Interpreting the effect of different concept sets, data domains and data provenances in cohorts from heterogeneous European data sources: examples of component strategy application from the EMIF and the ADVANCE projects

Rosa Giri1, Giuseppe Roberto1, Catrin Dodd1, Kaatje Boliers1, Alessandro Pasqua1, Lars Pedersen1, Miguel Angel Mayer2, Ron Herings3, David Ansell4, Peter Rijnbeek5, Rosa Gini6, Giuseppe Roberto1, Catrin Dodd1, Kaatje Boliers1, Alessandro Pasqua1, Lars Pedersen1, Miguel Angel Mayer2, Ron Herings3, David Ansell4, Peter Rijnbeek5, Rosa Gini6

1 (European University of Bolzano, Department of Data Science, Bolzano, Italy) 2 (Department of Maternal and Child Health, Hospital, Rotterdam, The Netherlands) 3 (PH Laboratory, Leuven, Belgium) 4 (Cancer Centre Oxford and Paediatric Care, Oxford, UK) 5 (Department of Clinical Epidemiology, Amsterdam UMC, Amsterdam, The Netherlands) 6 (Department of Internal Medicine, University Hospital of Navarra, Pamplona, Spain) 7 (Interuniversity EMIF-I2bLi3P and Emory University, Atlanta, Georgia, USA) 8 (EMIF project, Primary Care Research Unit, Delft, The Netherlands) 9 (European Centre for Disease Prevention and Control, Stockholm, Sweden) 10 (School of Public Health, Imperial College London, London, UK) 11 (Department of Internal Medicine, University of Antwerp, Antwerp, Belgium) 12 (Clinical and Translational Informatics, College of Medicine, University of Kentucky, Lexington, KY, USA) 13 (Rizzoli Institute, Bologna, Italy) 14 (SIDIAP database, Primary Care Research Institute Jordi Gol (IDIAP Jordi Gol), Barcelona, Spain) 15 (Allergo Jena, Jena, Germany) 16 (Cegedim, Paris, France) 17 (Department of Biostatistics and Epidemiology, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China) 18 (Laboratory of Medical Informatics, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China) 19 (EMIF-PCN project, Primary Care Research Unit, Delft, The Netherlands) 20 (Department of Public Health, Karolinska Institute, Stockholm, Sweden) 21 (Department of Pediatrics, retailer non-commercial software, University of Antwerp, Antwerp, Belgium) 22 (RIZIV/INAMI, Brussels, Belgium) 23 (National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands) 24 (Ministry of Public Health and Social Security, Mexico City, Mexico) 25 (Department of Public Health, Karolinska Institute, Stockholm, Sweden) 26 (Department of Pediatrics, retailer non-commercial software, University of Antwerp, Antwerp, Belgium) 27 (RIZIV/INAMI, Brussels, Belgium) 28 (National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands) 29 (Department of Public Health, Karolinska Institute, Stockholm, Sweden)

Identifying conditions in Europe

In a typical European data source based on electronic records, data may be only collected when a patient visits a primary care practice, or only when patients visit a hospital for inpatient care. When conducting a multi-national, multi-database study in Europe, medical conditions may be identified by different case identification algorithms. A process, called component strategy, was developed and tested in two European projects: EMIF and ADVANCE.

Define component strategies

Case-finding algorithms are split in simpler algorithms, each defined by a triple. The Unified Medical Language System (UMLS) was used to project concept sets to local terminologies and free text keywords.

Component set

Data domain

Data provenance

Diagnosis

Primary care practice

Diagnosis

Specialist practice

Emergency room

Diagnosis

Hospital

Diagnosis

Inpatient care

Inpatient diagnosis

Record linkage

Data provenance where was the information collected?

Data domain involved in diagnosis, drug, result from test, ...

Extract, compose and compare

All the data sources extract all the available components and compute the occurrence of components and of meaningful compositions in the study population. Occurrence is represented in the graphs below which enable comparisons. As an example of comparison: when a composite algorithm is OR B is represented, the share of the pink bar accounts for the share of subjects that would not be captured if B was not available: under suitable assumptions, this provides an estimate of sensitivity of A.

Record linkage DBs could extract diagnoses from inpatient, emergency care and death registry: they probably captured most of the cases in the underlying populations, although PPV of diagnoses may vary according to data provenance. Variability in cumulative incidence in primary care DBs was possible due to different local recording habits, such as use of free text. In one of the four primary care DBs, it was possible to extract diagnoses from inpatient care as well: assuming consistent PPV of primary and inpatient care diagnoses, sensitivity of primary care records was less than 60%.

In three data sources the IRs obtained by combining cohorts of specific and unspecified diagnoses were compatible with the notification, taking into account some misclassification in the unspecified diagnosis. In one data source the estimate was much higher, which might indicate under-notification of suspected cases normally confirmed to the relevant public health authority, or a higher rate of misclassification in the unspecified diagnosis in this data source. In another database, a high number of cases was recorded in the symptoms component, which in this data source was identified by a very specific string of free text (‘tos perusoide’) and the pooled estimate was compatible with the notification. In all the databases the percentage of cases recorded with a specific diagnosis was very low. The percentage of cases confirmed by a record of a positive diagnostic test was negligible in all the three databases where this data domain was available.

Application

In order to implement the component analysis in the OHDSI ecosystem, information on data provenance should be standardized. In a first attempt the simple classification in primary care practice, specialist practice, hospital, emergency room, and death registry, may lead to interpretable information. The impact of using SNOMED CT instead of UMLS should be assessed. The possibility of incorporating free text keywords in OHDSI concept sets should be explored in European data sources.

Conclusion

The nature of the disease under study is an important factor in the sensitivity and/or positive predictive value of a component. The systematic creation and comparison of component-based algorithms could be useful in the OHDSI ecosystem to improve the validity and efficiency of the data extraction. A systematic approach is needed in OHDSI to address the impact of the phenotypic definitions in a multi-database setting on study results.

Disclosure: This research was received support from the Innovative Medicines Joint Undertaking under ADVANCE grant agreement No: 115057