesis of the fragmented framework of Registries of rare diseases.

In particular, the use of different exploratory techniques of data analysis improved the ability in identifying three different types of registries, consequently characterized by different informative needs other than a common set of data elements.

A panel discussion by sharing expertises on different types of registries would be useful to define the pathway to build consensus on minimum common data element and specific common data elements for each type of registry. A further feedback by selected registries might help in refining the final proposal for the European platform.

References

1. EPIRARE – Work Package 6 (2012), Results of the Cluster Analysis on data of the EPIRARE Survey, Draft 14/12/2012.

2. A. Liaw, M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18-22.

3. T. Hothorn, P. Buehlmann, S. Dudoit, A. Molinaro, M. Van Der Laan (2006). Survival Ensembles. Biostatistics, 7(3), 355--373.

4. C. Strobl, A.-L. Boulesteix, A. Zeileis, T. Hothorn (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics, 8(25).

5. C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis (2008). Conditional Variable Importance for Random Forests. BMC Bioinformatics, 9(307).

6. C. Strobl, J. Malley, G. Tutz (2009) An Introduction to Recursive Partitioning: Rational, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. Psychological Methods. 14(4), 323-348.

APPENDIX: List of the 41 variables common to the three experiments of data mining

Q6: How many active cases/patients are included in your register?

Q10: Are your data used for pharmacovigilance?

Q11: What is the geographical coverage of the register?

Q12: What is the target population of the register?

Q14: Are standardized inclusion/exclusion criteria defined for the RD cases?

Q15_3: ICDO disease coding system is in use

Q15_5: ICD10 disease coding system is in use

Q16_3: Anthropometric data are collected

Q16_4: Socio-demographic data are collected

Q16_5: Genetic data are collected

Q16_6: Clinical data are collected

Q16_7: Medications, devices and health services data are collected

Q16_8: Patient-reported outcomes (e.g. quality of life data, Health status, etc) are reported

Q16_9: Family history data are reported

Q16_10: Birth and reproductive history data are reported

Q17: Is the date of the patient death collected?

Q19_9: data providers are mortality registers

Q22_2: An internal program automatically checks the type of response (automatic control) in order to avoid data entry mistakes

Q31_5: There is no structured training for new users

Q32: reasons for which register has been established

Q35: Do you ask for patient's written and informed consent to include his/her identifiable data in the register?

Q36_1: Register's general scope is provided on the patient information sheet

Q36_2: Register's specific research objectives are provided on the patient information sheet

Q36_3: The patient information sheet reports that the register is part of a network

Q36_7: The right to withdraw is provided on the patient information sheet

Q36_11: Possibility to be contacted for participating in clinical trials is provided on the patient information sheet

Q37_1: The patient information sheet has been revised by a local ethics committee

Q38: Is the patient's information sheet publicly available and easily accessible?

Q39_2: If a participant decides to withdraw from the register, he/she may withdraw the authorisation to future uses of the data (except already aggregated or published data)

Q39_6: If a participant decides to withdraw from the register, withdrawal is not possible as this is a mandatory public health register

Q44_4_1: The register has governing bodies composed by internal members with the function of Communication with the funding source, health care providers, patients, etc

Q45_1: Data providers are informed of the register activities

Q45_8: Public health policy makers are informed of the register activities

Q48_4: Is the register collaborating or sharing data with other registries, biobanks, centres of expertise?

Q54_2: The register was set up with funding coming from Regional Authority

Q55_2: Regional Authority is funding the register today

Q56: average yearly budget over the past 3 years

Q59_4: The main needs of the register are to motivate data providers

Q61: Do you expect your country to provide public funding for a centralised national register on RD?

Q64_3: Model documents (e.g. Informed consent form) is one of the services that should be offered by a EU platform for registries

Q64_7: Facilitated access to useful data sources is one of the services that should be offered by a EU platform for registries