Identifying cases of type 2 diabetes from heterogeneous data sources: strategy from the EMIF Project

G. Roberto[1], I. Leal[2], N. Sattar[3], K. Loomis[4], P. Avillach[2], P. Egger[5], R. van Wijngaarden[6], D. Ansell[7], S. Reisberg[8], M. Tammesoo[9,10], H. Alavere[9,10], A. Pasqua[11], L. Pedersen[12], J. Cunningham[13], L. Tramontan[14], M.A. Mayer[15], R. Herings[6], P. Coloma[2], F. Lapi[1], M.C.J.M. Sturkenboom[2], J. van der Lei[2], M. Schuemie[16,17], P. Rijnbeek[2], R. Gini[1]

[1]Agenzia Regionale di Sanità della Toscana, Osservatorio di Epidemiologia, Firenze, Italy;[2]Dept. of Medical Informatics, Erasmus University Medical Center, Rotterdam, Netherlands;[3]British Heart Foundation Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, United Kingdom,[4]Pfizer Worldwide Research and Development, Groton, Connecticut, United States,[5]GlaxoSmithKline, Worldwide Epidemiology GSK, Stockley Park West, Uxbridge, United Kingdom;[6]PHARMO Institute for Drug Outcomes Research, Utrecht, Netherlands,[7]The Health Improvement Network, Cegedim Strategic Data Medical Research Ltd, London, United Kingdom;[8]Quretec, Software Technology and Applications Competence Center, University of Tartu, Tartu, Estonia;[9]Estonian Genome Center, University of Tartu,[10]Tartu University Hospital,[11]Health Search, Italian College of General Practitioners and Primary Care, Firenze, Italy;[12]Dept. of Clinical Epidemiology, Aarhus University Hosptial, Aarhus, Denmark;[13]University of Manchester, Manchester, United Kingdom

[14]Arsenàl.IT Consortium, Veneto's Research Centre for eHealth Innovation, Treviso, Italy;[15]IMIMHospital del Mar Medical Research Institute and Universitat Pompeu Fabra, Barcelona, Spain;[16]Janssen Research & Development, Epidemiology, Titusville, New Jersey, United States;[17]Observational Health Data Sciences and Informatics, New York, New York, United States

Background

The European Medical Information Framework (EMIF) Project is establishing an EU wide information communication technology (ICT) infrastructure (EMIF-Platform) to facilitate the combination of a wide variety of existing health data from different European data sources and perform large epidemiological studies.

Objectives

To establish a set of standard algorithms for the identification of patients with type 2 diabetes (T2DM) across heterogeneous data sources, to describe the data source-tailored combinations of standard algorithms recommended by local experts, and to assess the impact of individual standard algorithms on the population of case identified across different data sources.

Methods

Eight data sources from six different European countries were included: three were primary care data sources (PCDs), three record linkage data networks (RLDs) data networks, and one hospital database (HD) (end date 1 Jan 2012) and one biobank (BD) (end date, 1 Jan 2009). PCDs and RLDs are population-based data sources, while HD and BD contain non-representative samples of the respective geographic catchment area.

A list of standard algorithms (*component algorithms)* for the identification of T2DM from the selected data sources was created. Each component algorithm was based on records from one specific data domain among: diagnoses (DIAG), drug prescription (DRUG), laboratory results (LABVAL) or utilization of diabetes healthcare services (TEST). The Unified Medical Language System (UMLS) was used for semantic harmonization of coding systems: pertinent medical concepts were identified and projected to local terminologies. Local experts chose the preferred combination of components for their data source (*recommended composite algorithm*) and provided a comment as reusable knowledge. Considering subjects 16+, all the person-time available at the index date (1st Jan 2012 for PCDs, RLDs and HD, 1st Jan 2009 for BD) was used in the case-identification algorithm.

In all data sources the total number of cases identified by the recommended composite algorithm was computed as a percentage of the data base population at the index date, and, among those, the percentage of cases identified by each extracted component algorithm was computed.

Results

For this analysis, the EMIF-Platform provided aggregated health data from a total of 12 million European citizens. The total number of cases identified through composite algorithms corresponded to a percentage of the individual data base population that ranged from 4.1% to 7.5% in RLDs, from 6.8% to 8.6% in PCDs, 3.5% in BD and 15.7% in HD. All composite algorithms used at least one DIAG-based component as inclusion criteria, except for one RLD that adopted a strategy based on DRUG only. DIAG-based components used as inclusion criteria contributed to the total number of cases identified for 93-100% in PCDs, 100% in both BD and HD and and 15-73% in RLDs. DRUG-based components identified from 81% to 100% of the respective total case population in RLDs, and from 58% to 83% in PCDs. LABVAL-based algorithms were adopted by only one PCDs, where they retrieved 46,5% of all cases. One RLD decided to identify T2DM patients using TEST-based algorithms which identified less than 44.1%. Since patients could be identified by more than one algorithm, the percentages reported above might overlap.

Conclusion

Case identification strategies have an important impact on the size and type of patient

Population. Harmonization of event identification is a substantial process, due to differences and availability of data in each source. The standardization approach proposed here allowed to benchmark results of individual component

algorithms from very heterogeneous data sources. In particular, our results showed how T2DM identification in the PCDs, HD and BD analyzed mostly relies on DIAG while in RLDs, where the presence of T2DM is manly inferred from drug utilization, DRUG-based algorithms identify the majority of retrieved cases.