Characterization of high-quality Rare Disease Registries by using a data mining approach

Coi A. (1), Santoro M. (1), Villaverde-Hueso A. (2), Lipucci M. (3), Gainotti S. (4), Taruscio D. (4), Posada M. (2), Bianchi F. (1)(5)

1) Istituto di Fisiologia Clinica, Consiglio Nazionale delle Ricerche, Pisa, Italy
2) Instituto de Investigación de Enfermedades Raras, Instituto de Salud Carlos III, Madrid, Spain
3) European Organisation for Rare Diseases (EURORDIS), Paris, France
4) Centro Nazionale Malattie Rare, Istituto Superiore di Sanità, Rome, Italy
5) Fondazione Toscana Gabriele Monasterio, Pisa, Italy

Introduction. Rare Diseases Registries (RDRs) are important epidemiological tools for health policy makers and researchers working in the field of low prevalence diseases. The quality of procedures used when a RDR is defined and also during the first steps of their development sets the basis for its success and it is at the same time the best way to guarantee the long-term sustainability. Therefore the quality of RDRs is one of the key questions to be assured and designed during the first steps of their design.
Aim of the study. To provide information useful to characterize high-quality RDRs by using an analytical approach
Methods. At first, a score of quality was defined by choosing a small set of variables derived by the EPIRARE Survey and related to quality assurance, quality control and quality assessment. In a second step, the random forest (RF) method was applied to the Survey data, so that, starting from the entire set of 223 variables, a subset of variables can be identified as the most informative to afford a reliable characterization of different levels of quality. In the third step, the presence of statistically significant associations between each variable identified by RF and the indicator of quality of RDR were checked with a Chi-square or Fisher exact test. Then the Cochran-Armitage test was also carried out to identify the presence of a linear trend.
Results. Out of the 223 variables RF identified a subset of 47 informative variables. Statistically significant associations are identified between 44 variables (out of 47) and the indicator of quality. A significant linear trend is observed for 43 variables, most of them showing a strong evidence ($p<0.001$).
This set of variables was useful to characterize high-quality RDRs that seem to pay much attention to: ethical and legal issues (protocol approved by Ethical Committee), governance (Main Governing Board executing all the main functions and composed by internal members and external experts), communication of the activities (scientific meetings, scientific journals, website), access to data and security, sustainability.
These findings are in line with the results of similar researches on disease registries [1], highlighting that quality is usually associated with a good oversight and governance mechanism and would benefit from a support in organization and management, information technology, epidemiology, and statistics.

References

1. Black N, Barker M, Payne M, Cross sectional survey of multicentre clinical databases in the United Kingdom, BMJ. 2004 Jun 19; 328(7454):1478-82.