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 Gini[1], Giuseppe Roberto[1], Caitlin Dodd[2], Kaatje Bollaerts[3], Alessandro Pasqua[4], Lars Pedersen[5], Miguel Angel Mayer[6], Ron Herings[7], David Ansell[8], Sulev Reisberg[9], Lara Tramontan[10], Gino Picelli[11], Consuelo Huerta[12], Elisa Martin-Merino[12], Talita Duarte-Salles[13], Gianfranco Spiteri[14], Emmanouela Sdona[14], Paul Avillach[15], Peter Rijnbeek[2], Miriam Sturkenboom[3,16]

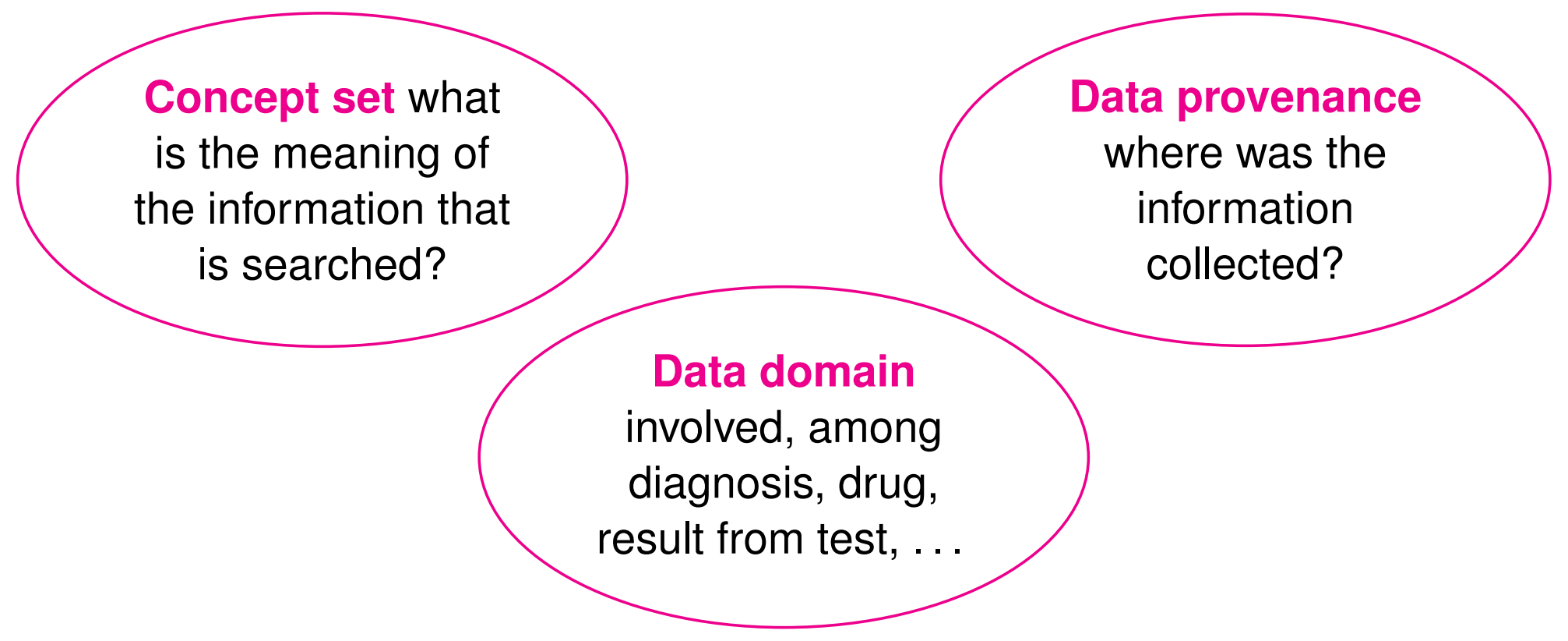
(1) Osservatorio di Epidemiologia, Agenzia regionale di sanità della Toscana, Florence, Italy; (2) Department of Medical Informatics, Erasmus Medical Center, Rotterdam, The Netherlands; (3) P-95, Leuven, Belgium; (4) Health Search, Italian College of General Practitioners and Primary Care, Florence, Italy; (5) Department of Clinical Epidemiology, Aarhus University Hospital, Aarhus, Denmark; (6) Hospital del Mar Medical Research Institute, Barcelona, Spain; (7) PHARMO Institute for Drug Outcomes Research, Utrecht, Netherlands; (8) The Health Improvement Network, Cegedim Strategic Data Medical Research Ltd, London, UK; (9) Quretec, Software Technology and Applications Competence Center, University of Tartu, Tartu, Estonia; (10) Arsenàl.IT Consortium, Veneto's Research Centre for eHealth Innovation, Treviso, Italy; (11) Pedianet, Padua, Italy; (12) Agencia Espaola de Medicamentos y Productos Sanitarios, Madrid, Spain; Institut Universitari d'Investigació en Atenció Primria (13) SIDIAP database, Primary Care Research Institute Jordi Gol (IDIAP Jordi Gol), Barcelona, Spain; (14) European Centre for Disease Prevention and Control, Sweden; (15) Department of Biomedical Informatics, Harvard Medical School & Children's Hospital Informatics Program, Boston Children's Hospital, Boston, MA, USA; (16) University Utrecht Medical Center, Utrecht, The Netherlands.

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 algorithms

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.



Concept set what is the meaning of the information that is searched?

Data provenance where was the information collected?

Data domain involved, among 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. Occurence is represented in the graphs below which enable comparisons. As an example of comparison: when a composite algorithm A **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.

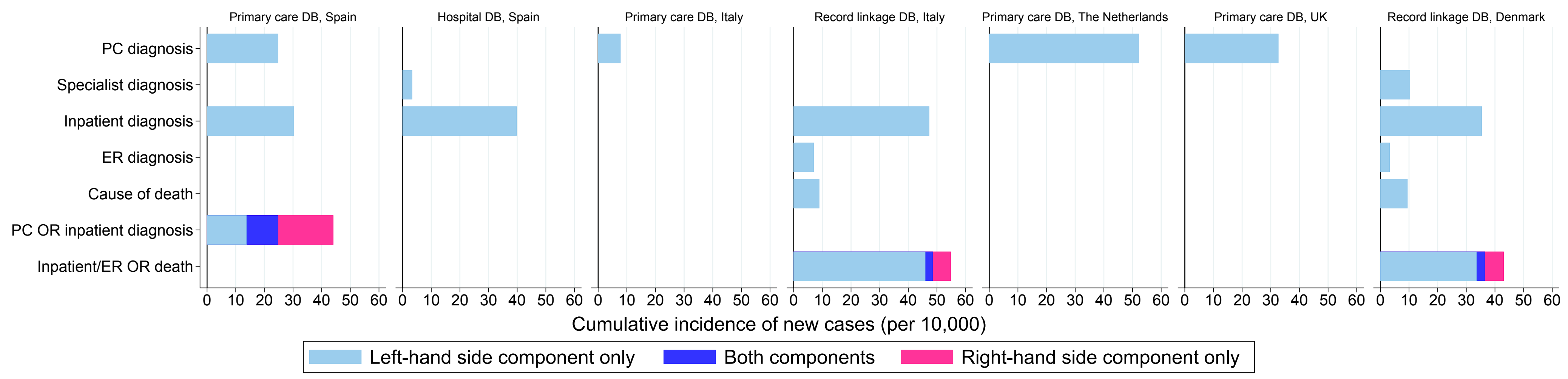
Interpret and decide

Decide which data-source tailored composite algorithm should be used. Obtain information on validity. Design sensitivity analyses using other algorithms.

An acute non-communicable disease: acute myocardial infarction

AMI is a cardiac emergency, which requires immediate medical attention and may lead to death before access to a medical facility. Six data sources from five European countries participating in the EMIF project were used. The one year cumulative incidence of new AMI cases in 2012 among subjects aged 45+, with at least two year of look-back and no previous record of an AMI diagnosis, was computed.

Component	Concept set	Data domain	Data provenance
PC diagnosis	(AMI)	Diagnosis	Primary care practice
Specialist diagnosis	(AMI)	Diagnosis	Specialist practice
Inpatient diagnosis	(AMI)	Diagnosis	Hospital
ER diagnosis	(AMI)	Diagnosis	Emergency room
Cause of death	(AMI)	Diagnosis	Death registry

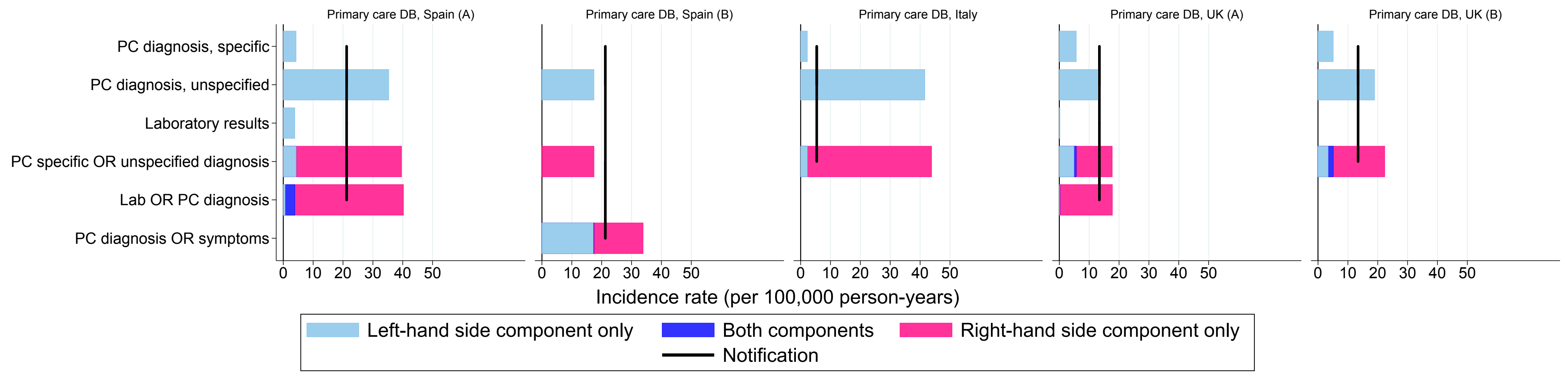


Record linkage DBs could extract diagnoses from inpatient, emergency care and death registry: they probably captured most of the cases in the underlying populations, although PPV of diagnoses may vary according to data provenance. Variability in cumulative incidence in primary care DBs was possibly due 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%.

An infectious disease: bordetella pertussis

Pertussis is an infectious disease of the respiratory tract, caused by the bacterium *Bordetella pertussis.* Only a specific laboratory tests can ascertain the specific microorganism responsible for the infection. National notifications of cases to the public health authority are reported annually by EU/EEA countries to the European Centre for Disease Prevention and Control (ECDC). In five databases participating in the ADVANCE project incidence rates per 100,000 person-years during the year 2012 and during the year 2014 were computed in the population aged 0-14 years. The incidence rates of the notified cases in the three countries were computed from the ECDC database.

Component	Concept set	Data domain	Data provenance
PC diagnosis, specific	(Bordetella pertussis)	Diagnosis	Primary care practice
PC diagnosis, unspecified	(Pertussis)	Diagnosis	Primary care practice
Laboratory results	(Positive result from a bordetella pertussis test)	Result from test	Primary care practice
Symptoms	(Symptoms of pertussis)	Symptoms	Primary care practice

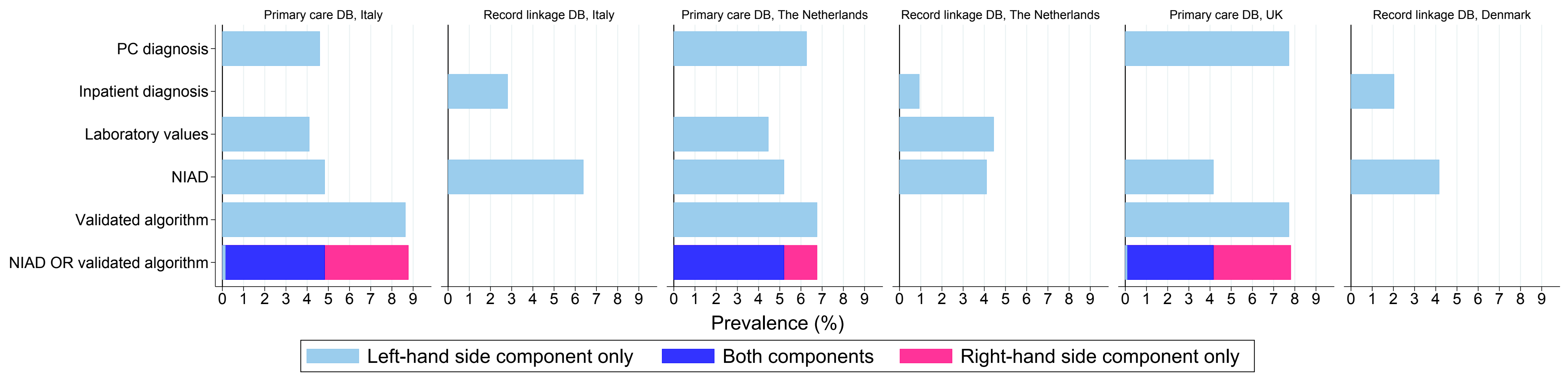


In three data sources the IRs obtained by combining cohorts of specific and unspecified diagnoses were compatible with the notification, taking into account some misclassification in the unspecified diagnosis. In one data source the estimate was much higher, which might indicate under -notification of suspected cases not finally 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.

A chronic disease: type 2 diabetes mellitus

T2DM is a chronic metabolic disease which needs diagnostic follow-up and, at a more advanced stage, regular pharmaceutical treatment with hypoglycemic drugs. T2DM care is typically provided by primary care physicians. In six population-based data sources from four European countries participating in the EMIF project prevalence of the cohorts at 1st January 2012 was computed in the adult population (16+).

Component	Concept set	Data domain	Data provenance
PC diagnosis	(T2DM)	Diagnosis	Primary care practice
Inpatient diagnosis	(T2DM)	Diagnosis	Hospital
Laboratory results	(Positive result from a T2DM test)	Result from test	Primary care practice
NIAD	(Non-insulin antidiabetic drugs)	Drug	Primary care practice or pharmacy



T2DM diagnoses in primary care could be extracted from 3 data sources, where a validated algorithm was also available. Between 54% and 76% of the cases detected by the validated algorithms were users of NIAD. The record linkage DBs could only extract diagnoses from inpatient care, which yielded non-realistically low prevalences. In those data sources, the component of NIAD may be used to detect T2DM cases: the sensitivity may be estimated between 54% and 76% from primary care DBs.

Application

In order to implement the component analysis in the OHDSI ecosystem, information on data provenance should be standardized. In a first attempt, a 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 empower 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 received support from the Innovative Medicines Joint Undertaking under ADVANCE grant agreement Nr: 115557

OHDSI Europe Symposium. Rotterdam, The Netherlands. March 23rd-24th, 2018.